SMM638 Network Analytics

Final Course Project

Group 5

Table of Contents

1 Introduction

Choices of decision making are extremely hard in this information blooming generation. In our case, football clubs face a choice of which team to trade with and which player to transit based on complex processes and criteria. However, bias, such as homophily, can act as a mental shortcut for decision making, the choice. People making choices based on conscious or unconscious bias can lead to drastically different results. Homophily is the "tendency for people to seek out or be attracted to those who are similar to themselves" (Oxford Reference, 2020). People observe homophily everyday, people become friends with friends that they share common friends with, or people are more likely to become friends when they like the same football team. Homophily have been widely studied in different social networks, and have been exemplified in many social relations. However, the conclusion of to what extent homophily affected choice affects the result are inconclusive. A study found that homophily choice constrain the variability of valuable information, while Hedge and Tumlinson's finding shows homophily choice results in doubling investment pay-off. (Erug & Gargiulo, 2012; Hedge & Tumlinson, 2018)

The dataset we chose revolves around the mobility network among professional football players, specifically containing data on 9 different leagues and a record of player transfers. What we particularly wanted to explore, as football fanatics ourselves, is the inter-organizational interactions and the extent of homophilous ties between teams within and outside of their respective leagues. Moreover, we wanted to explore how this would affect choice and performance, with the former involving observing whether transactions with teams, within or outside of their league, create tendencies - and the latter observing the effects on rankings of league tables, also delving into more social and cultural aspects.

To make things clearer and give an example, we would take Team A, Team B and Team C, all in the Premier League, and Team D in the Serie A. For homophily, we explore whether Team A has more ties and interactions with Team B or Team C since it is possible that a player being transferred into Team A from one of these teams has more experience in terms of being familiar with the playing style or other social/cultural factors, such as being brought up in the country of the league, being accustomed to the weather or speaking the same language of the country. However, for a player from Team D, a team outside of the Premier League, the situation may be different since there are known differences between the host countries of Serie A and Premier League in terms of culture, living conditions, play styles, and more. Consequently, we may expect that Team A would perform better more quickly, or in the short-term, if they had more ties within their league. Rather, with Team D, the impacts on performance may be experienced later due to the player possibly having to first adapt to the environment. To reinforce this argument, we explored a paper by D. Coates et al. (2020) that examined the player performance and concluded that it is optimal to have well-established relations with a small number of clubs, especially in a team's domestic league. All in all, we explore this hypothesis below in the analysis, highlighting several key and interesting findings.

2 Data Exploration and Transformation

2.1 Data Choices

We use several datasets in this project to highlight and analyze the characteristics of the task, which are homophily, choice, and performance. Homophily and choice are found through the dataset provided, on GitHub, of the mobility player network in conjunction with one self-collected dataset, 'football_dataset_final.xlsx', that provides us with information of the standardized team name and nationalities/country of origin of teams. More on the performance side, we used another self-created dataset, 'tier1_team_performance.xlsx', that provides league rankings of Tier 1 teams across time. We also decided to observe changes in homophily, choice, and performance over a 10-year period (2010 - 2020) since we assume that there may not be any significant ties between teams if we looked at a shorter or smaller time frame.

2.2 Data Transformation

For data transformation, the following steps were taken:

i. **Combined all the 8 leagues into a single dataframe:**

   This was done to simplify and efficiently access the different variables within the different leagues.

ii. **Focused on Tier 1 teams in 'club_name':**

   This way we were able to focus on the 260 teams, which were the Tier 1 teams of each country/league, omitting the English Championship (Tier 2 league of England), during the 2010 - 2020 period.

iii. **Filtered out lower-tier clubs from 'club_involved_name':**

   We did this to, again, only focus on the transactions and interactions between Tier 1 teams since we hypothesized that many of the top Tier 1 teams mainly acquire players from other Tier 1 teams within their country or overseas.

iv. **Incorporated self-collected datasets:**

   'football_dataset_final' dataset

   We decided to categorize the given leagues into their respective countries based on self-collected data online (UEFA, 2020), e.g. the Premier League would come under England, while the Serie A would come under Italy. This way, we counteract a prominent issue that might arise. The 'football_dataset_final.xlsx' file contains all the football clubs' nationalities, as well as the tiers in which these clubs belong to. Later, we collected all the football clubs that have played in Tier 1 of 8 countries between the season 2010/11 and 2019/20, then cross-checked all the club names from the dataset in the GitHub repository provided.

'tier1_team_performance' dataset

A secondary self-created dataset, 'tier1_team_performance.xlsx', contains the league rankings of all the 260 Tier 1 teams that are used to conduct the analysis - it includes the 10-year rankings for seasons 2010/11 to 2019/20. To analyze performance, we have collected each teams' wins and total matches played to calculate their winning percentage, then we also gathered their accumulated points for each season. Since the cycle for the latest football season has not yet finished, we did not include the current ranking for the 2020/21 season.

v. **Rectifying 'club_name' and 'club_involved_name' inconsistencies:**

The majority of the clubs had several inconsistencies in their names between the dataset provided. For example, 'Man Utd' would appear in the 'club_involved_name', while 'Manchester United' would appear in the 'club_name'. Using a for loop, we were able to transform the datasets such that the names of all clubs were consistent throughout by providing a coherent name. Then, there are also a few occasions where the clubs have more than one name in 'club_name' and 'club_involved_name', like 'Parma' has 3 different variants in the GitHub dataset: 'Parma Calcio 1913', 'Parma FC', and 'Parma'. Therefore, we had to find out all the possible different names that may appear in the dataset provided to modify them to a consistent name at the end.

vi. **Filter 'out' ties for Tier 1 teams:**

Since we focus on Tier 1 teams, which exist in both 'club_name' and 'club_involved_name', we are able to filter either the 'in' or 'out' ties since they are equivalent. For instance, if a player transfers from Juventus to Real Madrid, there would be one 'in' tie into Real Madrid and another 'out' tie out of Juventus - and since we just want to investigate undirected ties, we only require one. If it was the case of using lower tier teams, this would not be possible as the dataset only contains data on the 8 Tier 1 teams (and 1 Tier 2 team). Moreover, this halves the rows, and hence increases efficiency when calculating cosine similarities, as it takes up major computational power due to running considerable simulations.

3 Network Formation and Homophily

3.1 Network visualization

Premier League (England)	●	Serie A (Italy)	●
Eredivisie (Netherlands)	●	Liga Nos (Portugal)	●
Ligue 1 (France)	●	Premier Liga (Russia)	●
1 Bundesliga (Germany)	●	Primera Division (Spain)	●

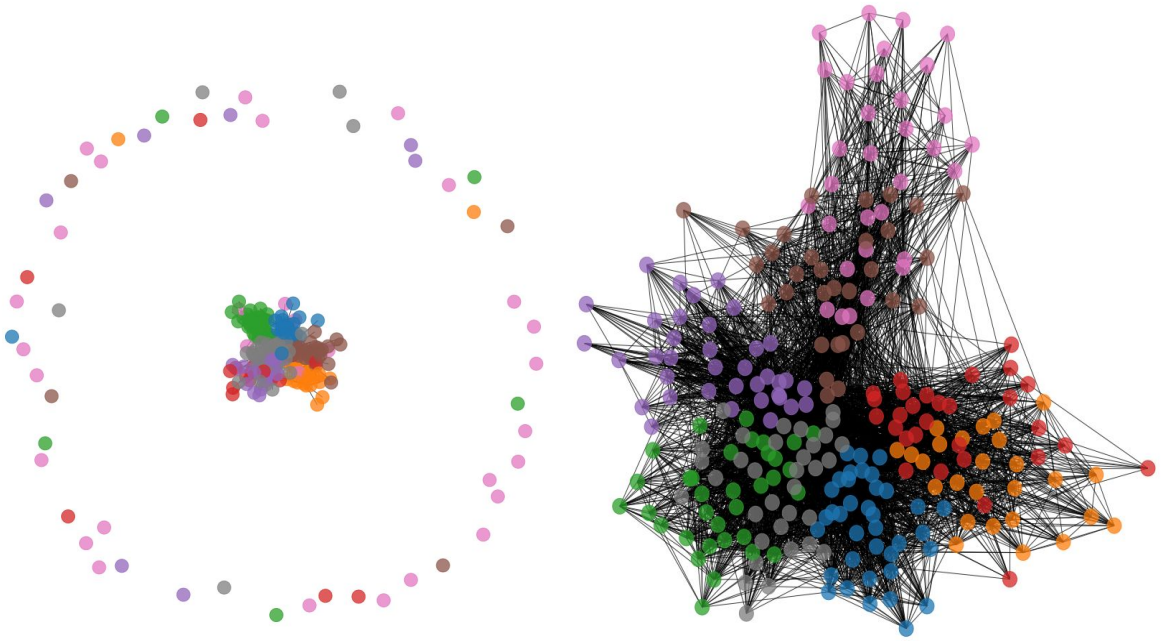


Figure 1 - Network Visualizations

Figure 1 demonstrates the formation of a network, indicating an adaptation of homophily choice for all Tier 1 teams in each country from base year 2010. In 2010 (visualization on the left), many teams have not yet traded with each other, thus less clustering can be observed. In 2020, (visualization on the right) we see an increase in trade between Tier 1 leagues. Additionally, a more obvious case of homophily can be seen, especially for the English Premier League (blue) and the Russian Premier Liga (purple). Even though we are able to visualize homophily, we have to mathematically prove its existence. In the next section, we conduct a probability calculation of homophily and cosine similarity to further prove the existence of homophily in our network.

3.2 Homophily calculation

To assess if the mobility network exhibits homophily, we used a similar method to one used in Networks, Crowds, and Markets (Easley & Kleinberg, 2010), which uses cross-variable probabilities. In our case, we examine the cross-league/cross-national probabilities as well as the within-league probabilities through the following equation:

$$1 = e^2 + d^2 + f^2 + g^2 + i^2 + p^2 + r^2 + s^2 + 2ed + 2ef + 2eg + 2ei + 2ep + 2er + 2es + 2df + 2dg + 2di + 2dp + 2dr + 2ds + 2fg + 2fi + 2fp + 2fr + 2fs + 2gi + 2gp + 2gr + 2gs + 2ip + 2ir + 2is + 2pr + 2ps + 2rs$$

where leagues include: e = England, d = Netherlands, g = Germany, i = Italy, p = Portugal, r = Russia, and s = Spain

Cross league trade probability: $2ed + 2ef + 2eg + 2ei + 2ep + 2er + 2es + 2df + 2dg + 2di + 2dp + 2dr + 2ds + 2fg + 2fi + 2fp + 2fr + 2fs + 2gi + 2gp + 2gr + 2gs + 2ip + 2ir + 2is + 2pr + 2ps + 2rs$

within league trade probability: $e^2 + d^2 + f^2 + g^2 + i^2 + p^2 + r^2 + s^2$

The equation above corresponds to the probability of the nodes connecting in the randomized network, essentially when nodes are equally likely to connect with one another. To test for homophily, we inspect whether the proportion of cross-league edges is less than the probability gained from the randomized network. We observe this from the figure below:

Figure 2

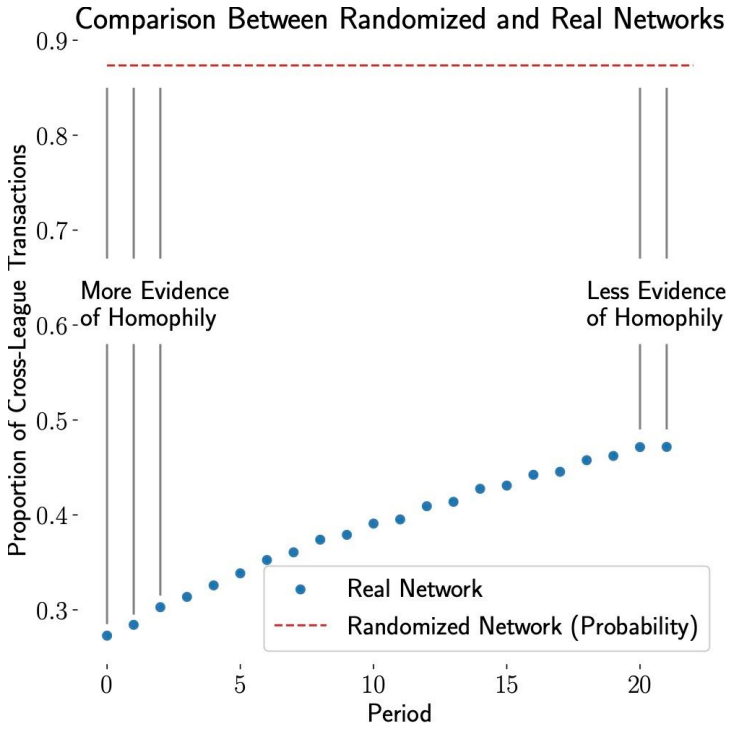


Figure 2 portrays the plot of both the real network and randomized network. The real network is based on the data we have used, while the randomized network is a situation where all of the teams are equally likely to trade (relating to the formula previously shown). As evident, the proportion of cross-league edges throughout the periods is lower than probability of having no homophily - thus, according to the condition outlined above, homophily exists in the mobility network. We can see that the gap between the real and randomized network gradually decreases, indicating homophily is also decreasing over time. An assumption we have made is that links between teams stay connected throughout the time periods, e.g. if Team A and Team B are involved in a transaction in 2010, that tie will still exist in 2020. We also relaxed and tested the assumption for when the links between teams do not stay connected (Appendix A) to be confident of our results and found that the outputs are the same (homophily decreasing).

To mathematically prove if cross-league homophily in the network exists, apart from the network visualizations in which the drawings (Figure 1) might present a homophilous distortion and the calculation executed in Figure 2 in which it holistically, not intricately, identifies that the network exhibits homophily, we measured the cosine similarities (Appendix B) of all the leagues in the year 2020, when the network's homophily level is relatively lowest. Using cosine similarities, we can iterate over a given probability, and hence obtain a more intricate and robust conclusion. To illustrate, the calculation indicates that a network should have approximately 86 cross-league edges out of 100; this number is

assumed to be fixed in the calculation done in Figure 2. However, the cosine similarity simulations are different each time: one iteration could indicate 84 and another 90. Hence, we obtain insightful statistical properties such as min, max, and mean (Appendix B).

Usually, a cosine similarity ranges between -1 to 1, where the former indicates a complete contrast and the latter a complete similarity. However, the comparison in our case is between non-negative vectors: domestic trades, international trades, and cross-national trades. Thus, the cosine index will be 0 to 1.

The comparison of the three mentioned elements is between the real network and the randomized network. Specifically, we analyse on a league-basis. If there is evidence of homophily at league-level, a small value of cosine similarity is expected. The results clearly illustrate that on a league-level, teams prefer to trade domestically than internationally as all the generated cosine values are close to zeros, indicating a stark difference from the randomized network.

4 Choice - Performance

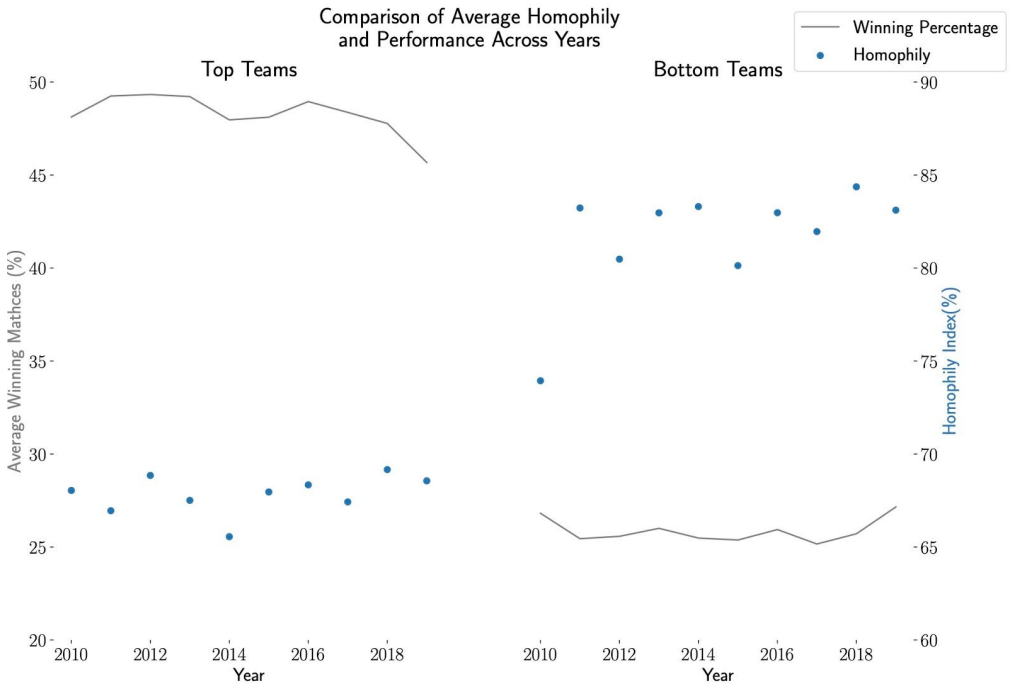
4.1 General Methodology

For the performance analysis, we decided to investigate the teams in the top half and bottom half of their respective league tables, sought out via the secondary self-created dataset, 'tier1_team_performance.xlsx', as a way of quantifying performance. In general, a common aspect of football leagues is that the bottom 3 teams are usually relegated into lower tiers, while in the next year, the top 3 teams of the lower tier get promoted. This would cause some loss or gain of links, though to counteract this issue, we decided to conduct our analysis on a yearly-basis, while categorizing teams into the 2 groups - top half and bottom half.

4.2 Scatter-Line Plot Analysis

Due to the argument of the relationships, we conduct an initial examination of the relationship between homophily indices (calculated by dividing all domestic transfer links by the total transfer links) and performance of the top half and bottom half teams aggregately. Figure 3 shows there could be a different tendency as to how homophily affects top rank and lower rank team performance. By comparing both scatter-line plots, no major relationships regard the change of homophily and performance each year for both top half and bottom half teams. This could mean that there is no relationship between the homophily between teams domestically and their performance. To fully discover the relationships between homophily and performance, regression analysis was conducted to deliver more statistical proof.

Figure 3



4.3 Regression Analysis

	Table 1 - Clubs in each Tier 1 leagues for each year					
Clubs	Coefficient	Standard error	p-value	R^2	AIC	Newey-West coefficient
All	0.4490	0.008	0.000	0.700	-175.3	9.4374e-5
Top	0.6336	0.012	0.000	0.786	-46.38	1.0633e-4
Bottom	0.3001	0.004	0.000	0.878	-1308	2.3677e-5

To statistically test for a relationship between homophily and performance, we conducted a regression analysis. The variable used for performance is the win percentage of teams since it accurately reflects a team's standings, much better than points due to variations in the size of some leagues e.g. the Primeira Liga consists of 18 teams where 34 games are played and a maximum of 102 points are attainable in a season, while the Premier League consists of 20 teams where 38 are played and a maximum of 114 points are attainable. Additionally, the win percentage accounts for matches apart from their main league such as UCL and Carabao Cup.

The results of the OLS regression in Table 1 illustrate the relationship with all teams in the whole network is positive with coefficient, 0.4490 - which can be interpreted as a higher level of homophily corresponding to increased performance or higher rank. The same method was applied to the top half and bottom half teams, in Table 2 and Table 3 respectively, which also show a positive relationship, though a stronger one with the top half teams. Despite the p-values being statistically significant and the R^2 values showing good fit, we uncover the

problem of autocorrelation and heteroskedasticity since we find that the Newey-West (N-W) coefficient is significantly lower in comparison to the general OLS coefficient. Particularly, Autocorrelation in this research might result from the accumulative good or bad performances of teams over years, which could lead to fitting the model to this cumulative performance. This may affect the biasedness of the coefficients and predictability of our model. The N-W coefficients, which provide consistent estimates in the presence of autocorrelation and heteroskedasticity, seen from the tables above imply that there is no correlation between homophily and performance.

5 Conclusion

5.1 Summary

Based on the results, the findings point towards there being no correlation between homophilous ties being beneficial or having a direct contribution to better results. In other words, increasing the diversity of transfers inwards of players from teams overseas or transferring players domestically does not have a clear effect on performances. A reason for this can be that there are other factors affecting performance, like we will mention in section 5.2, with the world of football, let alone the mobility/transfer network, being seemingly complex. This is a principally interesting result since it somewhat contradicts the initial hypothesis we had, stating that transferring players domestically within a league should ideally improve performance more so than transferring players from overseas. However, this can boil down to the specific structure of teams since we are not fully aware of teams' starting 11 squads, which essentially are the main determinants of performances - there are, of course, many other factors to consider.

5.2 Limitations

5.2.1 Homophily and Choice

Considering the nature of the mobility network, we did not give a close look to the weight of ties. The measurement of weight refers to the sensitivity of the network and is needed more consideration to avoid breaking the nature of the network. The accumulation and subtraction of any value between teams which have more than one transaction to each other can negative impact analysis.

Besides, although we found the existence of homophily in the network, however, the homophily choice is limited to only nationality, which may need a larger dataset to verify the result. Future direction can be to expand homophily choice, such as culture and religion, to delight the interaction of the football network.

5.2.2 Performance

With regards to the factor of performance, transfers may not be fully reflective of team performances since players transferred in perhaps may not play in the first year or may be injured for a certain period of time before they are played. Moreover, we do not know the structure of the individual teams based on the dataset, for example, whether a team's starting 11-squad is composed of mainly domestically or internationally transferred players. There

may also be managerial structure issues that also determine a large majority of the success of performances; for instance, if players, good or bad, are mismanaged or are not trained well, they are bound to not perform well.

We have to then consider the fact that we are limited to transfers within 8 leagues in Europe, although some well-established teams are composed of players that were transferred from teams in other areas of Europe, or other continents including South America and Africa. A prime example is a Brazilian footballer, Neymar Jr, who has been a massive contributor to the success of FC Barcelona in his early years, although he transferred to the team from a club in the Brazilian league in 2013 (BBC, 2013). Thus, that information is essentially omitted since we only consider interactions between Tier 1 teams within the 8 leagues.

5.3. Insights and Recommendations

Despite finding no link between homophily and performance, this may actually be quite insightful for managers and footballing lawmakers. For managers, it shows that when transferring players, it does not necessarily matter whether their previous team played in a domestic or international league, which is massive consideration for teams before acquiring players nowadays; thus, it perhaps emphasises the importance of management structures (as mentioned previously) and that players have unique reactions to different situations, not bound by socialisation and culture. For the footballing lawmakers, it may be useful in creating policies that nurture domestic talent, an example being England's 'Homegrown Player Rule' which requires there to be at least 8 domestic players in a full squad of 25 players (Goal.com, 2018). Extensive research within the topic may be able to aid in creating both domestic and international opportunities for football players at all levels.

In addition, we would recommend exploring a much wider span of leagues and, perhaps, also investigate the scouting networks within leagues. Particularly, we are likely to see strong homophily and performance links in under 21 teams where a massive composition of the teams include native players. Moreover, it may be interesting to investigate the social and cultural facets, as we did in this project, but more on a player-level rather than team-level. For instance, we could analyse the aspect of 'chemistry' that refers to players joining new teams based on factors such as having ethnic ties with incumbent players and managers or if the whole team is composed mainly of players from the same nationality, which would increase 'chemistry' and allow the team to play better due to enhanced socialisation and communication. A prime example is the team of Wolverhampton Wanderers (or Wolves) who appointed Nuno Espirito Santo, of Portuguese descent, in 2017 (BBC Sport, 2017) and, over the years of his tenure, bought a lot of Portuguese football players, factoring in that the players were likely to be drawn by the social tie with the manager and the incumbent Portuguese players. Consequently, Wolves ended up getting promoted into Premier League after finishing 1st in the English Championship during the 2017/18 season and, in the following year, finished in 7th place of the Tier 1 league (Sky Sports, 2018). Though, again, in order to conduct this type of analysis, a database with the nationalities of all players and the composition of teams, as well as starting 11 squads, needs to be collated. Moreover, one crucial caveat to this example is that different entities can react differently to a given environment, and hence continual experimentation might illustrate valuable and tailored

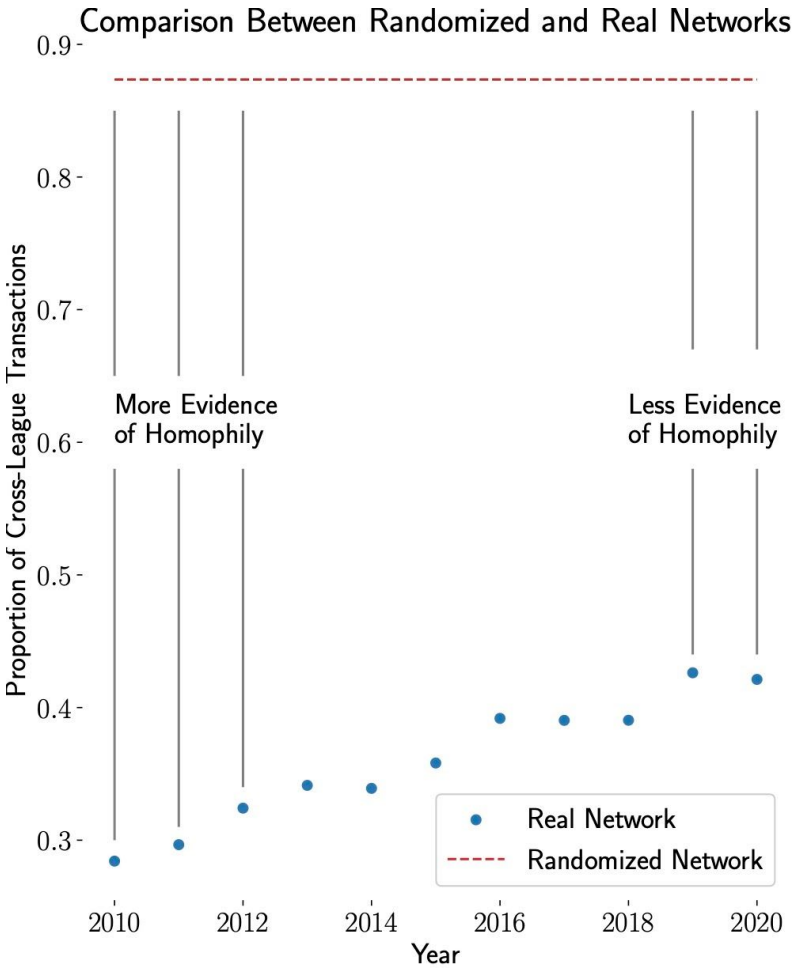
insights, just as A/B tests do for tech firms. For example, even though we found no correlation between performance and homophily, the average homophily indices of the top and bottom half teams are quite different, and the same case for performance. This phenomenon might stem from the fact that top teams have experienced a diminishing marginal return effect, meaning more heterogeneous or homogeneous teams do not tangibly affect performances anymore; this could be the same for bottom half teams. The way to test this may be an analysis conducted on a team that changes the team composition drastically in a short period.

The concept of closure can also be an interesting phenomena to investigate, both on a team-level and player-level. On the one hand, we would be able to notice trends in the categorisation of domestic teams or Tier 1 teams transferring players from a certain league or country. For instance, a few Bundesliga teams have acquired players from United States' Major League Soccer, with footballing giants Borussia Dortmund acquiring Christian Pulisic, and then in consequent years, Bayern Munich acquiring Chris Richards (Sports Illustrated, 2019). Thus, we see some form of team-level triadic closure but more on a rival basis and it would be quite interesting to investigate this concept further with additional data.

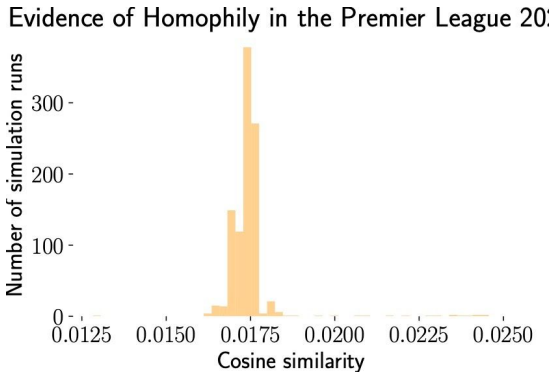
All in all, professional football is rather complex to truly and succinctly analyze since the factors to be considered, especially when looking at mobility/transfer networks, are extensive. Hence, further and more comprehensive analysis would ideally help understand and dissect the complexities of this network.

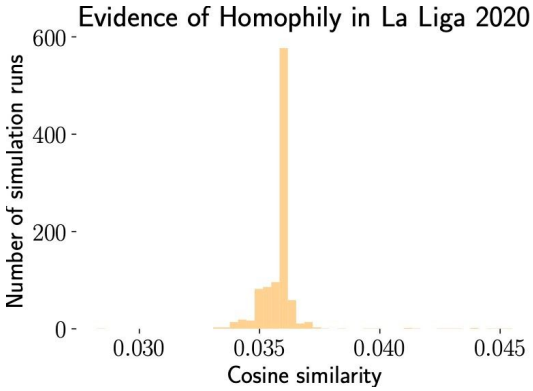
Appendix A

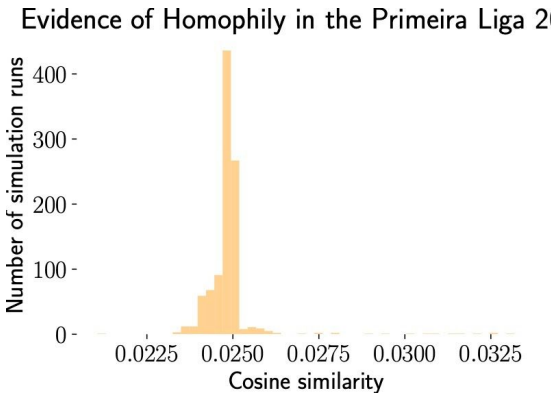
Figure 4

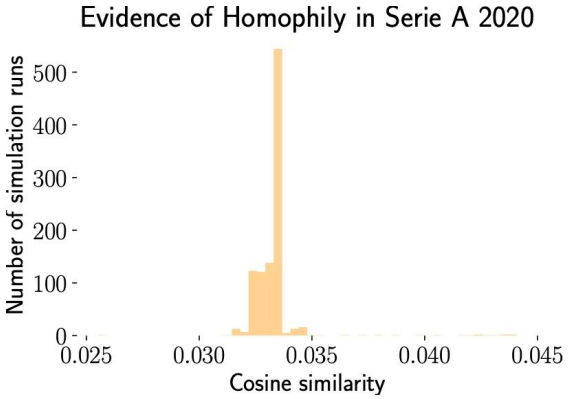


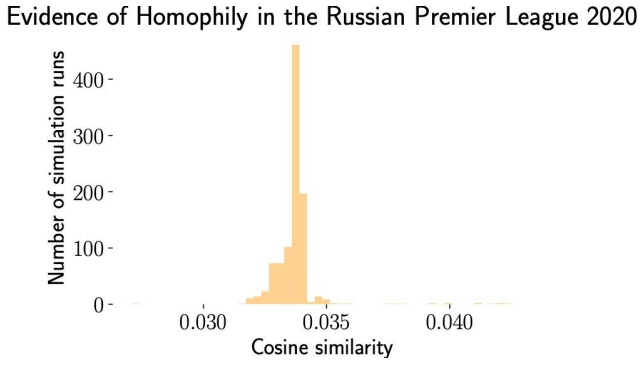
Appendix B

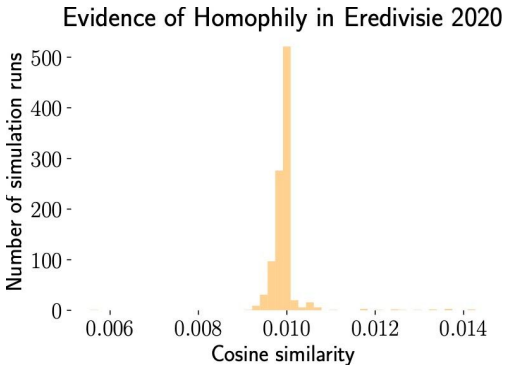

Evidence of Homophily in the Premier League 2020
Number of simulation runs
0
100
200
300
0.0125
0.0150
0.0175
0.0200
0.0225
0.0250
Cosine similarity

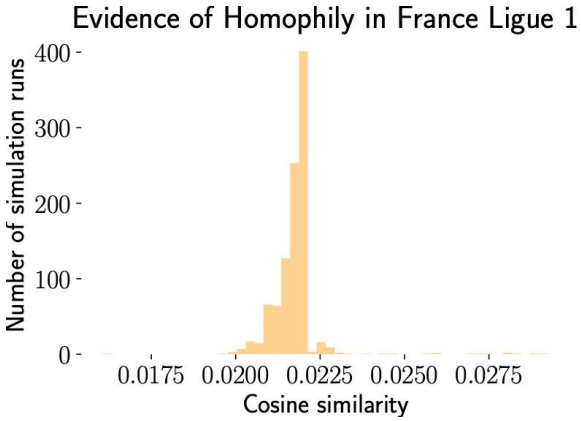

Evidence of Homophily in La Liga 2020
Number of simulation runs
0
200
400
600
0.030
0.035
0.040
0.045
Cosine similarity

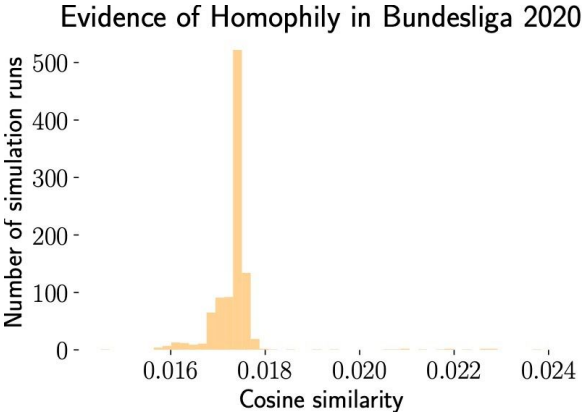

Evidence of Homophily in the Primeira Liga 2020
Number of simulation runs
0
100
200
300
400
0.0225
0.0250
0.0275
0.0300
0.0325
Cosine similarity


Evidence of Homophily in Serie A 2020
Number of simulation runs
0
100
200
300
400
500
0.025
0.030
0.035
0.040
0.045
Cosine similarity


Evidence of Homophily in the Russian Premier League 2020
Number of simulation runs
0
100
200
300
400
0.030
0.035
0.040
Cosine similarity


Evidence of Homophily in Eredivisie 2020
Number of simulation runs
0
100
200
300
400
500
0.006
0.008
0.010
0.012
0.014
Cosine similarity


Evidence of Homophily in France Ligue 1
Number of simulation runs
0
100
200
300
400
0.0175
0.0200
0.0225
0.0250
0.0275
Cosine similarity


Evidence of Homophily in Bundesliga 2020
Number of simulation runs
0
100
200
300
400
500
0.016
0.018
0.020
0.022
0.024
Cosine similarity

References

- BBC, 2013. *Neymar: Barcelona complete £49m signing of Brazil striker*. [Online] Available at: https://www.bbc.co.uk/sport/football/22760770 [Accessed 12 December 2020].
- BBC Sport, 2017. *Wolves Appoint Nuno As New Head Coach*. [online] Available at: https://www.bbc.co.uk/sport/football/40105588 [Accessed 17 December 2020].
- Coates, D., Naidenova, I. & Parshakov, P., 2020. *Transfer Policy and Football Club Performance: Evidence from Network Analysis*. [Online] Available at: https://search.proquest.com/openview/34d7fcf9e00c971ef21fd3db6fcc804b/1?pq-origsite=gscholar&cbl=28340 [Accessed 9 December 2020].
- Easley, D. & Kleinberg, J., 2010. *Networks, Crowds, and Markets: Reasoning About a Highly Connected World.* [Online] Available at: https://www.cs.cornell.edu/home/kleinber/networks-book/ [Accessed 14 December 2020].
- Ertug, G. & Gargiulo, M., 2012. *Does Homophily Affect Performance?*. [Online] Available at: https://flora.insead.edu/fichiersti_wp/inseadwp2012/2012-121.pdf [Accessed 11 December 2020].
- Goal.com, 2018. *Premier League Home Grown Players Rule – How Does It Work?* [online] Available at: https://www.goal.com/en-gb/news/premier-league-home-grown-players-rule-how-does-it-work/1mww3y06t775v1a7c6139l53ji [Accessed 17 December 2020].
- Hegde D., Tumlinson J., 2012 *Can birds of a feather fly together? Evidence for the economic payoffs of ethnic homophily*. Academy of Management Proceedings. Briarcliff Manor, NY 10510: Academy of Management, 2012, 2012(1): 13293.
- Oxford Reference. 2020. *Homophily*. [online] Available at: https://www.oxfordreference.com/view/10.1093/acref/9780191803093.001.0001/acref-9780191803093-e-591 [Accessed 12 December 2020].
- Sky Sports, 2018. *Championship (Sky Sports)*. [online] Available at: <https://www.skysports.com/championship-table/2017> [Accessed 17 December 2020].
- UEFA, 2020. *EUROPEAN LEAGUES & CUPS.* [Online] Available at: https://www.uefa.com/memberassociations/leaguesandcups/ [Accessed 4 December 2020].
- Sports Illustrated, 2019. Americans Abroad 2019-20: USMNT Stars To Watch. [online] Available at:

<https://www.si.com/soccer/2019/07/29/americans-abroad-preview-europe-pulisic-mckennie-weah-sargent-adams> [Accessed 17 December 2020].