Topological Analysis of Goal Scoring Patterns using Passing Networks

Mert Bildirici, Sam Borremans, Lauren Liu

github.com/Sam-B-Y/TDA-Soccer-Passing-Networks

Introduction

At the heart of a soccer team's strategy and performance is its passing network — a web of interactions that reflects how players connect and collaborate on the field. These networks hold valuable information about a team's style of play, major contributors, and overall effectiveness.

To uncover deeper structural patterns within these networks, Topological Data Analysis (TDA) can be used to study the "shape" of data by identifying its inherent patterns and structures. TDA has been successfully applied in sports like basketball [1] and hockey [2], where passing is crucial, offering valuable insights into team dynamics and performance.

With the growing application of data science in soccer [3], and particularly in analyzing passing networks [4], we wanted to explore how the topology of soccer teams' networks — especially their homology — correlates with their scoring. By examining these relationships, this study seeks to provide insights into passing strategies that can enhance goal-scoring outcomes in soccer.

This paper will look at the 2015/2016 season across Europe's top five leagues: the Premier League, the Bundesliga, Serie A, La Liga, and Ligue 1. It first introduces the construction and interpretation of passing networks, followed by an overview of the TDA applications employed. Finally, it analyzes the relationship between the homology of passing networks and the number of goals scored by a team, while also providing insights into the differences across these five leagues.

Passing Networks

The construction of these networks starts with raw match data, which we pulled from StatsBomb [5]. They collect passing data by using computer vision to track the ball's movement during a match, before manually verifying the data with human analysts [6]. Their data includes the location of the pass origin and destination, as well as the involved players, the type of pass, and the number of goals. Their data is free for certain games, including all games in the 2015/2016 season, which is why we chose to focus on these for this paper.

Players are represented as nodes, and the edges between nodes are normalized by dividing the frequency of passes exchanged between players during a match by the maximum number of passes in that match. However, unlike traditional passing networks where stronger connections are associated with higher weights, an inverted weighting scheme is adopted: the more frequent the passes between two players, the lower the weight of the edge connecting their nodes.

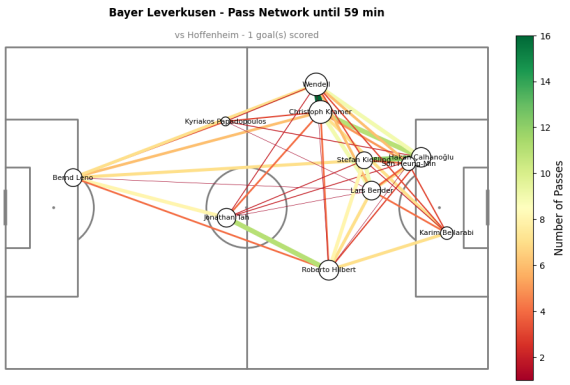

Figure 1: Example Passing Network Graph for Bayer Leverkusen

Filtration examines the evolution of a network's topology as a threshold parameter gradually increases, revealing features like connected components and loops at different scales. By inverting the weights, frequent passes (which signify stronger player relationships) appear first and persist longer, allowing us to capture the team's core structure in the early stages of filtration. This approach emphasizes the most critical interactions within the passing network and ensures the network captures a team's most essential collaborative framework.

Persistence

To quantify the structure of passing networks, we construct a *persistence diagram*, which encodes the

birth and death of topological features as a filtration parameter varies. In our setting, the nodes represent players, and edges represent (inverted) pass frequencies. By gradually "thickening" the network — starting from edges that represent the strongest (most frequent) passing connections, as these are closer to 0 in the inverted diagram, and progressively including weaker ones — we obtain a sequence of nested simplicial complexes known as a *filtration*. Tracking the homology groups of these complexes as the threshold changes reveals when particular topological features appear and vanish [7].

Formally, let $G = (V, E, w)$ be the weighted graph representing the passing network, where V is the set of players, $E \subseteq V \times V$ is the set of edges, and $w : E \to \mathbb{R}$ assigns each edge a weight based on the frequency of passes between the players. The edge weight $w(e)$ between players i and j is calculated using the formula:

$$w(e_{ij}) = \begin{cases} 1 - \frac{\text{count}_{ij} - \min(\text{count})}{\max(\text{count}) - \min(\text{count})} & \text{if count}_{ij} \neq 0 \\ \infty & \text{if count}_{ij} = 0 \end{cases}$$

where count_{ij} is the number of passes between players i and j, and min(count) and max(count) are the minimum and maximum pass counts across all pairs of players in the team for that match, respectively. This formula inverts the pass frequency, meaning that higher passing frequencies (stronger connections) result in lower edge weights, which helps prioritize the most frequent connections in the early stages of filtration. Furthermore, two players who have never passed the ball will not have an edge between them, as the weight of that edge is set to infinity, meaning it never appears.

We define a filtration parameter $\epsilon \in \mathbb{R}$. For each ϵ, consider the subgraph

$$G_\epsilon = (V, E_\epsilon), \quad \text{where } E_\epsilon = \{e \in E \mid w(e) \leq \epsilon\}.$$

To form a simplicial complex from a graph, we utilize the *Vietoris-Rips complex* construction. The Vietoris-Rips complex includes a simplex for any finite set of vertices where the pairwise distances (edge weights) are all below the threshold ϵ. For each ϵ, we construct the Vietoris-Rips complex VR_ϵ defined as:

$$VR_\epsilon = \{\sigma \subseteq V \mid \forall i, j \in \sigma,\ e_{ij} \in E \Rightarrow w(e_{ij}) \leq \epsilon\}.$$

If an edge e_{ij} is not present in E (i.e., $w(e_{ij}) = \infty$), then any simplex containing both i and j will not be included in VR_ϵ for any finite ϵ.

As ϵ increases, we obtain a nested sequence of simplicial complexes:

$$VR_{\epsilon_1} \subseteq VR_{\epsilon_2} \subseteq \cdots \subseteq VR_{\epsilon_m}.$$

From each complex VR_{ϵ_i}, we compute the homology groups $H_k(VR_{\epsilon_i}; \mathbb{F})$ over a field $\mathbb{F}$ (commonly $\mathbb{F} = \mathbb{Z}_2$). The k-th homology group is given by:

$$H_k(VR_{\epsilon_i}; \mathbb{F}) = \frac{\ker(\partial_k)}{\text{im}(\partial_{k+1})},$$

where $\partial_k : C_k(VR_{\epsilon_i}) \to C_{k-1}(VR_{\epsilon_i})$ is the boundary map on the k-th chain group. These homology groups detect topological features of dimension k: connected components ($k = 0$), loops ($k = 1$), and higher-dimensional voids ($k \geq 2$).

Persistent homology tracks these homology groups across the filtration. A topological feature (such as a loop) that is born at scale ϵ_b and dies at scale ϵ_d is represented as a point (ϵ_b, ϵ_d) in the persistence diagram:

$$Dg_k = \{(\epsilon_b, \epsilon_d) \mid \text{feature in } H_k(VR_\epsilon) \text{ for } \epsilon \in [\epsilon_b, \epsilon_d)\}.$$

From this analysis, we calculate *persistence statistics*, specifically the average and standard deviation of the life lengths of features in H_0 and H_1, which quantify the persistence of connected components and loops in the network:

- **Average Life Length of H_0:** Represents the average duration that connected components persist throughout the filtration. A higher average indicates that components merge more slowly, suggesting a more isolated team structure, with different players forming "triangles".
- **Average Life Length of H_1:** Corresponds to the average duration of loops or cycles in the network. Longer-lived loops can signify more stable subgroups within the team.
- **Standard Deviation of Life Lengths in H_0:** Measures the variability in the persistence of connected components. High variability may indicate inconsistent team connectivity.
- **Standard Deviation of Life Lengths in H_1:** Measures the variability in the persistence of loops. High variability can suggest fluctuating team strategies or cohesion.

In the context of soccer strategy, we interpret the following correlations between persistence statistics and the nature of passing networks:

- A higher average life length of H_0 suggests that connected components take longer to merge, indicating a more fragmented team structure.
- A higher average life length of H_1 implies that loops persist longer, which may correspond to stable passing subgroups within the team.

- Greater variability in the life lengths of H_0 and H_1 can indicate inconsistent team connectivity and strategy execution, or an evolving style of play during the season.

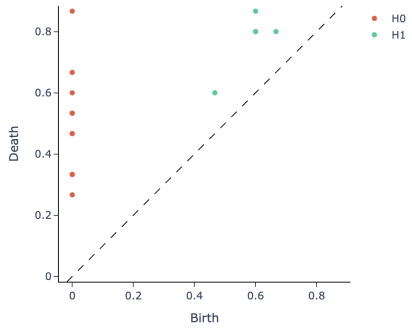

Figure 2: Example Persistence Diagram for a Team in a Single Game

We can note that some of these points have high multiplicity, as each player is "birthed" at 0, and sometimes "dies" at the same time with the simultaneous apparition of edges. Accounting for the multiplicity, this diagram has 10 H_0 points, which is what we expect from a team with 11 players (as there is one H_0 component that never dies).

TDA Methods

Using these persistence statistics, we aimed to draw conclusions about the homology patterns described above and the number of goals scored.

We calculated the average and standard deviation of the life lengths for both H_0 and H_1 from the persistence diagrams for all matches played by the same team in the 2015/2016 Serie A season. These statistics were then used to create scatter plots, where each point represents the average or standard deviation values for one team within a league.

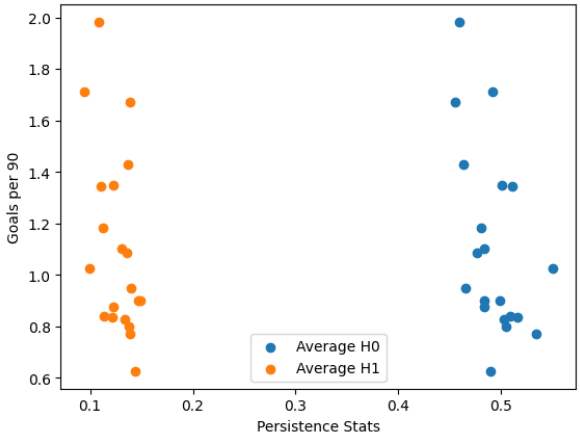

Figure 3: Plotting Average Life Length of H_0 and H_1 against Goals per 90 for Serie A

The standard deviation of the life lengths for H_0 and H_1 homology groups from the persistence diagrams was also plotted for each team, illustrating how fluctuations in these persistence statistics correspond to scoring outcomes.

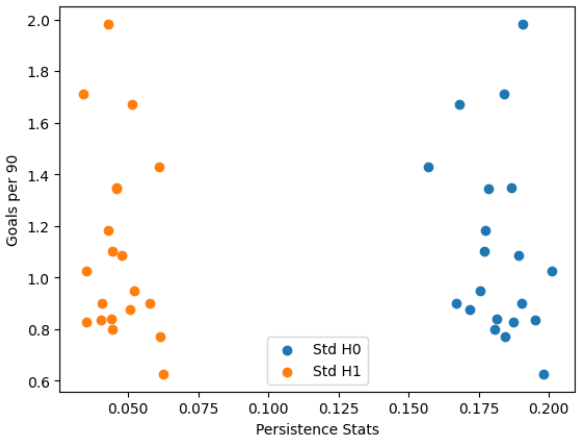

Figure 4: Plotting Standard Deviation of Life Lengths of H_0 and H_1 against Goals per 90 for Serie A

The correlations between these four persistence statistics and the target variable, Goals per 90, were then calculated.

Regression

A linear regression model was then trained using the average and standard deviation of the life lengths for H_0 and H_1, with goals per 90 as the target variable. For the Italian league, the model achieved an R^2 score of 0.642, which means that approximately 64.2% of the variability in goals can be explained by the persistence statistics derived from H_0 and H_1.

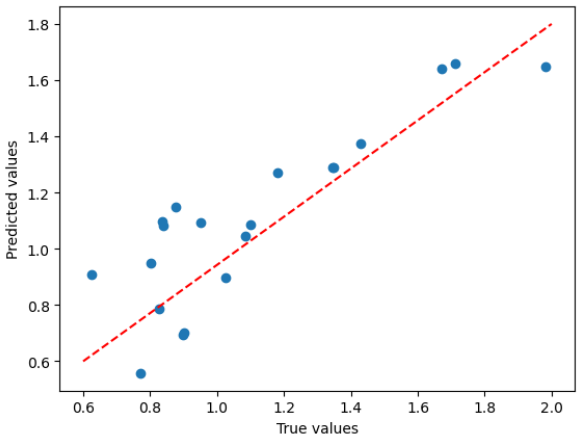

Figure 5: Plotted Predicted Goals per 90 against Actual Goals per 90

Comparison Across Leagues

We can extend our methodology beyond Serie A to the other four leagues, analyzing how homology impacts scoring with different playing styles and tactical approaches involved.

For example, as detailed above, leagues like Serie A utilize compact team formations and counterattacking play, which produces passing networks with fewer and more isolated clusters. In contrast, leagues like La Liga that emphasize fluid, high-possession styles may exhibit passing networks with more interconnected nodes and higher loop counts, indicative of sustained ball circulation and intricate passing sequences.

Variations in playing styles across leagues could influence the persistence statistics of passing networks in distinct ways, resulting in unique correlations between these statistics and scoring outcomes and making the number of goals scored easier to predict in certain leagues as compared to others.

The passing networks below represent Sampdoria in the 2015/2016 Serie A season and Barcelona in the 2015/2016 La Liga season, respectively.

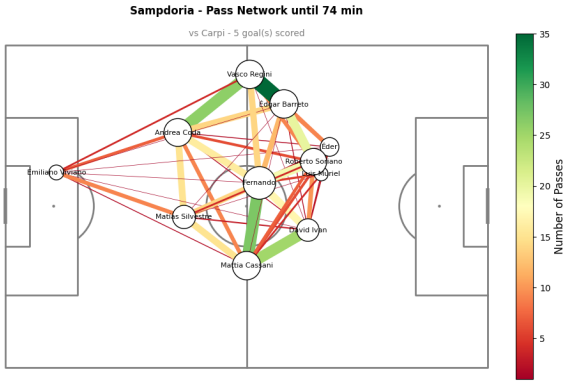

Figure 6: Passing Network of Sampdoria, against Carpi

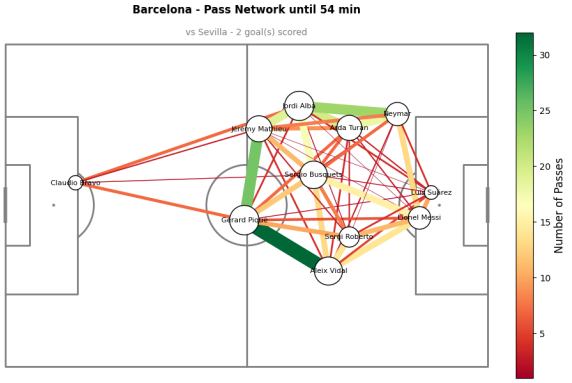

Figure 7: Passing Network of Barcelona, against Sevilla

Based on our methodology, we can obtain the correlation between each of the four persistence statistics and the target we are trying to predict (goals per 90), for each league. Additionally, we can calculate the R^2 of the linear regression that uses the four statistics to predict the target. The resulting table is below:

League	**Avg Life** H_0	**Avg Life** H_1	**Std Life** H_0	**Std Life** H_1	**Reg** $\mathbf{R^2}$
Bundesliga	-0.334	-0.283	0.412	-0.053	0.437
Premier League	-0.337	-0.123	-0.155	-0.229	0.184
Serie A	-0.483	-0.472	-0.226	-0.210	0.642
La Liga	-0.303	-0.422	0.511	-0.289	0.606
Ligue 1	0.143	-0.520	0.327	-0.338	0.346

Table 1: Correlation and Regression Analysis Across Leagues

Analysis

The table demonstrates that, as expected, the predictability of goal-scoring outcomes varies significantly across leagues.

In Serie A, the persistence statistics have the strongest relationship with the number of goals scored, with the linear regression model explaining 64.2% of the variance. The average life length of H_0 has a correlation of -0.483 with goals, while the average life length of H_1 shows a correlation of -0.472, both indicating moderate negative relationships. These results suggest that passing networks characterized by less repetitive passing patterns and more cohesive connectivity are associated with improved scoring opportunities. This finding aligns with Serie A's strategic style, which often emphasizes structured defensive setups and counterattacking play, relying on individual skill during offensive sequences to find the back of the net.

La Liga, despite its different playing style, shows similar correlations for the average life lengths of H_0 and H_1 with goal-scoring outcomes. However, the correlation between the standard deviation of the life length of H_0 and goals was a lot higher than other leagues, reaching 0.511. This positive correlation suggests that greater variability in team connectivity create more goal-scoring opportunities. This finding highlights the Spanish league's fluid and adaptive style, where versatility and the ability to adjust passing dynamics across different matches play a critical role in creating scoring chances. Overall, the linear regression model performs well for La Liga, achieving an R^2 value of 0.606.

The Bundesliga shows weaker correlations between persistence statistics and goal outcomes compared to Serie A and La Liga. The standard deviation of the life lengths of H_0 is the only feature with a meaningful correlation with goals, reinforc-

ing the idea that leagues with more dynamic and fast-paced gameplay require adaptability in passing structures to succeed.

In Ligue 1, the persistence statistics reveal some unique characteristics. The average life length of H_0 has a correlation of 0.143 with goals, making it the only league to with a positive correlation. This suggests that less cohesive structures—indicative of fragmented networks—may be advantageous for goal scoring in France. Additionally, the average life length of H_1 has the most negative correlation among all leagues, highlighting the negative role of stable passing triangles in this league. These observations indicate the importance of individual skills in Ligue 1, where players like Ibrahimović, Cavani, and Di Maria often create and convert scoring opportunities themselves.

Finally, the Premier League presents the greatest challenge for predictability, with the linear regression model achieving an R^2 of only 0.184. This result suggests that the Premier League's competitive and diverse tactical approaches lead to highly variable and adaptable passing structures, making it more difficult to link persistence statistics directly to goal-scoring outcomes.

Summary of Results

Serie A showed the strongest correlation between persistence statistics and the number of goals scored, while the Premier League showed the least correlation.

Due to differences in playing styles, tactical preferences, or other factors, a single pattern between the life lengths of H_0 and H_1 and the number of goals scored cannot be generalized across all leagues. However, the most useful statistic overall was the average life length of H_1, which seemed to remain the most consistent across all leagues.

Limitations

Several limitations of our analysis that could have impacted our results must be acknowledged:

- **Losing Positional and Directional Information:** Using persistence means we sacrificed all spatial information. We did not account for the player's average position and the distance from other players, for example, which plays a role in the passing network and could have impacted the number of goals. Furthermore, we summed the passes between two players to have a less sparse dataset, losing information about the direction of passes.
- **Passing Networks Restricted to Early Match Minutes:** We constructed passing networks only up until the first substitution. This decision was made to avoid adding new nodes to the graph, which would complicate the structure. However, it excludes important later phases of the match, including tactical adjustments and substitutions, potentially omitting significant changes in team dynamics.
- **Standardized Goals Estimate:** To standardize goals scored, we projected the scoring rate up to the first substitution across a full 90 minutes. While this approach ensures comparability, it assumes a constant scoring rate, which may not reflect a team's true performance over the course of a match.
- **Exclusion of Red Cards:** We did not account for red cards in our analysis. An early red card would remove a player from the field, effectively excluding them from the passing network. This could drastically alter the team's structure and performance but was not captured in our methodology.
- **Skewed Dataset:** Since many matches in the dataset recorded 0 goals scored, the data is skewed towards 0. This imbalance may bias the regression models and limit their ability to generalize, particularly in predicting matches with higher goal totals.
- **Limited Data Scope:** We included only leagues and seasons for which we had complete data for the entire season. While this ensured consistency, it restricted our analysis to a limited dataset. Access to more leagues and seasons could provide a broader perspective and more robust results. Also, inferring and drawing conclusions about a league from just a single season is not the most reliable, as it fails to account for variability across seasons, changes in team compositions, tactical evolutions, and other dynamic factors that influence league-wide trends.
- **Passing is Not All of Soccer:** Finally, our analysis focused exclusively on passing networks. While passing is a critical aspect of soccer, many other factors such as set pieces, individual skill, defensive tactics, team morale, and physicality contribute to scoring and overall performance. Thus, not all aspects of the game can be fully explained by passing metrics alone.

These limitations highlight the challenges of analyzing complex team sports like soccer and suggest several avenues for future research, including the incorporation of additional data, accounting for red cards, and integrating other aspects of the game beyond passing.

Conclusion

In this study, we applied TDA to soccer passing networks to investigate the relationship between network topology and goal-scoring outcomes. By scraping data from Europe's top five leagues during the 2015/2016 season, we utilized persistence statistics based on the life lengths of H_0 and H_1 to quantify team connectivity and passing loops. We plotted the correlation between the average and standard deviation of these life lengths against the number of goals scored, drawing conclusions about how network cohesiveness, consistency, and variability affect scoring in different leagues with different styles of play. We further used a regression model to analyze how much the scoring outcome depends on these persistence statistics.

Our findings revealed league-specific correlations between homological features and scoring, with Serie A showing the strongest predictive relationship, while other leagues, such as the Premier League, presented more variability. Furthermore, the average life length of H_1 contributed most consistently to the relationship between homology and the number of goals scored across all leagues.

Despite limitations like the loss of positional information and the focus on early match phases, the research highlighted the potential of TDA for understanding team dynamics and performance in sports. Future work can expand the data scope, incorporate additional game factors, and refine methodologies to build on these findings.

References

[1] J. Roehm, *An application of tda to professional basketball*, Mar. 25, 2021. [Online]. Available: https://www.youtube.com/watch?v=-cfp-tH-vIM.

[2] D. Goldfarb, "An application of topological data analysis to hockey analytics," *arXiv.org*, Sep. 25, 2014. DOI: 10.48550/arxiv.1409.7635. [Online]. Available: https://arxiv.org/abs/1409.7635.

[3] L. Lolli, P. Bauer, C. Irving, *et al.*, "Data analytics in the football industry: A survey investigating operational frameworks and practices in professional clubs and national federations from around the world," *Science and Medicine in Football*, pp. 1–10, May 14, 2024. DOI: 10.1080/24733938.2024.2341837. [Online]. Available: https://doi.org/10.1080/24733938.2024.2341837.

[4] J. M. Buldú, J. Busquets, J. H. Martínez, *et al.*, "Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game," *Frontiers in Psychology*, vol. 9, Oct. 8, 2018. DOI: 10.3389/fpsyg.2018.01900. [Online]. Available: https://doi.org/10.3389/fpsyg.2018.01900.

[5] S. "Github - statsbomb/statsbombpy: Easily stream statsbomb data into python." (), [Online]. Available: https://github.com/statsbomb/statsbombpy.

[6] Hudl Statsbomb — Data Champions. "Hudl statsbomb data — event data — hudl statsbomb." (Oct. 7, 2024), [Online]. Available: https://statsbomb.com/what-we-do/soccer-data/.

[7] M. E. Aktas, E. Akbas, and A. E. Fatmaoui, "Persistence homology of networks: Methods and applications," *Applied Network Science*, vol. 4, no. 1, Aug. 23, 2019. DOI: 10.1007/s41109-019-0179-3. [Online]. Available: https://doi.org/10.1007/s41109-019-0179-3.