IT University of Copenhagen | BSc in Data Science | 2021

Passing Networks and Performance

How Passing Affects Football Team Performance

Lasse Meinert Pedersen
BSc, Data Science
lasp@itu.dk

Supervisor:
Michele Coscia
Associate Professor of Computer Science
mcos@itu.dk

Abstract: Passing Networks in football (American English "soccer") have been subject to extensive research. One avenue of research that has not been thoroughly explored is how characteristics of said networks correlate with performance (quantified as expected goals (xG)) during games. Through thorough investigation, this project shows that during a regular season of 22 games in the Danish Superliga, having higher density, higher triadic closure, higher average degree and higher average clustering coefficient than the opposing team correlate positively with performance. This holds true for both individual teams, where 10 out of 12 teams show significant relationships between said measures and performance at $\alpha = 0.05$, and aggregated across the league, where the least significant relationship is significant at $\alpha = 1 \times 10^{-13}$.

1 Introduction

During the last decade, the field of football analytics has grown exponentially. Analyzing event data to evaluate players, teams and coaches is now industry standard and a crucial part of staying competitive. Passing networks are a natural extension of this movement because teams always pass the ball even when they do not score. In fact, the rarity of goals is the most notable feature of football that distinguishes it from other team sports (Gyarmati, Kwak, and Rodriguez, 2014).

Take the ball, pass the ball, take the ball, pass the ball. The immortal words of former FC Barcelona head coach, Pep Guardiola, later became the title of a documentary depicting Barcelona's dominance between 2008-2012. Barcelona's *pass to win* strategy made the team almost invincible during Guardiola's reign, and the team is considered to be one of the greatest teams of all time[1]. Opposing teams deployed counter-play strategies that involved limited possession and preyed on direct, counter-attacks with varying success.

Since Barcelona dominated the global football scene with their *pass to win* mentality, the game has evolved immensely. Pressing, work rate and defensive responsibilities are at the center of modern football; additionally, classical roles like the *10* is being reinvented[2].

While the game is constantly changing, a few questions remain the same: What separates successful and unsuccessful styles of play? Are possession dominant styles more successful across an entire league, or is it only true for individual teams (like the ones Pep Guardiola manages[3])?

This paper investigates how characteristics of passing networks correlate with performance during regular season games in the Danish Superliga 2020/2021. First, game data is sorted by timestamp and evenly split into six subsets containing passes and shots. Subsequently, six passing networks are constructed from each subset for each team and their characteristics are computed. Finally, the measures are correlated with performance. Performance is measured as the difference in expected goals (xG) between two teams and it is computed from the shots in each subset. In the passing networks, players are vertices and passes are directed edges connecting two players. The (directed) edges are weighted by the number of passes that travel (in that direction) between the connected players.

[1] The Greatest Teams of All Time: Barcelona 2008-12

[2] The Reinvention of the Number 10

[3] Man City crowned 2020-21 Premier League champions

2 Related Work

Passing networks have been subject to extensive research. The research distinguishes between three main types of passing networks: (i) player passing networks, where nodes are the players of a team, (ii) pitch passing networks, where nodes are regions of the field connected through passes made by players occupying them and (iii) pitch-player passing networks, where nodes are a combination of a player and his/her position at the moment of the pass (Buldú, Busquets, Martínez, Herrera-Diestra, Echegoyen, Galeano, and Luque, 2018). This paper utilizes the first type (i) of passing networks although the visualizations will show players in their *average* passing position throughout the game.

In passing networks, the importance of individual players (nodes) has been related to several network statistics: node degree - the number of passes made by a player (Cotta, Mora, Molina, and Merelo Guervós, 2011), eigenvector centrality - a measure of importance computed from the eigenvectors of the adjacency matrix (Cotta et al., 2011), closeness - a measure of how *close* each player is to their teammates (Peña and Touchette, 2012), and betweenness centrality - a measure of how *central* each player is in the ball's path when the ball moves within the team (Duch, Waitzman, and Amaral, 2010). The clustering coefficient has also been shown to quantify the impact of a player to the robustness of the passing network (Peña and Touchette, 2012) (Buldú et al., 2018).

Quantifying individual players' contributions or level of importance at *microscale* - analyzing the networks at node (player) or edge (pass) level - does not say much about the performance of the team. Instead, research has looked at the networks at *mesoscale* and *macroscale*. At *mesocale*, or the level of subcommunities (groups of players) within the networks, the surplus of certain types of passes between groups of players has been related to the success of a team (Gyarmati et al., 2014).

Finally, at *macroscale*, or the team level, a variety of network measures have been shown to be descriptive of the playing style and performance of teams. For instance, the network centroid's position and performance are related (the farther forward, the better). The team's average degree (mean number of passes) and the betweenness of players have also been related to performance: Duch et al. (2010) showed that the average betweenness of the top n players in a team can be used as an objective measure of performance, where an average betweenness (normalized log centrality) difference between teams $> .75$ means winning odds of 3:1 in favor of the better performing team. Similarly, Cintia, Rinzivillo, and Pappalardo (2015) found that the harmonic mean between average node degree and standard deviation of node degree correlates with success (Buldú et al., 2018).

The common denominator for the aforementioned *macroscale* research is that performance is measured by goals scored, shots on target, points accumulated or games won. In this sense, there is little research into how statistics of the passing networks *during* the game explain a team's chances of scoring *during* said game. As noted in Section 1, goals are rare in football. This rarity makes goals a sub-optimal indicator of (true) performance. Instead, this paper utilizes expected goals (xG) (Sections 3.1,3.2) as the performance indicator in an attempt to circumvent the rarity of the goal.

3 Data

The data is Wyscout Event Data provided by Brøndby IF (Danish Superliga team). Wyscout tracks the movement of the ball in all Danish Superliga games and creates data points; see Table 3.1) for each event. For this project only two types of events were of importance: passes and shots. Successful passes are all events where the ball travels between two teammates. This includes regular passes, throw-ins, free kicks, corners, etc. Shots are any attempt at scoring, even if the shot is blocked or off-target.

Table 3.1: Data

	1	*2*	...
eventId	665138045	665138046	
type	pass	pass	
minute	0	0	
second	3	6	
xG	0	0	
playerId	12345	54321	
playerTeam	SønderjyskE	SønderjyskE	
receiverId	54321	67890	
startX	50	29	
startY	49	64	
endX	29	67	
endY	64	5	

Most rows are self-explanatory in Table 3.1[4]: *eventId* is a unique identifier, *type* is either shot or pass (pass has several subtypes), *minute* and *second* are timestamps for the event, xG is the expected

[4]The original data has been transposed for readability

goal for this event (always 0 for passes), *startX, startY, endX, endY* are coordinates of where the ball starts and ends up. The coordinate system is placed as follows: on a horizontal field, the x-axis moves from left to right and the y-axis is inverted, moving from top to bottom. That is, (0,0) is the upper left corner, (100,0) is the upper right corner, (100,100) is the bottom right conner and (0,100) is the bottom left corner[5].

3.1 Expected Goals (xG)

Expected goals (xG) is a number between 0 and 1 that Wyscout attaches to every shot event[6]. The number is generated by a predictive machine learning model that assesses the likelihood of scoring from that particular spot on the pitch. Among other things, the model considers the location of the shot, the location of the assist and which body part the shooter uses[7]. A shot xG of 0,15 means that 15% of the time, the shot is expected to be converted to a goal. In essence, it is a measure of the quality of the chance created. When aggregated over a game, the team with the highest xG is expected to outscore its opponent.

3.2 xG as Performance Indicator

In this project, the difference between opposing teams' cumulative xG is used as a measure of performance. Having a higher cumulative xG means that the team is creating better chances of scoring and in that sense, outperforming its opponents. As noted in Section 1, the rarity of goals in football means that they cannot serve as a performance indicator, especially when dividing the game into subsets of events. xG solves this problem; there are *almost* always more shots on goal than actual goals. Simultaneously, xG expresses an objective measure of the quality of the goal scoring opportunity. When correlated with performance, xG outperforms possession, total shots ratio, goal difference and points scored, and it has been deemed the best single metric for understanding performance of a football team[8].

4 Methods

In an effort to make the full season analysis and the code digestible, a Python library of functions was developed during the project[9]. Alongside the library, *matchIds* for each of the Superliga's teams were extracted and stored in individual `csv`-files.

[5] Wyscout Glossary: Pitch Coordinates
[6] Theoretically, xG can be > 1 (see below)
[7] Wyscout Glossary: xG
[8] ASA: Single Game xG is here to stay
[9] Handed in alongside this paper

Finally, a Jupyter Notebook that imports the library and produces the results (Section 5) along with the anonymized data are both part of the self-contained hand-in.

4.1 Passing Networks

The passing networks are constructed in pairs of six per team per game (12 networks in total per game). Data from each game is sorted by timestamp and split evenly into six groups - *approximately* three per half. As discussed in Section 1 and Section 2, players are nodes and edges are successful passes between players. The edges are weighted by the number of successful passes travelling in a given direction between two team members.

The passing networks offer interesting insight. Consider the sequence of passing networks below. In Figure 4.1, the regular season average passing network of Brøndby IF is shown. This network shows the season average passing position for each player that has played[10] for Brøndby. The thickness of the arrows represents the number of passes between two players in the direction of the arrow. The *average passing position* is computed as the mean x,y coordinates from every pass that a player has made during the regular season.

Figure 4.1: Season Average Passing Network for Brøndby IF

The season average is aggregated over 22 games and in terms of football analytics, it is hard to gain any insights. It does, however, show the team's preferred formation and its overall interpretation of that formation.

Although the insights from regular season average networks are limited, passing networks from a single game paint a clear picture of the team's structure and style of play. Take for instance Brøndby IF's (Figure 4.2) fourth game of

[10] *completed or received a pass* for Brøndby IF

the season vs Randers FC (Figure 4.3). Brøndby's formation with 3 central defenders passing out from the back is very obvious. Likewise, Randers' two central defenders and two wide backs are also pronounced. Note that substitutes have not been filtered out and thus, there are more than 11 players in both figures. The subs were kept in the networks in an effort to better *catch* the effect of those substistutes in the affected passing networks.

Finally, the passing networks were split, offering *snapshots* of the development and position of the game (Figures 4.4,4.5). Here, the formations are harder to spot. It does, however, look like the two teams acted quite similarly throughout the game[11].

4.1.1 Network Statistics

The following statistics are computed for each of the six passing networks:

- Average Degree
- Average Clustering
- Density
- Triadic Closure

The opposing team's measure is always subtracted from a team's measure and the difference is then stored. Below are brief explanations of each measure.

Average Degree is the average number of distinct teammates that have passed the ball to a given player on the team. It is closely related to density when the number of nodes are fixed (which is true for passing networks until a substitution is made).

Average Clustering is the number of *neighboring* teammates that have also passed the ball between them (or the number of triangles around a player) (Buldú et al., 2018).

Density is a measure of how many teammates each player has passed the ball to. If every player has passed the ball to all his teammates, the density of the passing network is 1 (only possible until a substituion is made).

Triadic Closure is the fraction of all possible triangles present in the network. Essentially, if three players are connected by two passes:

$$Alice \xrightarrow{pass} Bob \xrightarrow{pass} Charlie$$

then triadic closure is the probability that there exists a pass

$$Alice \xrightarrow{pass} Charlie$$

that closes the triangle between the three players.

[11] Brøndby IF won the game at Randers, 2-1

4.2 Performance

When the games are split into six, the cumulative sum of xG is computed for each team and stored alongside the passing networks for each split.

4.3 Post-processing Data

As documented in Sections 4.1 and 4.2, each game was split into 6 and the cumulative xG and the network measures were stored for each team. The regular Superliga season 2020/2021 consists of 132 games (22 rounds of 6 games). So, in total, the analysis utilizes $6 * 2 * 132 = 1,584$ data points. Half of those share *absolute* y-value because the difference in xG is the same for each team in a given split (except for the sign).

5 Results

The following two sections explain how the network measures are correlated with xG on a season level (Section 5.1) and on a team-season level (Section 5.2). Finally, the results are used to predict the outcome of the post-season in Section 5.3.

5.1 Season as a Whole

As is apparent in Table 5.1 and Table 5.2, all four measures are statistically significant at α as low as 1×10^{-13}, with p-value 5.44×10^{-14} being the highest. It seems highly unlikely that the correlations are due to chance. Due to the low p-values, correcting for multiple hypothesis was deemed irrelevant.

Varying between 0.19 and 0.25, all of the correlation coefficients are positive and suggest the same thing: **Teams that pass the ball more frequently will create slightly more/better goal opportunities than those that do not**. Also, interesting to note is that, density is the best performing of the network measures across the board - with the lowest p-values (1.16×10^{-23}, 9.31×10^{-20}) and biggest correlation coefficients (0.25, 0.20) for both the Spearman and Pearson correlations. It seems that a well connected team, where players distribute the ball frequently to many of their teammates, is to be preferred to a team where players pass the ball to a small subset of their teammates.

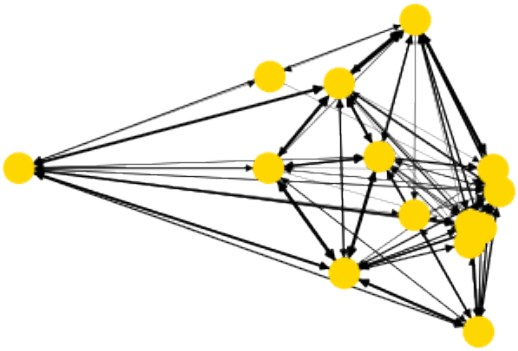

Figure 4.2: Brøndby IF (Week 4 vs RFC)

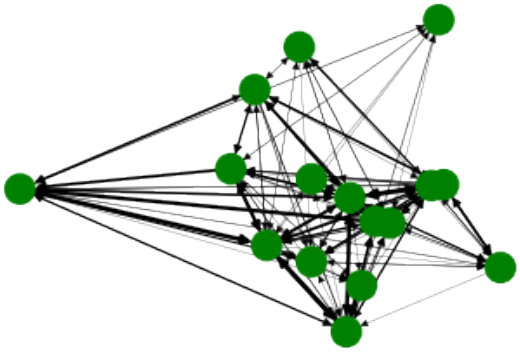

Figure 4.3: Randers FC (Week 4 vs BIF)

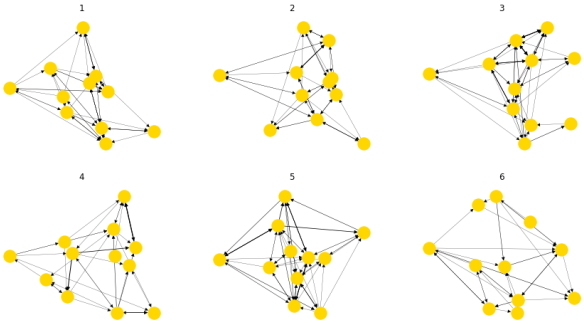

Figure 4.4: BIF Splits (vs RFC)

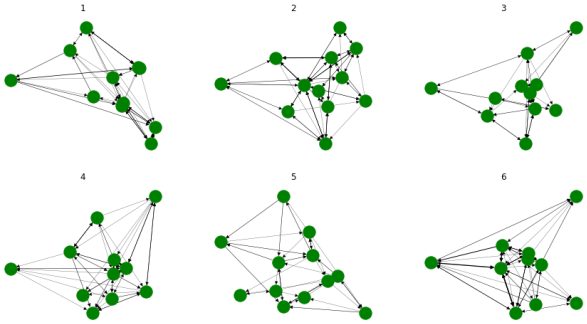

Figure 4.5: RFC Splits (vs BIF)

Table 5.1: Season Level Spearman Correlations

	Correlation	P-Value
Avg. Degree	0.24	6.15×10^{-22}
Avg. Clustering	0.21	5.49×10^{-17}
Density	0.25	1.16×10^{-23}
Triadic Closure	0.22	2.37×10^{-18}

Table 5.2: Season Level Pearson Correlations

	Correlation	P-Value
Avg. Degree	0.22	1.07×10^{-18}
Avg. Clustering	0.19	5.44×10^{-14}
Density	0.23	9.31×10^{-20}
Triadic Closure	0.20	3.97×10^{-15}

5.2 Team-Season

For the team season analysis, the number of data points heavily decrease and inevitably, so does the significance of the correlations. Instead of the 1,584 data points from the entire season, the number of data points per team is $6 * 22 = 132$.

Still, at $\alpha = 0.05$, ten teams (AGF, Horsens, København, Lyngby, Midtjylland, Nordsjælland, OB, Randers, SønderjyskE, Vejle) have significant correlations and of those ten, six teams (AGF, København, Lyngby, Nordsjælland, OB, Randers) have significant correlations for all eight.

On the other hand, two teams do not have a single significant correlation (AaB, Brøndby). Interestingly, Brøndby was top of the table by the end of the regular season and so it begs the question: how exactly Brøndby was so successful despite the lack of significance between passing and scoring chances?

Another interesting note is the relationship between the actual rank of the teams and how their *average correlation coefficients* rank. In Figure 5.1, the average correlation and significance are present along with the ranking of said correlations and the actual rankings. Note that the *average correlation coefficient* is a naive measure to show the teams' average correlation, while the aggregation does make the number itself hard to interpret. Nevertheless, it does offer some contrast between teams.

Of the top six teams in the actual rankings, Brøndby (1) and Midtjylland (2) do not have an *average correlation coefficient* in the top six of coefficients. The root mean squared error (RMSE) between the correlation rank and actual rank is ~ 4.7. In other words, predicting the season

Team	Correlation	Significance	Corr. Rank	Actual Ranking
Randers	0.323643	0.000161	1	5
Nordsjælland	0.318863	0.000504	2	6
København	0.317346	0.001877	3	4
Lyngby	0.256081	0.008819	4	11
AGF	0.249385	0.006126	5	3
OB	0.241889	0.005996	6	8
Horsens	0.192134	0.047728	7	12
Midtjylland	0.189404	0.041153	8	2
SønderjyskE	0.167392	0.072851	9	7
Vejle	0.140910	0.195415	10	10
Brøndby	0.117166	0.189419	11	1
AaB	0.084848	0.375062	12	9

Figure 5.1: Average Correlations and Ranking

rankings based on the strength of the teams' correlations alone would, on average, place every team almost 5 positions from where they really end up. Obviously, this makes correlation rank a bad predictor of season rank.

Instead, a better predictor must be the average density for each team. The strong positive correlations mean that teams are dependent on passes to perform. That is, teams with higher average density must be performing better – and in fact, they are. In Figure 5.2, the teams are ranked by their average density difference: the average difference between their own passing network's density and their opponent's passing network's density. Interestingly, the top three teams in this ranking hold the 1st, 2nd and 4th spot in the actual rankings. The RMSE between average density difference and regular season rank is ~ 3.6 and predicting the season rankings based on average density difference would, on average, place every team 4 positions from their true position. The average density difference predictions outperform correlation rank by more than one position in the table and below, the idea of predicting the outcome of the post-season with average density difference is entertained.

5.3 Predictions

If we were to predict the outcome of the upper-bracket of the post-season (10 games), where the top six teams (AGF, Brøndby, København, Midtjylland, Nordsjælland, Randers) play each other twice, based solely on their average density difference, Figure 5.3 would be our prediction.

Similarly, if we were to predict the outcome of the post-season for the lower-bracket, where the

Team	Avg. Density Dif.	Density Dif. Rank	Actual Ranking
Midtjylland	0.074750	1	2
København	0.065484	2	4
Brøndby	0.046147	3	1
Lyngby	0.028478	4	11
Nordsjælland	0.011643	5	6
Vejle	-0.000532	6	10
AaB	-0.000552	7	9
Randers	-0.015953	8	5
OB	-0.031848	9	8
AGF	-0.048958	10	3
SønderjyskE	-0.061990	11	7
Horsens	-0.066669	12	12

Figure 5.2: Average Density Dif. and Ranking

Team	Avg. Density Dif.	Density Dif. Rank	Actual Ranking	Predicted Rank
Midtjylland	0.074750	1	2	1
København	0.065484	2	4	2
Brøndby	0.046147	3	1	3
Nordsjælland	0.011643	5	6	4
Randers	-0.015953	8	5	5
AGF	-0.048958	10	3	6

Figure 5.3: Average Density Dif. and Ranking, Upper Bracket

bottom six teams (AaB, Horsens, Lyngby, OB, SønderjyskE, Vejle) play each other twice, Figure 5.4 would be our prediction.

Team	Avg. Density Dif.	Density Dif. Rank	Actual Ranking	Predicted Rank
Lyngby	0.028478	4	11	1
Vejle	-0.000532	6	10	2
AaB	-0.000552	7	9	3
OB	-0.031848	9	8	4
SønderjyskE	-0.061990	11	7	5
Horsens	-0.066669	12	12	6

Figure 5.4: Average Density Dif. and Ranking, Lower Bracket

6 Conclusions

Average degree, average clustering coefficient, density and triadic closure in passing networks are all slightly positively correlated with cumulated xG difference through the regular season of the Danish Superliga 2020/2021 at $\alpha = 1 \times 10^{-13}$. The four measures increase with the number of passes within a team and so, as highlighted in Section 5.1, teams that pass the ball more frequently than their opponents will, on average, outperform said

opponents in terms of quality/quantity of chances created.

Density is the best performing of the four measures with the lowest p-values and biggest correlation coefficients. This highlights that, in terms of accumulating better/more goal scoring opportunities, well connected teams where the ball travels frequently between many players is to be preferred to players only passing to a subset of their teammates. Alternatively, it could be that there is no benefit to making more passes but instead, teams with higher passing densities have more skillful players that are able to pass the ball with greater precision, thus making the team density higher and the team performance better.

Finally, accumulating higher xG does not automatically translate into a winning record. Football is multidimensional and there are many other variables that decide the outcome of a game. This is also obvious from the RMSE (~ 4.7) between correlation rank and actual ranking, and the RMSE (~ 3.6) between average density difference and actual ranking. That is, although the *density difference*-predictions better the *correlation rank*-predictions, neither are a good predictor of actual ranking.

6.1 Future Work

The post-season is, as this is being typed out, still underway and so one avenue of future work is to follow up on the predictions and the teams' correlations. How does playing against the other top teams affect each top 6 team's style of play? Are the games in the post-season different from those in the regular season like in the NBA (Masaru and L., 2010)? If they do change, how exactly?

Another interesting line of research is to look at the passing networks and how they interact. In the code library[12], there are building blocks for analyzing *leaks* between opposing teams from the passing data. In the library, a leak is when one team loses (leaks) the ball to the opposing team and specifically, it is an edge connecting players from opposite teams. These leaks act as proxy for how tight the competition is during a game - if the ball frequently travels between competing teams, the game must be competitive. Similarly, if the ball switches sides infrequently, one team must be dominating possession. This could also be tied into the analysis discussed above, where the number of leaks during the post-season could be compared to the same number for the regular season.

Finally, the library also holds infrastructure to build signed networks from the data. In essence, signed passing networks are networks in which every positive edge is one that connects two teammates and every negative edge is one that connects opposing players. Here, it could be interesting to introduce social network-scientific measures like frustration (a measure that the library has the code to compute).

[12] handed in alongside this article

References

Javier M. Buldú, Javier Busquets, Johann H. Martínez, José L. Herrera-Diestra, Ignacio Echegoyen, Javier Galeano, and Jordi Luque. Using network science to analyse football passing networks: Dynamics, space, time, and the multilayer nature of the game. *Frontiers in Psychology*, 9:1900, 2018. ISSN 1664-1078. doi:10.3389/fpsyg.2018.01900. URL `https://www.frontiersin.org/article/10.3389/fpsyg.2018.01900`.

Paolo Cintia, Salvatore Rinzivillo, and Luca Pappalardo. A network-based approach to evaluate the performance of football teams, 09 2015. URL `https://www.researchgate.net/publication/280520469_A_network-based_approach_to_evaluate_the_performance_of_football_teams`.

Carlos Cotta, Antonio Mora, Cecilia Molina, and Juan Merelo Guervós. Fifa world cup 2010: A network analysis of the champion team play. *CoRR*, abs/1108.0261, 08 2011. doi:10.1007/s11424-013-2291-2.

Jordi Duch, Joshua S. Waitzman, and Luís A. Nunes Amaral. Quantifying the performance of individual players in a team activity. *PLOS ONE*, 5(6):1–7, 06 2010. URL `https://doi.org/10.1371/journal.pone.0010937`.

László Gyarmati, Haewoon Kwak, and Pablo Rodriguez. Searching for a unique style in soccer. *CoRR*, abs/1409.0308, 2014. URL `http://arxiv.org/abs/1409.0308`.

Teramoto Masaru and Cross Chad L. Relative Importance of Performance Factors in Winning NBA Games in Regular Season versus Playoffs. *Journal of Quantitative Analysis in Sports*, 6(3):1–19, July 2010. doi:10.2202/1559-0410.1260. URL `https://ideas.repec.org/a/bpj/jqsprt/v6y2010i3n2.html`.

Javier López Peña and Hugo Touchette. A network theory analysis of football strategies, 2012.