A Novel Network Dataset Based on Football Players' Co-Appearances

Ahmad Zareie[1][0000−0002−2081−8112] and Rizos Sakellariou[2][0000−0002−6104−6649]

[1] Department of Computer Science, The University of Sheffield, Sheffield, UK
a.zareie@sheffield.ac.uk
[2] Department of Computer Science, The University of Manchester, Manchester, UK
rizos@manchester.ac.uk

Abstract. This short paper presents a dataset that contains a network graph, which captures relationships between football (soccer) players depending on whether they have been teammates. Most specifically, the dataset captures the co-appearance of players as members of the same squad in the UK Premier League over 30 seasons, spanning from the 1992-93 season to the 2021-22 season (inclusive). Such a dataset can be considered as a collaboration network graph and can facilitate a wide range of research applications. In sports analytics, it can be used to study player movement between teams, the impact of such movement on the network, and the career trajectories of players. In network science, it can provide a real-world example of network structures, allowing researchers to explore theories related to network evolution, formation, stability, and resilience. Additionally, this dataset can aid the development of machine learning models aimed at predicting team performance, player synergy, and the potential success of future team compositions.

Keywords: Network Graph · Football Players Networks · Affiliation Networks · Collaboration Networks

1 Introduction

The Premier League, officially known as the Football Association Premier League Limited, is the highest level of men's English football (soccer). Founded on February 20, 1992, the league operates in seasons, typically running from August to May. The inaugural season occurred in 1992-1993. In each season, initially 22 and later (from the season 1995-96) 20 football clubs are playing a match twice against each other. Each football club, or simply a team, is composed by a squad of several players, some of whom will appear in a match. Players can move between teams.

This short paper presents a dataset, which models the relationships among football players who have played in the Premier League as a network graph. A relationship is defined between pairs of players who have been members of the same team in one season *and* have appeared in at least one match of this

team in that season. In the graph, each player is represented as a node and the relationships between pairs of players are represented as edges. This dataset relates to a form of a *collaboration network* [3], which may be useful to analyze, understand or even use to optimize collaboration dynamics in order to form strong teams [8,14].

The remainder of this short paper first describes briefly some related work into similar networks and then explains the approach that was followed to build the network dataset. Section 4 presents a short analysis of the characteristics of the generated network dataset, including a range of statistics obtained from the network that was constructed. Finally, Section 5 concludes the paper and provides some suggestions for additional work.

2 Related Work

It has been suggested that Social Network analysis has significant potential in sports [13]. In fact, our dataset can be considered as an *affiliation network* according to the classification in Table 5 of [13]. Similar networks to the one we describe in this paper have been mentioned in different studies [11,3,7,9]. The earliest work we found that closely resembles our approach to generate a graph studies a network of NBA basketball players but only for one season [3]. To the best of our knowledge, most of these datasets are not publicly available. Some studies [6,5] deal with network datasets that model the relationships between teams in sports leagues (including the teams of the Premier League for the season 2013-14, a dataset which is available from [2]) but they do not include players as we do. Finally, more recently, interest is growing in using network analysis to examine events within a match, such as passing [4,10].

3 Methodology

Historical data for the 30 Premier League seasons (including teams, games, players, etc) is readily available through different sources (eg, news articles, magazines, football statistics enthusiasts and so on). We first gathered information on players' appearances for each team every season. Then, by considering each individual player as a unique node in a graph an edge is drawn with all teammates of this player (that is, everyone who was in the same team in the same season and has appeared in at least one match of this team in that season). Note that this definition of an edge does not imply that two players connected with an edge have also played together in the same match. It only means that they have been players of the same team over a period of time and each player has appeared in at least one match, so the two players co-appear in the roster of all players with some game participation in one season. The weight of each edge reflects the number of seasons the two players appeared on the same team. Since football players may change teams mid-season, the weights of edges can be decimal numbers, indicating partial seasons spent together.

The resulting graph, which models the relationships between players in the Premier League over 30 seasons as an undirected weighted graph, is connected (so it has exactly one component) with 5,295 nodes (players) and 173,932 edges (relationships). Detailed statistics related to this graph are given in the next section. The file containing the graph is named and it is part of a dataset which is available to download from:
`https://github.com/A-Zareie/Premier-League-Graph/`.

The *Premier League Graph* is a tab-separated CSV file with three columns: Source, Target, and Weight. Each row in the file represents an edge in the graph, describing a relationship between two players (source and target, represented by an id-number), and the weight of their connection.

In addition to the main file, which represents the graph of co-appearances of football players over the 30 seasons of the Premier League, we also release a dynamic version as part of the dataset. The dynamic version consists of 30 different tab-separated CSV files (with three columns as described above), named Graph_01 to Graph_30, each serving as a snapshot from the first season (1992-93) until the end of each of the 30 seasons (hence, Graph_01 represents the season 1992-93, Graph_02 covers the seasons 1992-93 and 1993-94 and so on). These files express the evolution of player relationships during the 30-year period, as each snapshot file captures co-appearances of players until the end of a specific season, allowing for the analysis of temporal changes in player relationships. Clearly Graph_30 is the same as the main file in the dataset (*Premier League Graph*). The 30 files are also available to download from the link already provided.

4 Statistics

Table 1 shows the value of the structural statistics commonly used by the KONECT project [1] for the *Premier League Network*. A short description of each statistic is presented in the table; for detailed descriptions of each statistic, refer to the KONECT project [1].

In contrast to many networks found in KONECT, it is worth noticing the high value of *Claw count* and *Triangle count* in our network. While some networks in KONECT may exhibit higher values for these properties, they are often larger, directed and/or multiple-edge networks. The relatively low *Gini coefficient* in our network, compared to similar datasets in KONECT, underscores the equitable distribution of node degrees, highlighting a notable departure from typical network structures. Furthermore, the proposed network demonstrates a notably high *Algebraic connectivity*, measured at 4.23484, particularly noteworthy given its size. This exceeds values observed in network datasets of a similar magnitude, indicating a robust level of interconnectedness within the network. Finally, another noteworthy aspect is the high *Spectral norm* of 0.84116 in our network. While there are networks in established repositories with greater spectral norms, they often entail directed or multiple-edge networks, making such a high value in a single-edge undirected network noticeable.

Table 1: The structural statistics of the Premier League Network

Statistic	Value	Description
Size	5,295	The number of nodes in the graph.
Volume	173,932	The number of edges in the graph.
Wedge count	17,211,958	The number of wedges in the graph. Wedges contain three nodes, only two of which are connected.
Claw count	801,457,705	The number of claws in the graph. Claws contain a central node connected to three other nodes.
Triangle count	2,230,255	The number of triangles in the graph. A triangle refers to three nodes all connected to each other.
Square count	85,507,019	The number of squares in the graph.
4-Tour count	753,251,848	The number of tours of length four that start and end at the same node.
Maximum degree	356	Maximum number of edges connected to any node.
Minimum degree	20	The minimum number of edges connected to any node.
Average degree	65.69669	The average number of edges connected to a node.
Fill (Density)	0.01241	The probability that two randomly chosen nodes are connected.
Maximum weighted degree	633.00000	The maximum sum of the weight of edges connected to any node.
Minimum weighted degree	19.50000	The minimum sum of the weight of edges connected to any node.
Average weighted degree	97.51653	The average sum of the weight of edges connected to a node.
Maximum k-core number	55	The maximum value k such that the graph contains a subgraph with all nodes having a degree greater than or equal to k.
Average path length	2.75027	Average distance between all pairs of nodes. The number of edges in the shortest path between two nodes is considered as the nodes' distance.
Diameter	5	The maximal distance between any two nodes.
Gini coefficient	0.35479	Degree inequality within the graph.
Relative edge distribution entropy	0.97541	A measure of the equality of the degree distribution.
Power law exponent	2.00785	A measure of the slope of the degree distribution.
Degree assortativity	0.05902	A measure of the degree distribution between pairs of connected nodes.
Algebraic connectivity	4.23484	How well connected the overall graph is.
Average local clustering coefficient	0.72948	The average clustering coefficient of nodes.
Global clustering coefficient	0.38873	The probability that two incident edges form a triangle by a third edge.
Spectral norm	223.45704	The largest absolute eigenvalue of the adjacency matrix of the graph.
Non-bipartivity	0.84116	The extent to which the graph deviates from a bipartite graph.

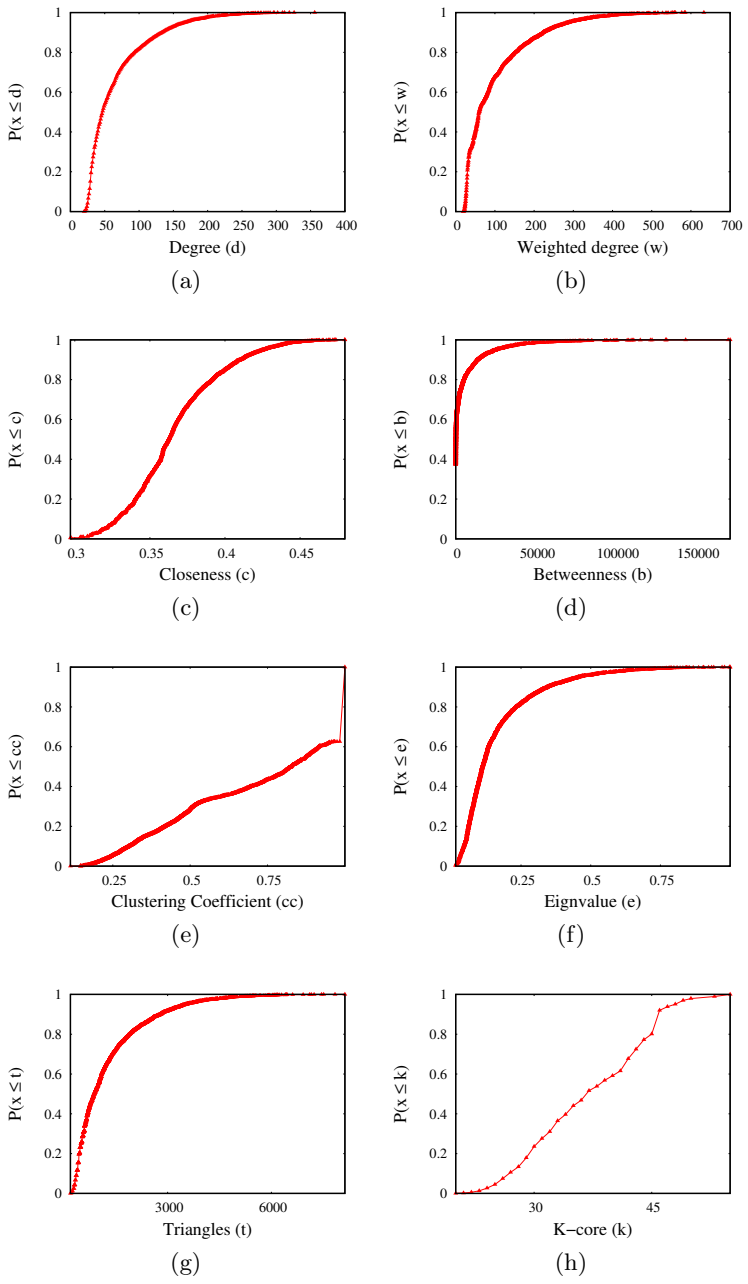

Fig. 1: Cumulative distribution of structural property metrics.

To analyze further the properties of the proposed network, we consider the structural properties associated with each node. Every node in the network graph is associated with a set of metrics based on the node's structural properties. Figure 1 shows the normalized cumulative distribution of different metrics in the network. A detailed description of these metrics can be found in [12]. In each plot shown in Figure 1, the x-axis represents different values of a metric, while the y-axis represents the probability that a randomly chosen node has a value less than or equal to a given value for this metric.

Figure 1(a) shows that 80% of nodes have a degree value ranging between 25 to 100, underscoring a trend of low inequality in node degrees. This consistency is also evident when examining the weighted degree metric in Figure 1(b). Likewise, as shown in Figure 1(c), approximately 90% of nodes exhibit a closeness value within the range of 0.3 to 0.4. As shown in Figure 1(d), the majority of nodes (approximately 80%) exhibit a low betweenness centrality close to zero. This indicates that a small subset of nodes serves as crucial bridges for most shortest paths between all pairs of nodes in the network. A similar distribution is observed for the eigenvalues and triangle metrics, as seen in Figures 1(f) and (g), where 80% of nodes display relatively small values, while only 20% show larger values. However, Figures 1(e) and (h) reveal a different distribution for the clustering coefficient and K-core metrics, indicating a more uniform distribution of nodes across the various values of these metrics.

5 Conclusion

This short paper presented a new network dataset suitable for social network analysis and research. This dataset captures co-appearances of football players in the Premier League over 30 seasons. Details of the statistics and distribution of structural metrics were provided to give an initial insight into the network characteristics. The dataset allows researchers to explore player collaborations, dynamics and network structures using data from one of the most competitive football leagues globally.

Future work could enhance this dataset in different ways. First, one can capture relationships in different ways, for example by considering only players who have actually played together in the same match, which would capture a more meaningful and accurate type of interaction. Second, one can integrate additional data, such as individual player statistics, team characteristics, matches and seasons' outcomes, to enable more detailed and diverse analyses. Third, one can analyze trends over time, which, at first instance, could consider the snapshots of the different seasons provided. In addition, comparative studies of similar graphs from different leagues or periods of time or even different sports could provide broader insights into team sports dynamics. Overall, we hope this dataset can motivate and provide an opportunity for further research into both sports analytics and network science and enable interesting applications.

Acknowledgments. This work was carried out when the first author was a PhD student at the University of Manchester, supported by a Dean's Doctoral Scholarship.

References

1. http://konect.cc/
2. http://konect.cc/networks/league-uk1-2013/
3. Boginski, V., Butenko, S., Pardalos, P.M., Prokopyev, O.: Collaboration networks in sports. Economics, Management and Optimization in Sports pp. 265–277 (2004). https://doi.org/10.1007/978-3-540-24734-0_16
4. Buldú, J.M., Busquets, J., Echegoyen, I., Seirul.lo, F.: Defining a historic football team: Using Network Science to analyze Guardiola's F.C. Barcelona. Scientific Reports **9**(1) (Sep 2019). https://doi.org/10.1038/s41598-019-49969-2
5. Fanuel, M., Suykens, J.A.: Deformed laplacians and spectral ranking in directed networks. Applied and Computational Harmonic Analysis **47**(2), 397–422 (2019). https://doi.org/10.1016/j.acha.2017.09.002
6. Girvan, M., Newman, M.E.: Community structure in social and biological networks. Proceedings of the National Academy of Sciences **99**(12), 7821–7826 (2002). https://doi.org/10.1073/pnas.122653799
7. Kooij, R., Jamakovic, A., Kesteren, F.v., Koning, T., Theisler, I., Veldhoven, P.: The dutch soccer team as a social network. Connections **29**(1), 2009 (2009)
8. Lappas, T., Liu, K., Terzi, E.: Finding a team of experts in social networks. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 467–476. KDD '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1557019.1557074
9. Onody, R.N., de Castro, P.A.: Complex network study of brazilian soccer players. Physical Review E **70**, 037103 (Sep 2004). https://doi.org/10.1103/PhysRevE.70.037103
10. 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) (Oct 2019). https://doi.org/10.1038/s41597-019-0247-7
11. Pardalos, P.M., Zamaraev, V.: The impact of social networks on sports. Social Networks and the Economics of Sports pp. 1–8 (2014). https://doi.org/10.1007/978-3-319-08440-4_1
12. Wasserman, S., Faust, K.: Social network analysis: Methods and applications. Cambridge University Press (1994). https://doi.org/10.1017/CBO9780511815478
13. Wäsche, H., Dickson, G., Woll, A., Brandes, U.: Social network analysis in sport research: an emerging paradigm. European Journal for Sport and Society **14**(2), 138–165 (2017). https://doi.org/10.1080/16138171.2017.1318198
14. Yu, S., Zeng, Y., Pan, Y., Chen, B.: Discovering a cohesive football team through players' attributed collaboration networks. Applied Intelligence **53**(11), 13506–13526 (Oct 2022). https://doi.org/10.1007/s10489-022-04199-4