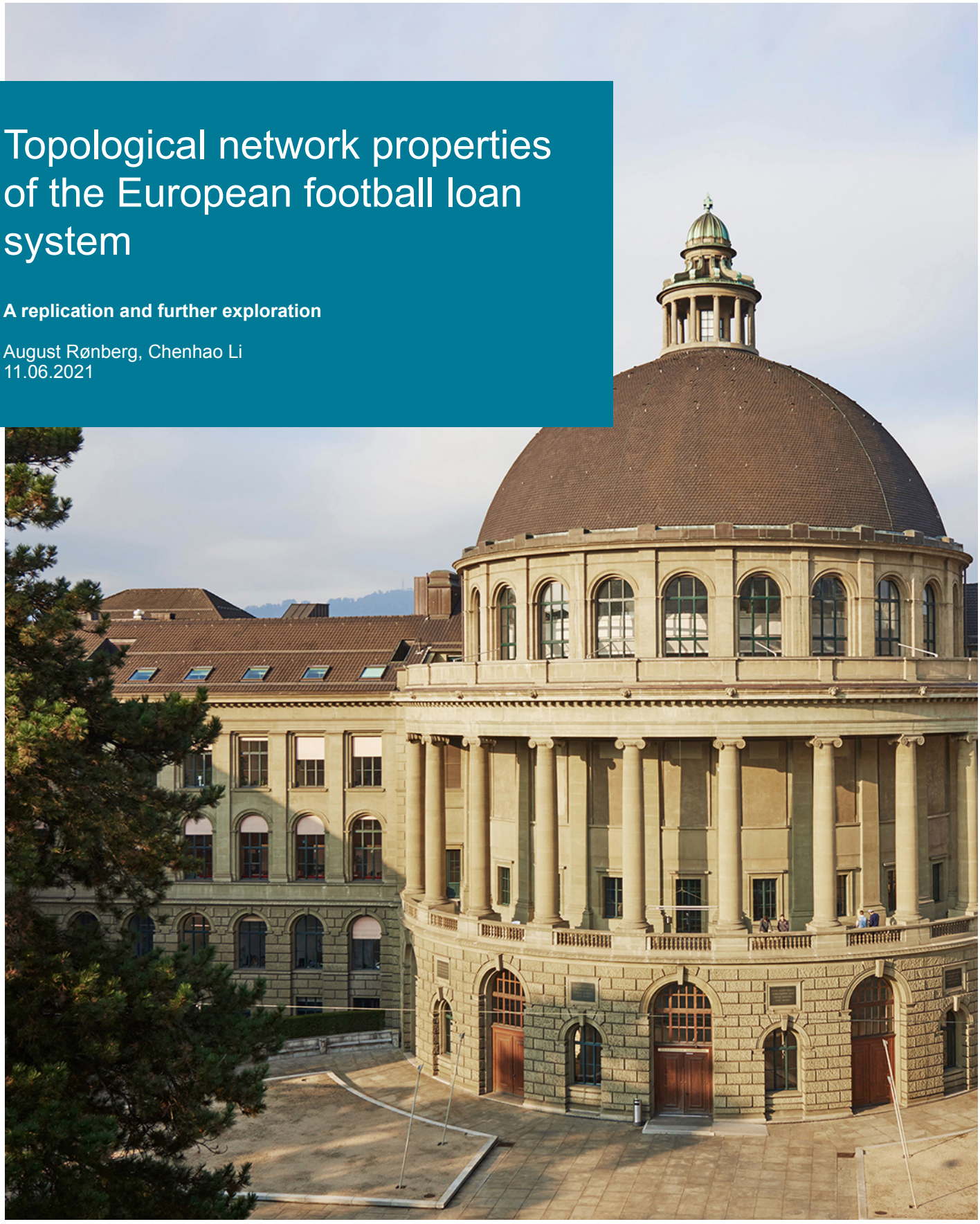
ETH zürich

Topological network properties of the European football loan system

A replication and further exploration

August Rønberg, Chenhao Li
11.06.2021



ETH zürich

Table of contents

1. Main content of the reference work **3**

2. Method replication and data visualization **4**
2.1. Data collection and processing 4
2.2. Data visualization and extension 4
2.2.1. Overview 4
2.2.2. Analysis of difference 5
2.2.3. Graph construction and extension 6

1. Main content of the reference work

Main content
Please refer to the reference work[1] and our slides.

1. Bond, A. J., Widdop, P., & Parnell, D. (2020). Topological network properties of the European football loan system. European Sport Management Quarterly, 20(5), 655-678.

2. Method replication and data visualization

To check the reliability of the interpretation concluded by the reference work, we replicated the data, visualized and analyzed it. We scraped the data while paying attention to certain caveats that the original reference missed, thus basing our replication and own experiments on a more rigorously and accurately collected dataset. Furthermore, to check whether the finding derived in the work could actually be applied to theoretically infer topological properties of the loan network in the future, we also extended the time horizon until last season of the European leagues and made a comparison of the network constructed using the previous data and that of recent years.

2.1. Data collection and processing

Collecting data is the first step of replication. Remember our assumptions:

- Only clubs playing in the top-5 European leagues are considered, i.e., edges representing loan transactions are taken into account when either of the two clubs involved is from the top-5 European leagues
- Permanent transfers, free agent moves, length of the loan, a player returning from loan, or whether a loan deal is cut short are not considered

Note that special attention has been paid to some caveats during the collection of the data. In the first place, the cases not considered in the assumptions as mentioned above have been excluded. Then the duplicated loan records, where a loan transaction is recorded twice if both involving clubs are in the top-5 European leagues, have been removed. Note that the case differs when only one of the clubs is in the top leagues.

It happened that sometimes the names of certain clubs do not unify over the seasons from the webpage of the data resource. For example, Inter Milan is sometimes written as Inter for short and sometimes in Italian as Internazionale. Therefore, specific methods have been applied to deal with this problem where a unique ID of the clubs instead of their names has been sourced to identify a club.

With all caveats well addressed, a scraping code has been developed for general use.

2.2. Data visualization and extension

2.2.1. Overview

The comparison of the reference and our replication has been made. Regardless of the swapped color for some leagues, the same network structure as concluded in the reference work can be observed, where the connection within the Premier League and Serie A presents a much denser structure than the others.

Figure 1 Comparison of network structure between reference and replication. Note the clearly denser edge distribution within Premier League and Serie A.

Global measures

Global measures	reference	replication
Number of nodes	1,105	1,132
Number of edges	5,331	5,777
Density	0.007	0.005
Global clustering coefficient	0.158	0.154

It can be observed that our replication does not match exactly the reference, where more nodes and edges have been included, while the density and the global clustering coefficient is slightly smaller. In the next section, we will analyze the factors that might cause such a gap in global measures.

2.2.2. Analysis of difference

In the original assumption, it was claimed that only the clubs in “Top-5 European leagues” have been taken into account. But this is not clear how this concept is defined. Intuitively, one would consider the case that in each season, the clubs in the top leagues are different due to the promotion and relegation. However, in the main data source of the reference work, soccerway, the data is provided only for clubs that are currently playing in top-5 European leagues, regardless whether they played in these leagues over the past seasons, which means no promotion and relegation have be considered.

What we used for the replication is the data from transfermarkt, which is also mentioned in the reference work as a data resource. On this website, the data is organized in a way that complies with our intuition where clubs in the top leagues vary each season and promotion and relegation have been considered.

This difference of data sources is probably the reason why the measures do not match exactly. And it explains our observation well – if no promotion and relegation have been considered, the same clubs will always appear over the years and the resulting data is therefore more consistent and thus concentrated. And that's why there are fewer nodes and edges but a higher density and global clustering coefficient in the reference work.

People may ask, why not directly use the data from soccerway the same as the reference work did. However, remember the organization of the data provided by soccerway, the clubs that were playing in the top leagues were different at the time when the reference work was done as from today. Thus, it would be relatively involving to retrieve the data of such clubs.

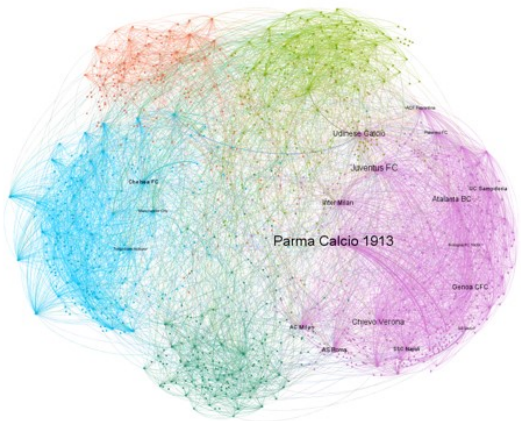
Therefore, an exact replication not possible and not necessary, if in addition the potential updates of records are considered, which happen frequently. Since there is no original data provided, we don't pursue a perfect replication but to get a general sense of how the analysis method could reflect the topological structure of the overall network.

2.2.3. Graph construction and extension

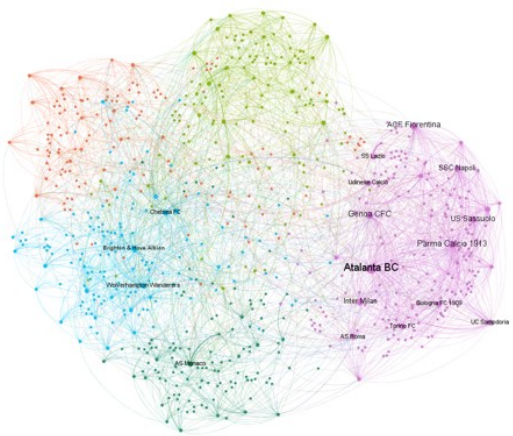
The local measure analysis broke the network down even further and looked to analyze the position clubs take within the network. With all the data in hand, we constructed graphs with Networkx and used Gephi for visualization and analysis. In the following part, topological structures of the network in view of different time spans for each local measure are compared. In particular, the graph on the left-hand side is the visualization of the data over 2010 to 2017, which is exactly the time span considered in the reference work. In addition, we considered the time horizon from year 2018 to 2020, to verify if the topological structure of the network developed by the reference work by the end of 2017 can be used to predict future configurations that have not been considered yet.

The leaders of each measure are labeled with their names and the magnitude of the measure is given by the size of the labels.

Out-degree
It can be seen that Serie A remains to be the most active market and the out-degree leading clubs are still mainly from Italy and England. The leaders remain almost the same but with a slightly different order. Because the difference between the clubs is originally very small and thus small errors can already lead to large change of order.



2010 - 2017



2018 - 2020

Figure 2 Comparison of network structure in view of different time spans based on out-degree.

In-degree

The pattern of in-degree leaders is similar to the one observed in the out-degree setting in the sense that Serie A remains to be the most interconnected loan market, meaning that the Italian clubs not only actively take part in lending but also in borrowing players from other leagues and the same league. And this structure remains relatively stable over the seasons.

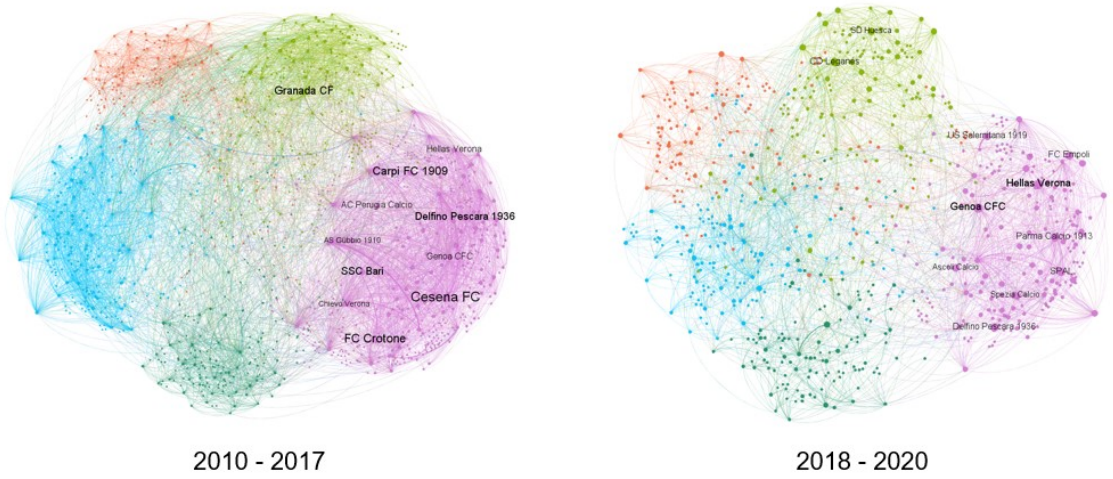


Figure 3 Comparison of network structure in view of different time spans based on in-degree.

Closeness centrality

In the closeness centrality setting, result turns to be unstable, meaning that the list of the clubs that have the closest access to other clubs keeps changing. Therefore, it is not reasonable to simply derive a general conclusion on the closeness centrality leaders without specifying a specific time period.

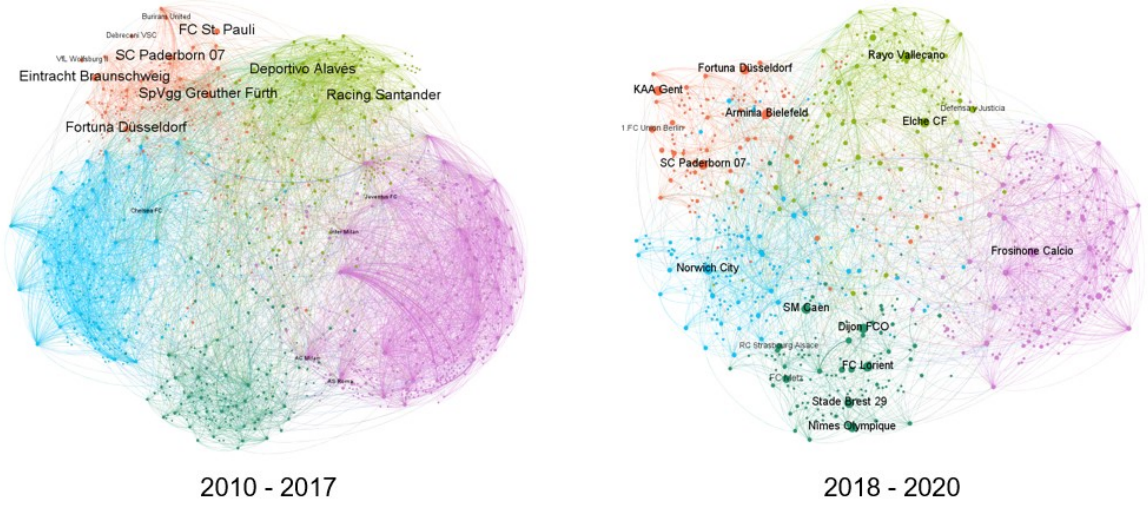


Figure 4 Comparison of network structure in view of different time spans based on closeness centrality.

Betweenness centrality

In betweenness centrality setting, interestingly, the result is again stable and the leaders remain almost the same up to a slightly different order.

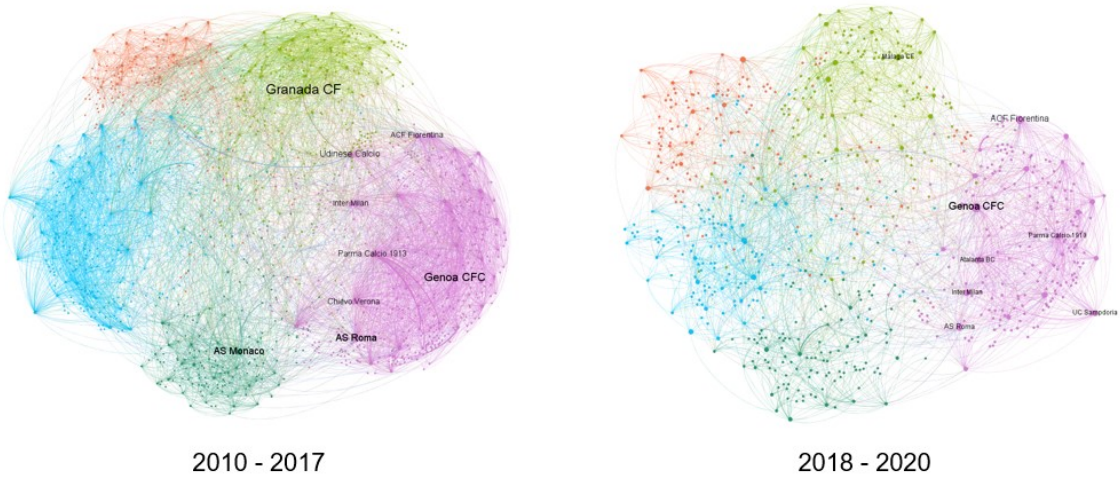


Figure 5 Comparison of network structure in view of different time spans based on betweenness centrality.

3. Predicting loan networks and testing them a CUG test

3.1. Generating graphs

We computed a prediction matrix based on the first 10 seasons which works like this:
The prediction matrix predicts club i loaning a player from club j with probability p=k/10 where k is the number of seasons between 2010 and 2019 where club i loaned a player from club j. This means that if club i never loaned a player from club j in the 10 seasons between 2010 and 2019 the prediction matrix will never generate a graph where club i loans a player from club j and if club i loaned a player from club j in most of the seasons between 2010 and 2019 the prediction matrix will rarely generate a graph where club i doesn't loan a player from club j.
The generator which uses the prediction matrix randomly selects an edge and "rolls a die" based on the probability of that edge to see whether or not to include that edge. This is repeated until the wanted amount of edges in the graph has been reached.

We fix the density of the generated graphs so that all generated graphs match the density of the observed 2020 season. This also means that we are bound to have some isolates in the generated graphs as some new clubs appear in 2020 that haven't been included before.

Finally we calculate relevant metrics on the generated graphs so that we can compare them to the observed graph for season 2020.

3.2. Performing the CUG test

We tested the generated graphs using a conditional uniform graph test also known as a CUG test. The CUG test works by generating a large sample of predicted graphs (we chose to generate 1000 samples), then some relevant metrics are calculated for each of the generated networks, and finally these metrics are compared to the same metrics calculated on the observed 2020 graph.

The relevant metrics in this case are the ones we have already discussed such as in/out degree, triad census, and degree, closeness, and betweenness centralities.

Then we looked at the distribution of the 1000 generated scores for a metric where the y axis is the frequency of a score and the x-axis is the value of the scores, and compared the distribution with the calculated score for the observed 2020 graph. We used the p-value to do this as it tells how well our predicted graphs fit the observed 2020 graph. The p-value is the area under the curve to the right of the score of the observed 2020 graph. If the p-value is very small it means that most of our generated graphs scored lower in that metric than the observed graph. If of the p-value is very large it means that most of our generated graphs scored higher in that metric than our observed 2020 graph.

3.2.1. Interpreting CUG test results for in-degree

It can be observed that 50/454 (11%) clubs that loaned players from other clubs in 2020 never loaned(or loaned as many players) from other clubs in any of the generated graphs. These are likely exclusively clubs that never loaned players from other clubs before season 2020. We also see 43/454 (9%) of clubs that never received loans or as many in 2020 as in the generated graphs. These are likely clubs which

received loans once or more in the previous seasons but never in 2020. Finally if we look at the remaining 361/454 (80%) of the clubs which received loans in previous season(s) and in 2020, we see that 308/454 (68%) of all the clubs were well approximated (when considering a double sided test with significance level of a=0.05) which is 305/358 (85%) of the clubs which received a loan in at least one seasons 2010-2019 and in 2020.

3.2.2. Interpreting CUG test results for out-degree

Here it can be observed that 28/454 (6%) clubs that loaned out players to other clubs in 2020 never loaned out(or loaned out as many players) in any of the generated graphs. These are likely exclusively clubs that never loaned out players to other clubs before season 2020. We also see 297/454 (65%) of clubs that never loaned out or loaned out as many players in 2020 as in the generated graphs. These are likely clubs which loaned out players once or more in the previous seasons but never in 2020. Finally if we look at the remaining 129/454 (28%) of the clubs which loaned out players in previous season(s) and in 2020, we see that 102/454 (22%) of all the clubs were well approximated (when considering a double sided test with significance level of a=0.05). This shows that it is interestingly far more difficult to predict outgoing loans based on previous seasons than it is to predict incoming loans based on previous seasons.

3.2.3. Interpreting CUG test results for triad census

Numbers 1-16 represent the 16 different triad census structures of the MAN convention:

1, 5, 8-16: No isolates in observed but several in generated graphs was to be expected. This is because some of the clubs that are a part of the loan system in season 2020 made no loan transactions that were recorded in the 2010-2019 transactions which our model makes predictions based on. Therefore they are always predicted to have a total degree of 0 and become isolates in the generated graphs. It also means that there are far more empty triads in the generated graphs.

2, 4: Since the generator picks at random based on the prediction matrix we see that teams that have traded through many of the previous years end up having loan transactions having a larger weight than observed in season 2020. This is despite the fact that those teams might have simply traded few players consistently instead of many over the previous seasons.

3, 7: Reciprocity estimated really well, which is because reciprocity is also well represented in prediction matrix. Reciprocity being captured well hints at the stability of the strategic alliances.

3.2.4. Interpreting CUG test results for degree centrality

It can be observed that 63/454(14%) clubs that loaned out/received players from other clubs in 2020 never loaned out/received players(or loaned out/received players as many players) in any of the generated graphs. These are likely exclusively clubs that never loaned out/received players or never loaned out/ received players from other clubs before season 2020. We also see that 5/454 (1%) of clubs that never loaned out/received players or as many in 2020 as in the generated graphs. These are likely clubs which loaned out/received players once or more in the previous seasons but never in 2020. Finally if we look at the remaining 385/454 (85%) of the clubs which loaned out/received players in previous season(s) and in 2020, we see that 314/454 (69%) of all the clubs were well approximated (when considering a double sided test with significance level of a=0.05) which is 315/385 (82%) of the clubs which loaned out/received players in at least one of seasons 2010-2019 and in 2020. ("loaned out/received" means either loaned out or received or both)

3.2.5. Interpreting CUG test results for closeness centrality

Here it can be observed that 97/454 (21%) of clubs had higher closeness centrality scores in 2020 than in any of the generated graphs. These are likely clubs that never loaned out players (or just loaned to few not well connected clubs) before 2020. We also see that 297/454 (65%) of clubs had a higher closeness centrality score in all of the generated graphs than in 2020. These are likely clubs that never loaned out players (or loaned out players to few not well connected clubs) in 2020 while having loaned out players in one or several of previous years. This is not very unlikely as the loan system must have changed at least a little over the 10 years from 2010 to 2019 and since the probability matrix prob basically is those 10 years compressed into one matrix which means that the graph generator is likely not to respect that some pairs of ties never happened in the same season (negative edge correlation). On the other hand the model also does not respect if in every season where club a loans from club b, club b also loans from club a (positive edge correlation). In other words the model does not reflect edge correlations but rather edge probabilities. This results in that the model generates graphs that are better connected than what can realistically be expected.

Finally if we look at the remaining 60/454 (13%) clubs we see that only 56/454 (12%) of the clubs were well approximated (when considering a double sided test with significance level of a=0.05) with respect to their closeness centrality. This is very low because the graph generator only is conditioned of the density of the observed graph which doesn't capture more complex structures and interdependencies that affect the value of centrality indices such as closeness centrality.

3.2.6. Interpreting CUG test results for betweenness centrality

It can be observed that 7/454 (1%) of clubs had higher betweenness centrality scores in 2020 than in any of the generated graphs. These are likely clubs that never loaned out players and received players before 2020. We also see that 341/454 (75%) of clubs had a higher betweenness centrality score in all of the generated graphs than in 2020. These are likely clubs that never loaned out and received players(or loaned out players to few not well connected clubs and received players from few not well connected clubs) in 2020 while having loaned out and received players in one or several of previous years. This is not very unlikely as the loan system must have changed at least a little over the 10 years from 2010 to 2019 and since the probability matrix prob basically is those 10 years compressed into one matrix which means that the graph generator is likely not to respect that some pairs of ties never happened in the same season (negative edge correlation). On the other hand the model also does not respect if in every season where club a loans from club b, club b also loans from club a (positive edge correlation). In other words the model does not reflect edge correlations but rather edge probabilities. This results in that the model generates graphs that are better connected than what can realistically be expected.

Finally if we look at the remaining 107/454 (24%) clubs we see that only 83/454 (18%) of the clubs were well approximated (when considering a double sided test with significance level of a=0.05) with respect to their closeness centrality. This is very low because the graph generator only is conditioned of the density of the observed graph which doesn't capture more complex structures and interdependencies that affect the value of centrality indices such as betweenness centrality.

Appendices

1. data_scraping.py - script for scraping data from transfermarkt.com.

2. json - folder containing .json files scraped using data_scraping.py.

3. graph_generation.py - script for reformatting .json files into .gexf files which were used for graph visualization in Gephi.

4. gexf - folder containing .gexf files created using graph_generation.py.

5. graph_generation2.ipynb - script for reformatting .json files into .graphml files which was used for graph generation in Analysis CUG.R.

6. Analysis CUG.R - script for graph generation and CUG testing.

7. Presentation - pdf containing slides from presentation.