A network-based approach to evaluate the performance of football teams

Paolo Cintia, Salvatore Rinzivillo, and Luca Pappalardo

[1] Department of Computer Science, University of Pisa, Italy
[2] KDD Lab, ISTI, National Research Council (CNR), Italy

Abstract. The striking proliferation of sensing technologies that provide high-fidelity data streams extracted from every game, induced an amazing evolution of football statistics. Nowadays professional statistical analysis firms like ProZone and Opta provide data to football clubs, coaches and leagues, who are starting to analyze these data to monitor their players and improve team strategies. Standard approaches in evaluating and predicting team performance are based on history-related factors such as past victories or defeats, record in qualification games and margin of victory in past games. In contrast with traditional models, in this paper we propose a model based on the observation of players' behavior on the pitch. We model a the game of a team as a network and extract simple network measures, showing the value of our approach on predicting the outcomes of a long-running tournament such as Italian major league.

1 Introduction

Thanks to the sensing technologies that provide high-fidelity data streams extracted from every game, in recent years football statistics have evolved in an amazing way. Nowadays professional statistical analysis firms like ProZone and Opta provide data to football clubs, coaches and leagues, who are starting to analyze these data to monitor their players, search for new talented ones, improve team strategies, and ensure themselves competitive advantage versus their peers. These football Big Data, which describe in great detail the behavior of teams during the games, pave the road to understand, model and possibly predict the complex patterns underlying sports success. An intriguing question is whether and how these data can be used to capture the performance of a team during a game: what are the features of the strongest teams? Can we extract from the data reliable measures of the performance of a team that correlate with its success during a competition?

Standard approaches provide a history-based answer to these questions: they assess the strength of a team using information about past victories or defeats, record in qualification games and other global competitions and margin of victory in past games. In this paper we provide a different point of view on the problem: in contrast with history-based prediction techniques, we describe the performance of a team by observing its *behavior* on the pitch as captured by

football data extracted from games. We show that this data-driven approach provides a description of the performance that shows an interesting correlation with the success of that team during the competition.

Starting from the list of frequent events occurred in the game – passes, crosses, assists, goal attempts – we model each football team as a complex system and infer a network whose nodes are players or zones on the pitch, and edges are movements of ball between two nodes, also labeled with weights to represent the amount of interactions among any pair of nodes. We describe the performance of a football team during a game by means of three simple measurements: the mean degree of a network's nodes, a proxy for the volume of play expressed by a team in a game, the variance of the degree of a network's nodes, a proxy for the diversity of play expressed by a team in a game, and a combination of the two. We observe a correlation among these performance indicators and the success of team, and therefore set up a simulation on the games of the FIFA World Cup 2014 and the Italian Serie A 2013/2014. The outcome of each game in the competition was replaced by a synthetic outcome (win, loss or draw) based on the network indicators of the teams in all the past games of the competition. We compare the outcomes of our simulation with the outcomes of two null models: a naive model which just sets the outcome of the game randomly, and a history-based model which assigns the victory to the team with the highest rank in recent official rankings. We observe that our approach outperforms the other models for long-running competitions as the Italian Serie A.

Football analytics has only begun to scratch the surface in the quest to understand, measure and predict performance. Our indicators have proven to be a good proxy of the performance of a team. If simple indicators like ours exhibit surprising connections to the success of teams, then a more complete view which includes defense strategy and movements without the ball has the potential of revealing hidden patterns and behavior of superior quality.

2 Football Data

We have football data about two football competitions: (i) the FIFA World Cup 2014 with 32 teams and 64 football games; (ii) the Italian Serie A 2013/2014 with 20 teams and 380 football games. In our dataset, a football game is described by a sequence of events on the field – passes, crosses, assists, goal attempts and so on. Each event consists in a timestamp, the player who generates the event, the position of the ball on the field when the event is generated, the position of the ball on the field when the event ends, the outcome of the event (completed or failed). Table 2 shows some examples of events as stored in our dataset, while Figure 1 shows the total number of football events in all the games of our datasets.

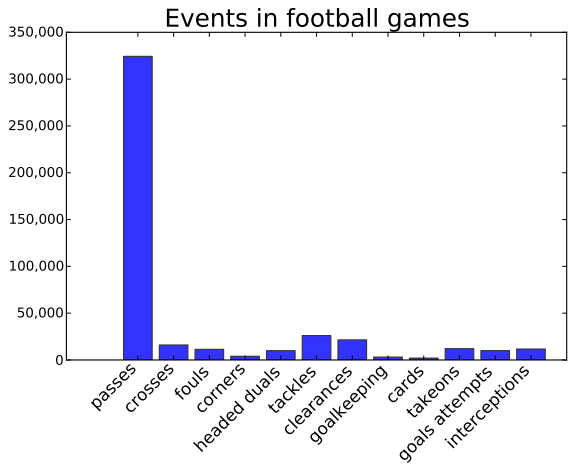

Fig. 1. The total number of events per category occurred in all the games of our datasets. We observe that passes are the most frequent events and constitute the 80% of all the events during a football game.

event	time	player	origin	destination	outcome
pass	17:24	Messi	(65.4, 20.2)	(67.8, 44.1)	completed
attempt	18:12	Messi	(49.8, 10.5)	(115.8, 10.5)	failed
assist	45:00	Pirlo	(65.87, 22.1)	(65.9, 30.6)	completed
cross	78:54	Tevez	(110, 31.1)	(115.7, 30.2)	completed

Table 1. Example of events of a football game. The event "pass" in the first row identifies a completed pass made by a player from position (87.2, 40.4) of the field; the event "attempt" indicates a failed goal attempt from position (49.8, 10.5).

3 A football team as a passing network

We describe the performance of a team in terms of *passing activity* using the information about pass events, the most frequent events occurring during a football game (Figure 1). We first represent the behavior of a team during a game by two kinds of *passing networks*. In the player passing network nodes are players and edges represent ball displacements between two players. In this type of network the number of nodes is constant across the different games and teams, while the density of edges and the networked structure define the passing strategy of a team during the game. Figure 2 shows a visualization of a player network extracted from a game by Juventus. We also introduce a zone passing network, a weighted directed network where nodes are zone of the pitch and edges represent ball displacements between the two zones (Figure 3). Formally, the zone passing network of team A is a weighted directed graph $G_A = (V, E)$, where V is the

set of zones (obtained by splitting the pitch into cells of size 11m×6.5m, 100 cells totally) and E is the set of edges, where an edge (z_1, z_2) represents all the passes, assists, or crosses started from zone z_1 and ended in zone z_2. In a zone passing network the number of nodes (zones on the pitch) varies across the teams and the games, allowing to detect significant differences in the passing strategy of the same team across different games. The player passing network and the zone passing network are abstractions of the team's behavior that synthesize the passing history during a game in a compact model, and it can be constructed efficiently from the event data. They can be used to determine hot zones of the pitch (zones where a team prefers to play) or at what extent the team uses short distance or long distance passes, to detect preferred positions for players, crosses, assists or shots, and even to understand how much predictable the strategy of a team is.

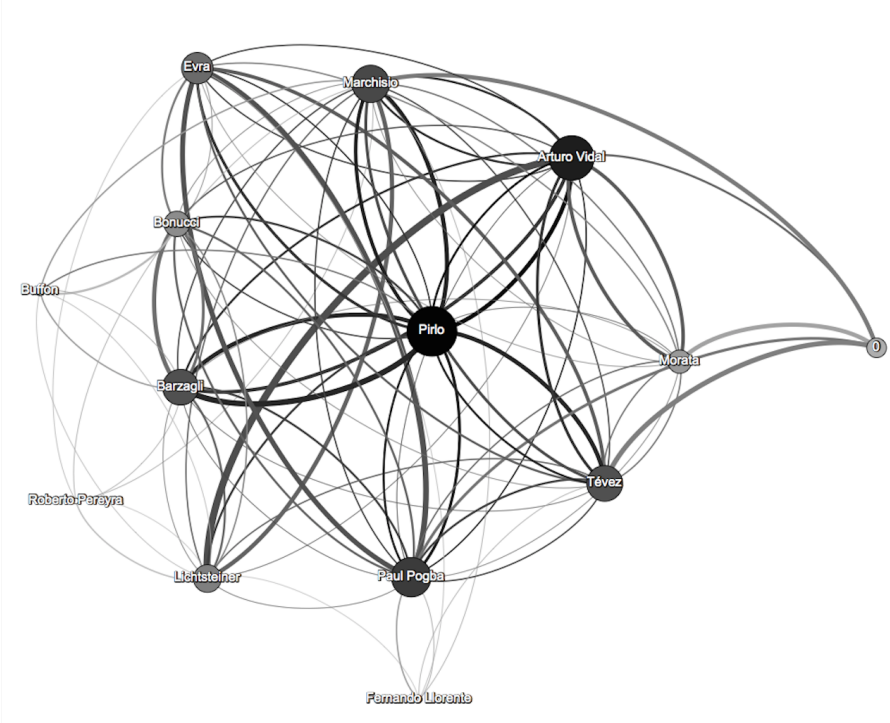

Fig. 2. A player passing network extracted from a game of Juventus in Serie A 2013/2014. Node are players, directed edges represent passes between players. The size of an edge is proportional to the number of passes between the players. Node 0 indicates the opponent's goal, and edges ending in 0 node represent goal attempts.

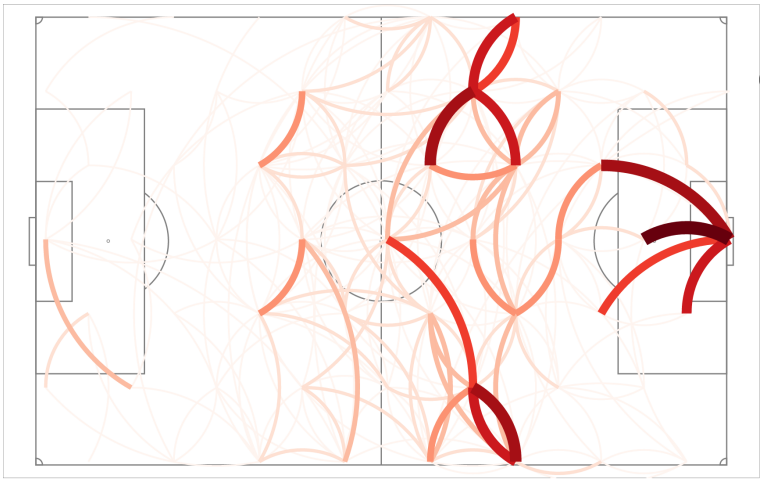

Fig. 3. Zone passing network of Argentina extracted from the FIFA World Cup 2014 semi-final. Node are zones on the field, directed edges represent passes performed by any players between two zones. The size of an edge is proportional to the number of passes between the zones.

4 Evaluating the performance of a football team

We describe the performance of a team T during a football game i by three network measures extracted from its player (zone) passing network: (i) the mean of nodes' degree μ_i^T, a measure of the passing *volume* expressed by the team during the game; (ii) the standard deviation of nodes' degree σ_i^T, a measure of the passing *heterogeneity* expressed by the team during the game; (iii) and a combination of the two measures by their harmonic mean $H_i^T = 2/(1/\mu_i^T + 1/\sigma_i^T)$. We compute the three network measures of teams for every game and observe a correlation among the proposed network measures and the success of a team during the competition (see Figure 4), and therefore set up a simulation experiment to validate our approach. In the simulation the outcome of each game of a competition (World Cup or Italian league) is predicted according to the value of the network measures of the two teams in all the past games of the competition. We simulate the outcome of game i of a football team T with the following two steps procedure:

1. for each of the two teams, we compute the three exponentially smoothed means of previous performances of the team $\bar{\mu}_{i-1}^T$, $\bar{\sigma}_{i-1}^T$, $\bar{H}_{i-1}^T$;
2. we compare the predicted measures of the two teams setting the team with the highest measure as winning.

At game i the performance history of team T is described as a list $L = P_1, \ldots, P_{i-1}$ where $P_{i-1} = (\mu_{i-1}, \sigma_{i-1}, H_{i-1})$. We build a prediction for the performance of a team at game i by computing the exponentially smoothed

means of previous performances of each team $\bar{\mu}_{i-1}$, $\bar{\sigma}_{i-1}$, $\bar{H}_{i-1}$. The exponential smooth is used to weight the recent past and take into account the recent shape of the team. We validate our model against two null models: the 48-26-26 model and the ranking model. In the 48-26-26 model the outcome is extracted randomly from a probability distribution computed on our data: 48% is the probability of a win for the home team, then 26% is the probability of a draw, 26% is the probability that the away team wins. The ranking model is a history-based model where the winner of a game is the team who ended the previous tournament in the highest standing. We take the FIFA official rankings updated to may 2014 for World Cup and the final rankings of Italian league 2012/2013 for Serie A 2013/2014. Table 2 provides the result of our experiments, where the values for null models are the means over 100 experiments. We observe that $\bar{H}$ is the measure that produces the best results for our model. The 48-26-26 model is the worst one predicting the outcome of games only in about the 30% of times. Our model outperforms both 48-26-26 model (30%) and ranking model (48%) for Serie A, reaching the best performance (0.53%) when using the player passing network. In contrast, for FIFA World Cup 2014 we have performance lower than the ranking model. Our results suggest that the proposed network measures are able to describe the performance of teams, adding predictive power with respect to the outcomes of games especially for long running competitions.

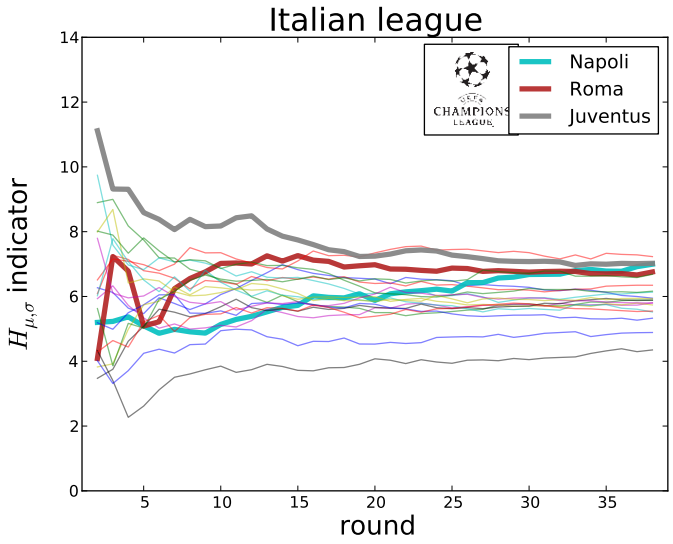

Fig. 4. Evolution of H indicator in player passing networks of Italian league. We highlight the three teams which achieved the qualification to Champions League. We observe that the strongest teams show the highest values of H measure.

Model	Serie A 2013/2014		FIFA World Cup 2014	
	Players network	**Zones network**	**Players network**	**Zones network**
$H_{\mu,\sigma}$	0.53	0.48	0.36	0.35
μ	0.44	0.44	0.31	0.34
σ	0.48	0.48	0.36	0.34
48-26-26 model	0.30		0.31	
ranking model	0.48		0.51	

Table 2. Prediction accuracy for Serie A 2013/2014 and World Cup 2014. We observe that our model performs better in long running competition like the Italian Serie A.

5 Related Work

In the last decade data science have entered the world of sports increasing its pervasiveness as the technological limits were pushed up. Many works exploit data mining or network science techniques to understand the complex patterns of success in both individual and team sports. Cintia et al. developed a first large scale data-driven study on cyclists' performance by analyzing data about workout habits of 30,000 amateur cyclists downloaded from popular fitness social network application [4]. The authors show that cyclists' wourkouts and performances follow a precise pattern and build an efficient training program completely learned from data. Hollinger analyzes NBA basketball games and propose the Performance Efficiency Rating, a measure to assess players' performance by combining the manifold type of data gathered during every game (i.e. pass completed, shots achieved, etc.) [5]. In the context of tennis, Terroba et al. present a pattern discovery exploration to find common winning tactics in tennis matches [6]. Smith et al. propose a Bayesian classifier for predicting baseball awards, prizes assigned to the best pitchers in the Major League Baseball. The model is correct in the 80% of the cases, highlighting the usefulness of underlying data on describing sports results and performances [7].

In the context of football, the possibility of observing strategies and decisions of teams by means of football data is attracting the interest of scientists and football teams [8]. Borrie et al. used T-Pattern detection to find similar sequences of passes from games [9]. Gudmunsson and Wolle analyzed and clustered players' sub-trajectories using Frechét distance as similarity measure [10]. The same authors encoded and mined typical sequences of passes by using suffix trees [11]. Still looking at the problem from a data mining perspective, Bialkowski et al. extracted players' roles over time by clustering spatio-temporal data on players' positions during a game [12]. Gyarmati et al. mined frequent motifs from teams passing sequences in order to classify team playing style [13]. They discovered that Barcelona Football Team, the most awarded team in the last decade, has unique passing strategy and playing style. Horton et al. performed a supervised learning of passes efficiency involving domain expert to rate the features of a pass between two players [14]. Lucey et al. built a shot outcome prediction method which considers strategic features like defender positions extracted from spatio-temporal data [15]. Taki and Hasegawa [16] introduced the dominant re-

gion model, a geometric model based on Voronoi spatial classifications where the football pitch is divided into cells owned by the players that reach every point of the cell before any other player [17]. Fujimura and Sugihara further developed the concept of dominant region defining an efficient approximation for region computations. Gudmunsson and Wolle built a passes analysis based on passing options computations, revealing the ability of a player to enforce and maximize the dominance of his team [11]. Other approaches represent football players as nodes of a passing network where passes are links. Peña and Touchette for example analyzed the games of FIFA 2010 World Cup through network analysis tools, showing that the two teams that reached the final (Spain and Netherlands) show the two highest values of average clustering [18]. Clemente et al. made a density evaluation of teams playing network showing how network metrics can be a powerful tool to assess players connections, strength of such links and help support decision and training processes [19].

6 Conclusion and Future Works

In this paper we describe the performance of a team during a football game by means of network indicators. We observe that these indicators correlate with the success of teams and them to predict the outcomes of the games in FIFA World Cup 2014 and Italian Serie A 2013/2014. We compare our results with the outcomes of two null models observing that our model performs better on longer and complex competitions like the Italian major league. As future work, we plan to include information about defensive events – tackles, goalkeeping actions, recoveries of ball and so on – and information about the movements of players without the ball. Defensive actions are crucial in the strategy of a team and they can make the description of a team's game more realistic. Moreover, studying the behavior of players without ball is crucial since it is known that most of the time (around 80% of the time) players move without the ball. Second, we plan to build other network features on the football networks to build models and classify the outcome of a game: which features are the most predictive of the outcome of a football game?

Acknowledgements

The authors wish to thank TIM and Mariano Tredicini for supporting part of our research. We also thank Marco Malvaldi, Fabrizio Lillo, Dino Pedreschi, Fosca Giannotti, Daniele Tantari, Adriano Bacconi and Maurizio Mangione for their insightful discussions. We also must thank Max Pezzali and Edoardo Galeano for the useful suggestions about the nature of football.

References

1. M. Lames and T. McGarry, "On the search for reliable performance indicators in game sports," *International Journal of Performance Analysis in Sport*, vol. 7, no. 1, pp. 62–79, 2007.

2. Fifa. [Online]. Available: www.fifa.com
3. R. Schumaker, O. Solieman, and H. Chen, *Sports Data Mining*. Springer, 2010.
4. P. Cintia, L. Pappalardo, and D. Pedreschi, "Engine matters: A first large scale data driven study on cyclists' performance," in *Data Mining Workshops (ICDMW), 2013 IEEE 13th International Conference on*. IEEE, 2013, pp. 147–153.
5. J. Hollinger, "The player efficiency rating," 2009.
6. A. Terroba, W. Kosters, and J. Vis, "Tactical analysis modeling through data mining," in *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval*, 2010.
7. L. Smith, B. Lipcomb, and A. Smikins, "Data mining in sports: Predicting cy young award winners," *Journal of Computing Sciences in Colleges archive*, vol. 22, 2007.
8. T. Reilly and A. M. Williams, *Science and soccer*. Routledge, 2003.
9. A. Borrie, G. K. Jonsson, and M. S. Magnusson, "Temporal pattern analysis and its applicability in sport: an explanation and exemplar data," *Journal of sports sciences*, vol. 20, no. 10, pp. 845–852, 2002.
10. J. Gudmundsson and T. Wolle, "Towards automated football analysis: Algorithms and data structures," in *Proc. 10th Australasian Conf. on Mathematics and Computers in Sport*. Citeseer, 2010.
11. J. Gudmonsson and T. Wolle, "Football analysis using spatio-temporal tools," *Computers, Environment and Urban Systems*, vol. 47, pp. 16–27, 2014.
12. A. Bialkowski, P. Lucey, P. Carr, Y. Yue, S. Sridharan, and I. Matthews, "Large-scale analysis of soccer matches using spatiotemporal tracking data," 2014.
13. L. Gyarmati, H. Kwak, and P. Rodriguez, "Searching for a unique style in soccer," *arXiv preprint arXiv:1409.0308*, 2014.
14. M. Horton, J. Gudmundsson, S. Chawla, and J. Estephan, "Classification of passes in football matches using spatiotemporal data," *arXiv preprint arXiv:1407.5093*, 2014.
15. P. Lucey, A. Bialkowski, M. Monfort, P. Carr, and I. Matthews, "quality vs quantity: Improved shot prediction in soccer using strategic features from spatiotemporal data." MIT Sloan Sports Analytics Conference, 2014.
16. T. Taki and J.-i. Hasegawa, "Visualization of dominant region in team games and its application to teamwork analysis," in *Computer Graphics International, 2000. Proceedings*. IEEE, 2000, pp. 227–235.
17. M. De Berg, M. Van Kreveld, M. Overmars, and O. C. Schwarzkopf, *Computational geometry*. Springer, 2000.
18. J. L. Peña and H. Touchette, "A network theory analysis of football strategies," *arXiv preprint arXiv:1206.6904*, 2012.
19. F. M. Clemente, M. S. Couceiro, F. M. L. Martins, and R. S. Mendes, "Using network metrics in soccer: A macro-analysis," *Journal of human kinetics*, vol. 45, no. 1, pp. 123–134, 2015.