ARPALIANGEAS et BOUTELOUP 17/03/2024

Graph Analysis Project:
Study of a network formation mechanism

In this project, you will investigate a dyadic mechanism driving the formation of a real network. This mechanism can be related to assortativity or dissortativity on a node attribute (similar or dissimilar individuals tend to be connected), to the sociability or popularity of a node attribute (some individuals connect to more people or receive more connections), or to another network variable (individuals connected through another network tend to be connected). Please leave the instructions in italic. Attach your code as a RmD file.

1. Problem Statement

Describe the network you are studying. What mechanism are you interested in and why is it interesting or relevant to study it in this network? (max 200 words)

The network under examination represents competitive interactions among players within a rugby team, where connections (edges) signify direct competition between players occupying the same position on the field. For the graph structure, it means an edge exists if two players have played the same position in the same team. And each player (node) has a couple of attributes like his preferred position. This analysis is pivotal for comprehending how certain factors influence the structure of the competitive network and, consequently, the team's effectiveness and performance.

How players choose with who they will compete ? Players are rational and will then not compete for a position they don’t like. They will not also compete for a position secured by someone else. By secured we mean that a player is already essential for this position. The mechanism we are then interested in is an homophilic selection based on the preferred position. (As a recall there is 15 players on the field for 10 different kind of positions with each some particularities).

Studying this homophilic selection mechanism is crucial as it helps reveal any patterns in players' preferences for particular positions. We aim to discern whether there's a correlation between players' preferred positions and the roles they ultimately fulfil on the field. Although we lack precise information on each player's assigned position due to its variability. But we can do the following hypothesis : if a player is surrounded by others who favour the same position as him, then it might mean that he is on the right place in team because he competes with people with the same preferences. That would also mean that the coach does not take any incoherent tactical decision.

It's noteworthy that players performing in their favoured positions are often more motivated and inclined to deliver their best performance. This underscores the significance of understanding the team's dynamics to gauge its overall effectiveness and performance on the field.

2. Expectations

Formulate expectations about the causes and consequences of this mechanism. Why do you expect this mechanism to occur? What network characteristics do you expect as a result from this mechanism? (max 300 words).

The underlying cause of this mechanism lies in the natural affinities players have towards competing for their preferred positions. Players are naturally drawn to positions that align with their skills, interests, and strengths. Additionally, the competitive spirit inherent in sports motivates players to excel in their chosen positions, driving them to seek out opportunities to showcase their abilities. Consequently, players are more inclined to compete for positions they are passionate about, even if it means facing tougher competition.

As a result of this mechanism, we anticipate the emergence of distinct clusters within the team, characterized by players who frequently vie for the same positions. These clusters represent cohesive groups of players who share common objectives and experiences, fostering a sense of camaraderie and solidarity among teammates. Within these clusters, we expect to observe high levels of transitive relationships, where players who compete against the same opponents are more likely to be connected. This reflects the cohesive nature of the team's competitive network.

Moreover, we anticipate a high assortativity coefficient, indicating a tendency for players with similar preferred positions to compete against each other more frequently. This phenomenon, known as homophily, further reinforces the clustering of players based on their position preferences.

Furthermore, we anticipate variability in the degree distribution across clusters. Positions that require specialized skills or carry greater responsibilities may attract fewer players, leading to smaller clusters for these positions compared to others. This diversity in cluster sizes highlights the nuanced dynamics at play within the team's competitive structure.

3. Research design

Explain how to define a CUG test to find evidence for the mechanism you are studying in any given graph. Clearly specify the reference model you chose and define your test statistic with equations. (max 300 words)

The CUG test involves generating multiple random graphs to create a distribution of expected network characteristics. By comparing these simulated distributions with the actual network's features, we can assess whether the observed characteristics are statistically significant or occur by random chance.

To define a CUG (Cluster-Uniformity-Graph) test for detecting evidence of the homophilic selection mechanism in any given graph, we first specify the reference model. Our reference is the Erdős–Rényi model, commonly used as a null model in network analysis. In this model, a graph is chosen uniformly from the collection of all graphs which have n nodes and M edges. We write the model as follow : $G(n,m)$.

The model $G(n,m)$ assigns a uniform probability to all graphs g = (v,e) such that $|e|$ = m and a null probability otherwise

$$\boldsymbol{P(G = g|m) = \begin{cases} 1/\binom{n(n-1)/2}{m} & if\ |e| = m \\ 0 \end{cases}}$$

Put differently, within the entire population of networks that can be generated, any network with n nodes and m edges has an equal chance of being selected.

The test statistic we use is the Edge Index (EI), which measures the proportion of connections between nodes with the same attribute. In our case, the attribute of interest is the preferred position. The EI index is calculated as the ratio of the number of connections between players with the same preferred position to the total number of edges in the network.

$$\textbf{EI} = \frac{\textbf{Number of edges between nodes with the same preferred position}}{\textbf{Total number of edges in the network}}$$

Hypotheses:

- $H0$: our statistic is less prevalent or equally present than in a random network
- $H1$: our statistic is more prevalent

To conduct the CUG test, we generate a large number random graphs based on ER model (e.g., 10000) with the same number of nodes and edges as the observed network. For each random graph, we compute the EI index. We then compare the observed EI index in the real network to the distribution of EI indices obtained from the random graphs. An empirical p-value is calculated as the proportion of random graphs with EI indices greater than or equal to the observed EI index.

A significant p-value (e.g., below a chosen threshold, such as 0.05) suggests that the observed network has more connections between nodes with the same preferred position attribute than expected by random chance alone. This finding provides evidence supporting the operation of the homophilic selection mechanism in the network.

Explain how to define a QAP test to find evidence for the mechanism you are studying in any given graph. Clearly define all your statistical variables with equations. (max 300 words)

The QAP (Quadratic Assignment Procedure) test is employed to examine the presence of structural patterns or associations between nodes in a network. Similar to the CUG test, it involves comparing observed network characteristics with those expected under a null model. Here's how to define a QAP test for detecting evidence of the homophilic selection mechanism:

Reference Model: Our reference model for the QAP test is the random permutation model. In this model, the positions of nodes and their attributes are randomly shuffled while preserving the network's structure. The random permutation model allows us to generate null distributions of network characteristics that account for network structure but randomize node attributes.

Test Statistic: The test statistic used in the QAP test measures the strength of the association between nodes' attributes and the existence of a tie. For the homophilic selection mechanism, we can use the Mantel statistic based on Pearson's product-moment correlation. This statistic quantifies the linear relationship between the presence of connections and the similarity of node attributes. Alternatively, a logistic regression model can be employed to estimate the probability of a tie between nodes based on their attributes. The logistic regression model captures the likelihood of a connection between nodes given their attributes.

The Mantel statistic measures the correlation between two distance matrices. It is calculated as the Pearson correlation coefficient between the entries of the two distance matrices. Mathematically, it can be represented as:

Mantel(X,Y) = $$\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_{ij}y_{ij}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}x_{ij}^2}\times\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}y_{ij}^2}}$$

Where:

- X and Y are the two distance matrices.
- wij represents the weight associated with the pair of observations i and j.
- xij and yij are the entries of the distance matrices X and Y respectively.

In the logistic regression we are interested in computing $g(X) = \theta 0 + \theta 1 Z + \varepsilon$

Hypotheses:

- $H0$: There is no significant association between node attributes and network connections.
- $H1$: There is a significant association between node attributes and network connections.

Procedure: To conduct the QAP test, we first generate a large number of random permutations of the network while preserving its structure. For each permutation, we calculate the test statistic (e.g., Mantel statistic or logistic regression estimate) between node attributes and network connections. We then compare the observed test statistic in the real network to the distribution of test statistics obtained from the random permutations under H0. An empirical p-value is calculated as the proportion of random permutations with test statistics greater than or equal to the observed test statistic.

Interpretation: A significant p-value (e.g., below a chosen threshold) indicates that the observed network exhibits a stronger association between node attributes and network connections than expected by

random chance alone. This finding would provide evidence supporting the presence of the homophilic selection mechanism in the network, where nodes with similar attributes are more likely to be connected.

4. Data collection

Describe your data. Which real network did you select, how did you collect and store the data? Provide a graph visualization and three descriptive measures of the network that are useful for your analyses. (max 300 words)

The data for our analysis consists of the composition of the Toulouse ElectroGaz Club (TEC) rugby team during the 2023-2024 competition season. For each match, we recorded the starting positions of the 15 players. Players occupying the same position as starters in the same team (there is two teams) during the season are considered connected or in competition with each other. This data was collected manually. We also asked questions to players individually to get their age, rugby's experience, preferred position, and first ever position.

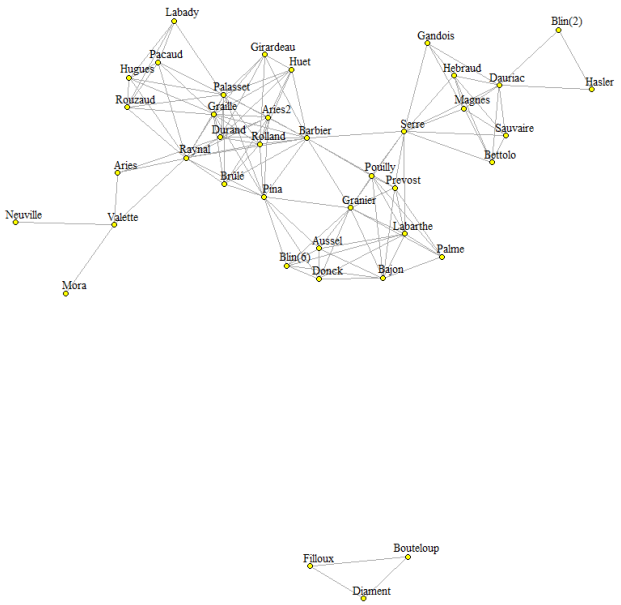
Our dataset comprises data from two teams: the first team and the reserve team, spanning across two distinct periods. It's noteworthy that certain players may have participated in both teams, resulting in potentially higher connectivity for those players. Additionally, the dataset covers both outward and return phases of the competition. During the return phase, we anticipate a potential reduction in the number of players participating, as teams typically stabilize their lineup if they achieve satisfactory results. This expectation suggests that the network structures may vary between the outward and return phases, reflecting the evolving dynamics within the team. Therefore, we anticipate observing differences in graph characteristics based on the phase of the competition and the team's performance.

In order to visualize our graph effectively, we produced five visualizations corresponding to different team and period combinations.

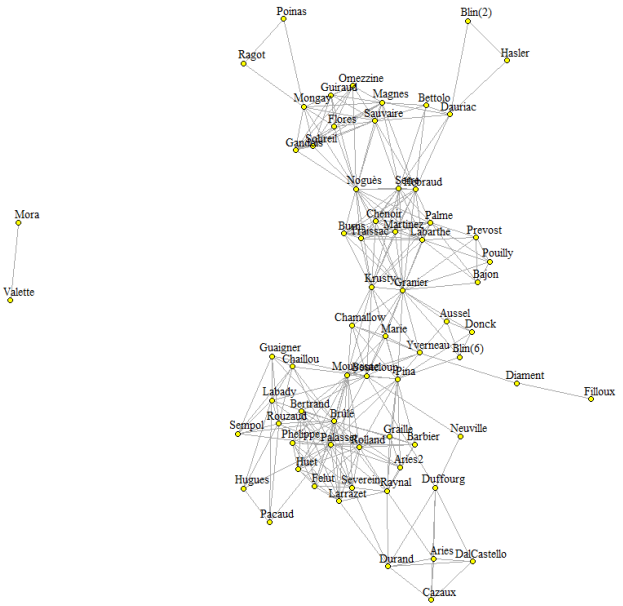
Regarding descriptive measures of the network, we focused on metrics not directly aligned with our primary interest in preferred positions but which provide valuable insights into overall network characteristics. First, the Average local efficiency measures the efficiency of information exchange within local neighborhoods of nodes in the network. It quantifies how well information can be transferred within clusters or communities of nodes. A higher average local efficiency suggests that local neighborhoods are well connected and can efficiently exchange information. In this case, the average local efficiency value is approximately 0.83, indicating that within local neighborhoods, the rugby network demonstrates relatively high efficiency in information exchange. Secondly, the Transitivity measures the tendency of nodes in the network to form clusters or triangles. It quantifies how often nodes that are connected to the same node are also connected to each other. A higher transitivity value suggests a higher likelihood of clustering, indicating that nodes tend to form tightly interconnected groups or communities. In this case, the transitivity value is approximately 0.63, indicating that the rugby network exhibits a moderate level of clustering, with nodes forming triangles or clusters at a relatively frequent rate. Finally, the Mean Distance in the rugby network is calculated to be approximately 3.94. This metric represents the average shortest path length between all pairs of nodes in the network. It provides insight into the overall connectedness and accessibility within the network, with lower values indicating shorter average distances and potentially greater efficiency in information flow or interaction between nodes. Additionally, the maximum shortest path length in the network is found to be 10. This measure identifies the longest shortest path between any two nodes in the network, highlighting the maximum distance required to traverse the network from one node to another. Understanding these distance metrics helps assess the network's structure and the ease of communication or interaction between its components.

Note: We are interested in the network corresponding to the two teams during the second period.

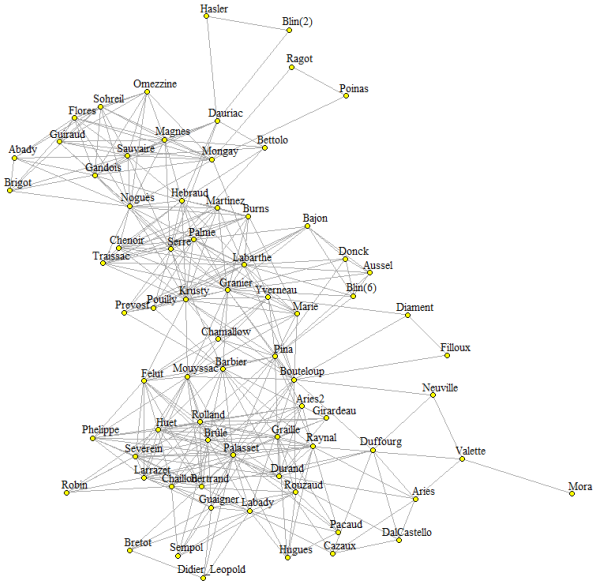
Graphique du réseau de rugby - Période 1 et 2 - Equipe 1



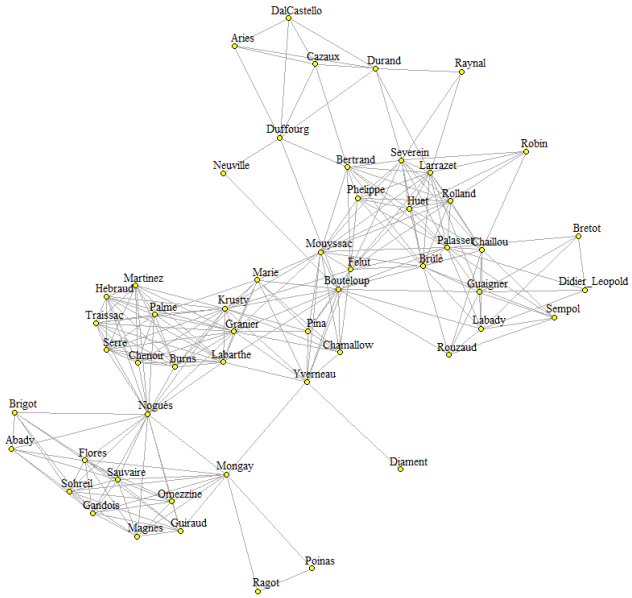
Graphique du réseau de rugby - Période 1 - Equipe 1 et 2



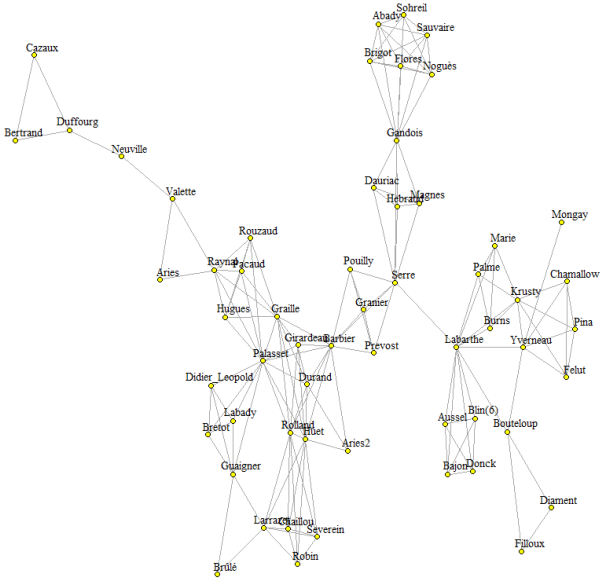
Graphique du réseau de rugby - Période 1 et 2 - Equipe 1 et 2



Graphique du réseau de rugby - Période 1 et 2 - Equipe 2

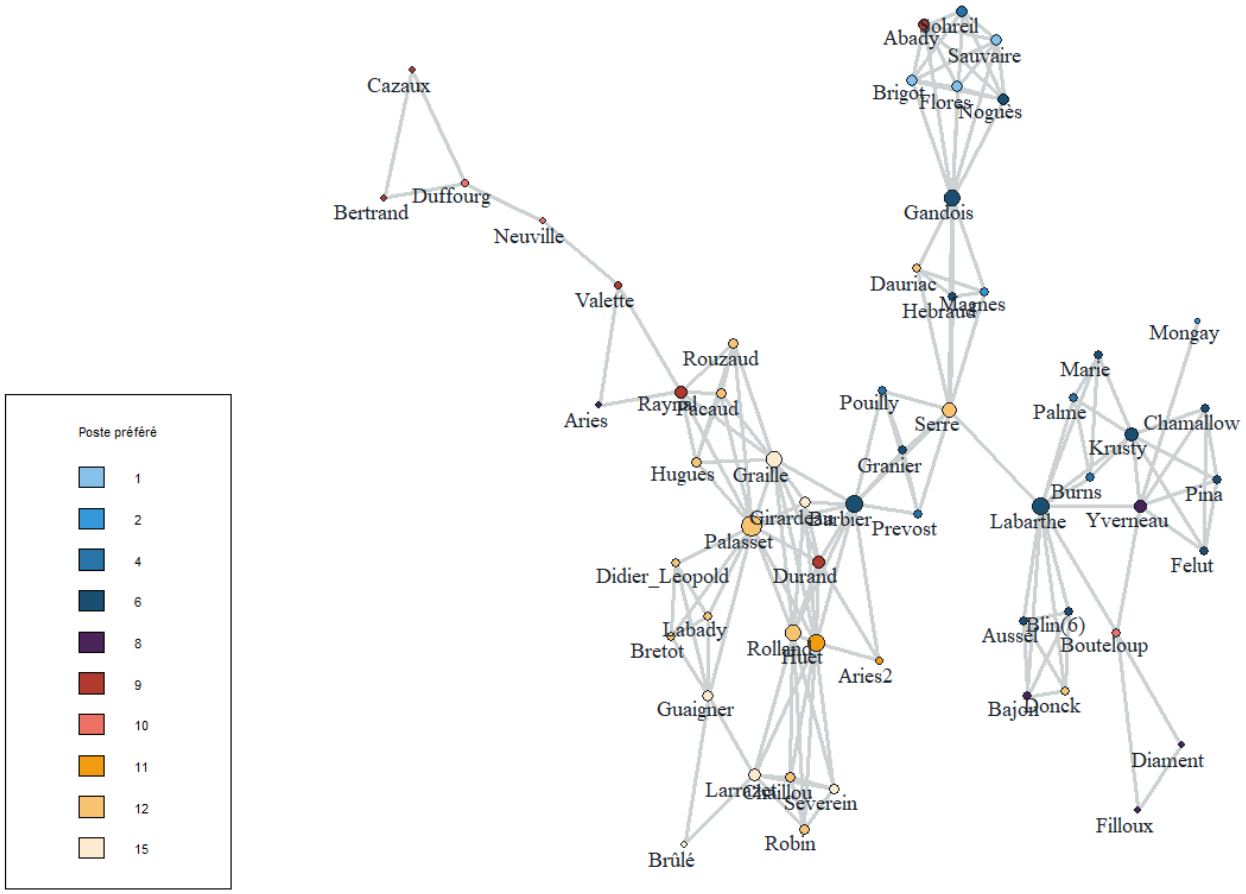


Graphique du réseau de rugby - Période 2 - Equipe 1 et 2



Finally, the visualization we prefer is the following one :

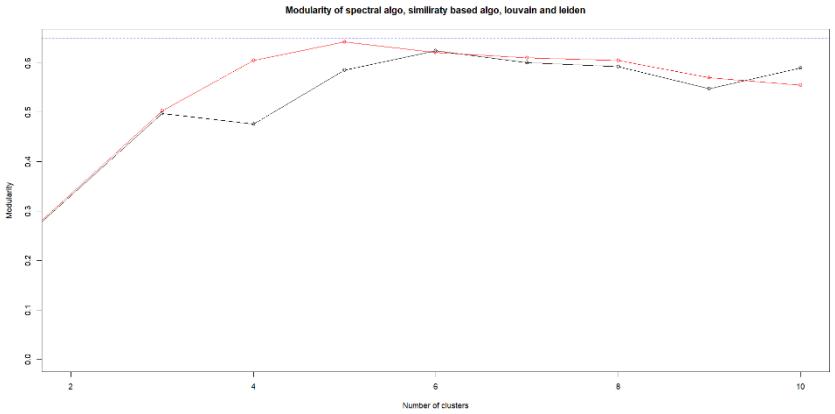
Graphique du réseau de rugby - Période 2 - selon le poste préféré, taille des noeuds proportionnelle à leur degré



5. Exploration and Analysis

Try identifying communities in your graph, using the algorithm of your choice. Explain your choice, describe the communities you found and provide a graph visualization of the communities. (max 300 words)

We explored various clustering algorithms to identify communities within our rugby team network, focusing on similarity-based clustering with the "ward.D2" method, spectral clustering, and the Louvain algorithm. To evaluate the quality of the clusters produced by each algorithm, we calculated the modularity indicator for the top 10 clusters generated by these methods. The resulting graph illustrates the modularity scores for each algorithm and the number of clusters considered. The modularity score tells us how well the network is partitioned into communities. A positive modularity indicates that the network has more within-community edges than expected by chance, suggesting a good community structure. Knowing that we want to select the algorithm that maximize this score.

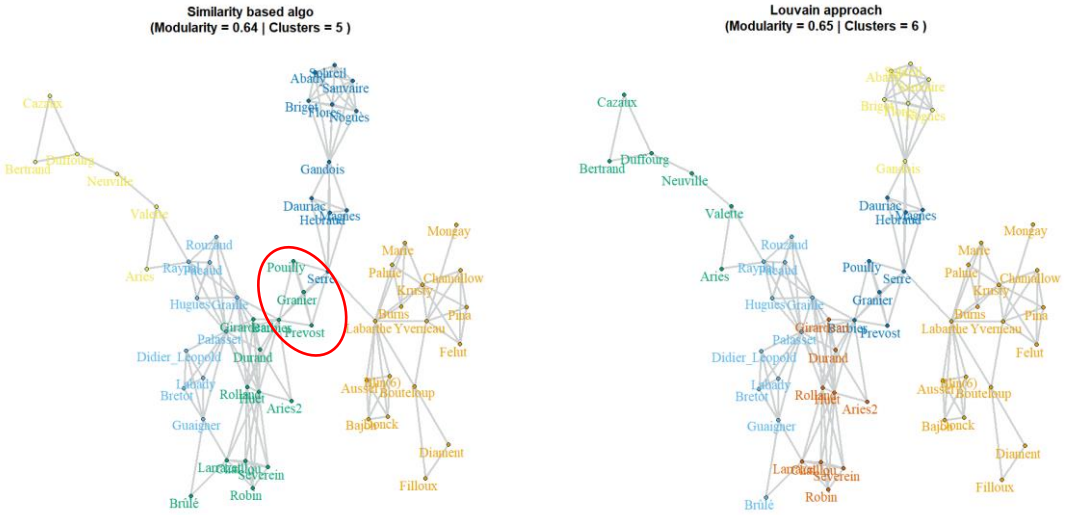


Red : similarity based algorithm

Black : spectral clustering

Blue : Louvain algorithm

Among the algorithms tested, the Louvain algorithm and the similarity-based clustering emerged as the most promising. The Louvain algorithm yielded 6 distinct communities, while the best fit with the similarity-based clustering identified 5 cohesive groups. Visual representations of these results are provided for further insights.



Based on our knowledge of the team dynamics and individual players, we determined that the Louvain algorithm is the most relevant for our analysis. The 6 communities identified align with the following player roles:

- Light orange for the flankers and number 8, with only one exception (players 6, 7, and 8)
- Yellow for the props of team 2 (players 1 and 3)
- Dark blue for the forwards of team 1 (players 1, 2, 3, 4 and 5)
- Dark orange for the wings and fullbacks (players 11, 14, and 15)
- Light blue for the versatile backs and centre (players 12 and 13)
- Green for the scrum-halves and fly-halves (players 9 and 10)

Does the distribution of individuals in the communities match your expectations regarding the mechanism you are studying? You can choose the statistic and/or visualization that best shows this. (300 words)

Community	Preferred Position 1	2	4	6	8	9	10	11	12	15
1	0.0	5.9	11.8	47.1	23.5	0.0	5.9	0.0	5.9	0.0
2	0.0	0.0	0.0	0.0	0.0	9.1	0.0	0.0	63.6	27.3
3	0.0	0.0	0.0	0.0	16.7	50.0	33.3	0.0	0.0	0.0
4	42.9	0.0	14.3	28.6	0.0	14.3	0.0	0.0	0.0	0.0
5	0.0	12.5	25.0	37.5	0.0	0.0	0.0	0.0	25.0	0.0
6	0.0	0.0	0.0	0.0	0.0	11.1	0.0	22.2	33.3	33.3

Preferred Position	Community 1	2	3	4	5	6
1	0.0	0.0	0.0	100.0	0.0	0.0
2	50.0	0.0	0.0	0.0	50.0	0.0
4	40.0	0.0	0.0	20.0	40.0	0.0
6	61.5	0.0	0.0	15.4	23.1	0.0
8	80.0	0.0	20.0	0.0	0.0	0.0
9	0.0	16.7	50.0	16.7	0.0	16.7
10	33.3	0.0	66.7	0.0	0.0	0.0
11	0.0	0.0	0.0	0.0	0.0	100.0
12	7.7	53.8	0.0	0.0	15.4	23.1
15	0.0	50.0	0.0	0.0	0.0	50.0

We can analyze the distribution of individuals across communities by examining the cross table, which depicts the relationship between preferred positions and communities. Upon closer inspection, it becomes evident that certain positions are predominantly represented within specific communities. This observation aligns with our expectations regarding the homophilic selection mechanism under study.

A noteworthy finding emerges from the cross table: the largest representation in community 1 corresponds to position 6. Similarly, upon examining the distribution, we observe that position 6 is most prevalent in community 1. This congruence between the two tables reinforces our hypothesis that players tend to cluster with others who share their preferred position.

However, there is a notable exception: community 5, depicted in dark blue. Here, we observe a diverse representation of preferred positions, these player facilitate connections between the yellow, orange, and other communities. This versatility suggests that players within community 5 may possess positions in the team different from their preferred ones.

By attributing each cluster the most represented position as the primary one and calculating the proportion of players correctly allocated to their respective clusters based on their preferred positions, we find that 46% belong to the correct cluster. It's a satisfying result counting that there are only 6 clusters for 10 different positions. And if we use the reorganized version of the attribute 'preferred position' (as explain in the appendix) then it it becomes 60% good belonging for 6 clusters and 6 positions.

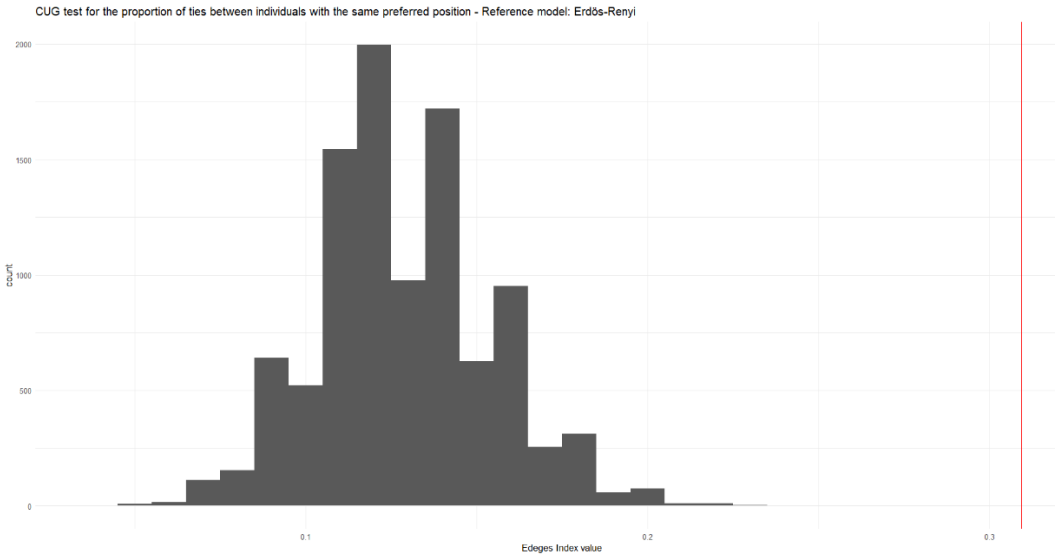
In summary, the distribution of individuals across communities largely corresponds to our expectations regarding the mechanism under study. The prevalence of specific positions within certain clusters and the versatility observed in others support the notion of homophily in team composition, where players tend to gravitate toward others with similar playing preferences.

Perform the CUG test and the QAP test previously described. Report all test statistics and empirical p-values. (max 300 words)

About the CUG test :

The outcome of the Edges Index in the original network is 0.3.

We applied the methodology described in question 3 to 10,000 random graphs, each consisting of 58 nodes and 152 edges. The resulting distribution of the network's characteristics is as follows:



We obtain an empirical p-value of 0, indicating strong evidence against the null hypothesis. Consequently, we reject the null hypothesis and conclude that our statistic is statistically significant. Specifically, the observed number of edges between nodes with the same preferred position is significantly higher than expected at random. This finding provides compelling evidence for a clear tendency of individuals to connect preferentially with others who share the same preferred position, supporting the hypothesis of homophilic selection based on preferred position in the network.

If we use the reorganized version of the attribute ‘preferred position’, results are identical.

About the QAP test :

To conduct the QAP test, we utilize the reorganized version of the attribute 'preferred position'. This reorganization facilitates the construction of a similarity matrix based on the assumption that each of the six different positions carries an equal distance from one another in terms of their characteristics. This similarity matrix assigns values ranging from -1 to 1, where -1 indicates maximum dissimilarity between players' preferred positions, while 1 indicates identical preferences.

A logistic regression model was employed to investigate the relationship between network connectivity in a rugby team and the similarity in preferred positions among players. The NetLogit model estimated the coefficients for the intercept and preferred positions, represented as attributes x0 and x1. The results indicate that while the coefficient for preferred positions (x1) was statistically significant ($p < 0.05$), the coefficient for the intercept (x0) was not. The coefficient for preferred positions was estimated at 1.62, showing a significant positive association between the presence of edges in the network and the similarity

in preferred positions. This suggests that players with similar preferred positions are more likely to be connected in the network.

A Mantel test based on Pearson's product-moment correlation was conducted to further explore the relationship between network structure and preferred positions. The Mantel statistic (r) was calculated as 0.18, indicating a positive correlation between the two matrices. The associated p-value was determined to be lower than 0.001, demonstrating statistical significance. Additionally, the observed Mantel statistic fell well above the 99th percentile of the distribution of permuted statistics, providing further evidence of a significant association between network connectivity and preferred positions.

Conclusion: The results from both the logistic regression and Mantel test provide robust evidence supporting the presence of a homophilic selection mechanism within the rugby team. Players with similar preferred positions exhibit a higher likelihood of being connected in the network, indicating a tendency towards homogeneity in network connectivity based on positional preferences. This insight sheds light on the underlying dynamics of team formation and player interactions within the rugby team.

6. Interpretation and conclusions

Summarize what we learn from your community detection and statistical tests. Are they all in line with your expectations? (max 300 words)

Our community detection analysis revealed that the detected communities align closely with player attribute 'preferred positions' within the rugby team. This observation suggests that players with similar characteristics tend to cluster together within the network, consistent with our expectations of homophilic selection in team composition.

Furthermore, both the CUG (Cluster-Uniformity-Graph) test and the QAP (Quadratic Assignment Procedure) test provided significant results. The CUG test indicated a higher-than-expected proportion of connections between players sharing the same preferred position, reinforcing the notion of homophily in the team structure. Similarly, the QAP test, utilizing the Mantel statistic based on Pearson's correlation coefficient as well as the logistic regression, demonstrated a significant association between node attributes (preferred positions) and network connections.

Overall, these findings are in line with our expectations regarding the mechanism of homophilic selection in team formation. The significant results from both the community detection analysis and the statistical tests provide robust evidence supporting the hypothesis that players with similar attributes, such as preferred positions, are more likely to interact and form connections within the rugby network.

In conclusion, our community detection analysis and statistical tests not only confirm the presence of homophily in the rugby team but also validate the effectiveness of our analytical approach in uncovering underlying patterns and dynamics within the network.

Discuss the potential value of performing a Multiple QAP regression in your case (without doing it). (max 200 words).

Performing a Multiple QAP regression could offer valuable insights in our case by allowing us to examine the combined effects of multiple node attributes on network connections while accounting for the network's structure. By including additional predictor variables such as player age, experience, or playing position alongside preferred position, we can assess how different attributes collectively influence the formation of connections in the rugby team network. This approach would provide a more comprehensive understanding of the factors driving network formation and could reveal nuanced relationships between player characteristics and network topology. Additionally, conducting a Multiple QAP regression would enable us to control for confounding variables and better elucidate the unique contributions of each attribute to the observed network patterns, enhancing the depth and richness of our analysis.

In our case it would be as follow :

$$g(X) = \theta 0 + \theta 1 Z1 + \theta 2 Z2 + \theta 3 Z3 + \varepsilon$$

Where:

- g(X) represents the observed matrix of network connections.

- Z1, Z2, and Z3 are similarity matrices representing node attributes such as preferred position, first-ever position, and experience.
- $\theta 0, \theta 1, \theta 2$, and $\theta 3$ are the regression coefficients.
- ϵ represents the error term.

By controlling for variables such as first-ever position and experience, we can assess the unique contribution of preferred position to network connections.

Discuss the limitations of this study and identify possible ways to improve or enrich your study. Could this type of analysis be used to provide insight into or answer answering real-life problems? (max 200 words)

Although this study offers valuable insights into the rugby team's network dynamics, it is constrained by certain limitations. Primarily, the reliance on self-reported player attributes introduces the possibility of bias or inaccuracies. Moreover, the analysis focuses solely on a restricted set of attributes, such as preferred position and experience, neglecting other potentially significant factors like player skill level or team dynamics. Lastly, the manual data collection process, involving scrutiny of each game to compile the necessary cross table, leaves room for oversight or errors, possibly leading to incomplete or inaccurately recorded data.

To improve the study, data collection methods could be refined to include more objective measures of player attributes, and a broader range of attributes could be explored. Additionally, incorporating longitudinal data to track changes in network dynamics over time could provide a more comprehensive understanding of team dynamics.

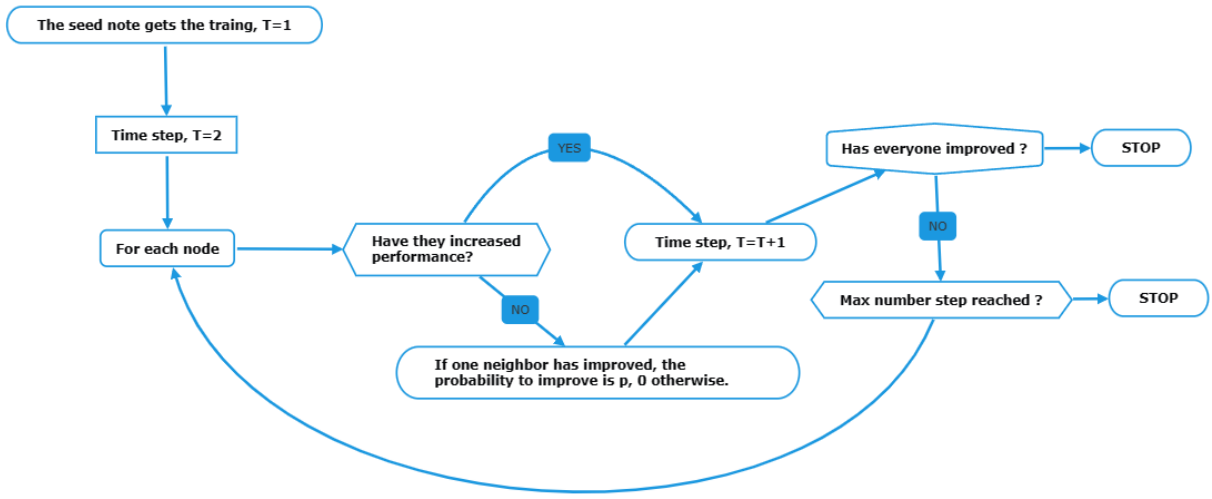
Despite these limitations, this type of analysis has the potential to offer valuable insights into real-life problems, such as optimizing team composition or identifying influential players within a network. By understanding the factors that drive network formation and evolution, organizations can make informed decisions to enhance team performance and collaboration.

Study of a contagion process in a network

We are interested in the spreading of knowledge. The situation is as follows: the TEC rugby club has the opportunity to send one of its players for an intensive training program with professionals in the sport. The potential benefits of such a program for the chosen player are immense. Their performance, skills, and knowledge will significantly improve. However, we must not forget what we are studying here: competition. This means that players competing with the chosen player will have to put in extra effort to catch up, thereby increasing their own performance. Additionally, they will directly and indirectly benefit from what the chosen player gains during the training program. Indirectly, through observation and imitation, and directly through the transmission of knowledge by the chosen player. The performance gains of the chosen player will therefore have a chain impact on the entire team. The question then becomes: which player should be chosen to maximize the impact on the team?

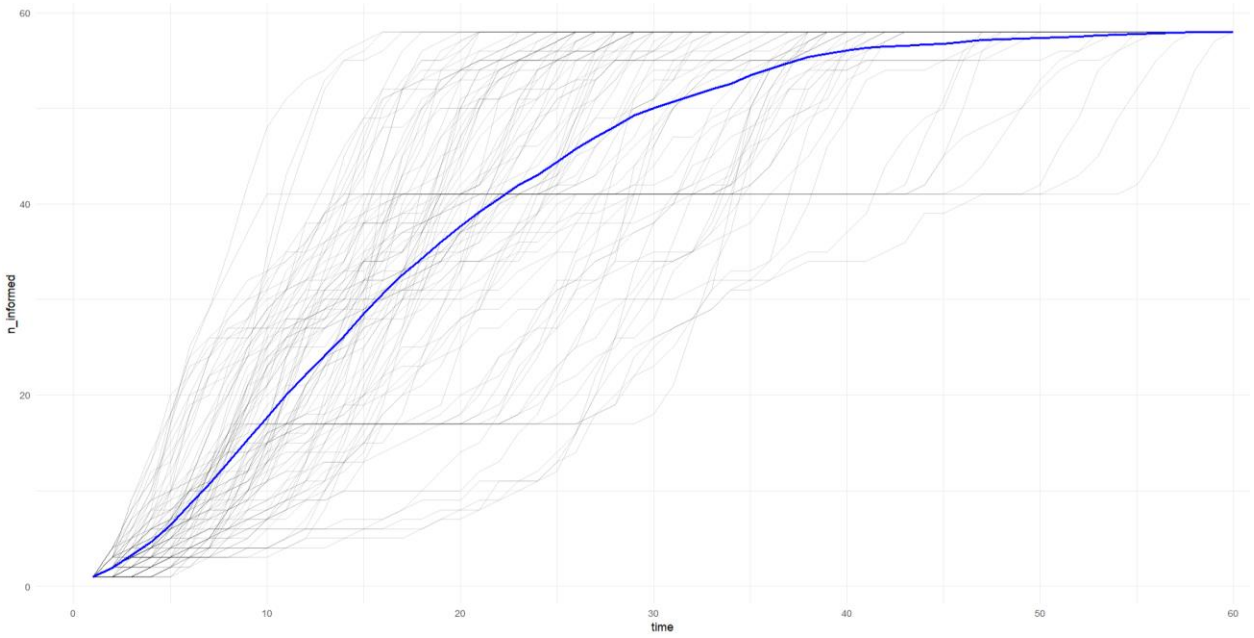
The scenario described appears to be a case of simple contagion. In simple contagion, individuals are influenced by a single contact or exposure to adopt a behavior or idea. However, in complex contagion, multiple exposures or contacts are required for adoption to occur. In the context of the rugby club scenario, having one competitors that has taken advantage of this training or its repercussions is enough.

By undergoing this iterative process, the ABM offers valuable insights into the intricate dynamics of knowledge dissemination and performance enhancement within the rugby club setting. The starting position will be given randomly, the probability to transfer skill to other player is defined as p and the method is applied to the network we have studied so far. The schema as follow represents the process where each step T represent 1 training.

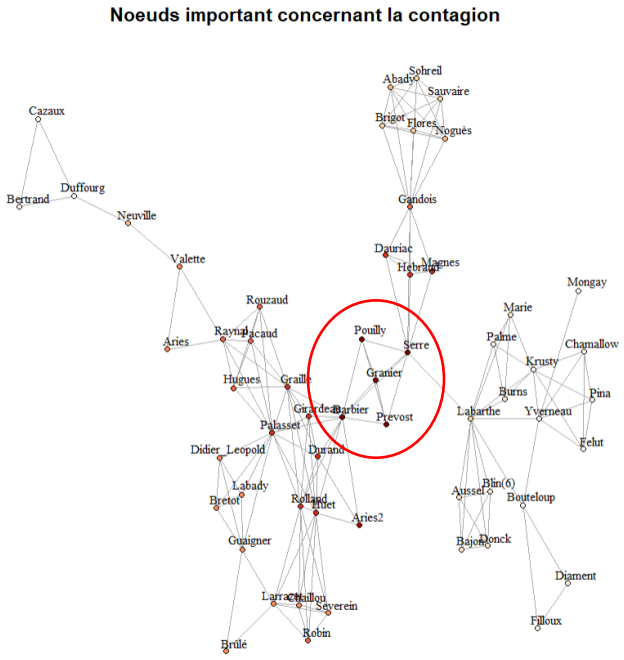


In our simulations, we will explore varying probabilities (p) to observe their effects. Determining whether p=0.5 is preferable to p=0.25 poses a challenge, as it would imply that, on average, skill transfer between players occurs over four training sessions compared to two. Additionally, we will investigate specific starting positions to assess their impact on skill transmission dynamics. All of that in order to determine if it's advantageous for the club and which player to select.

By running several time the process of contagion with different starting position and p=0.5, we get the following representation. In blue we can see the average time needed to reach a certain number of player.



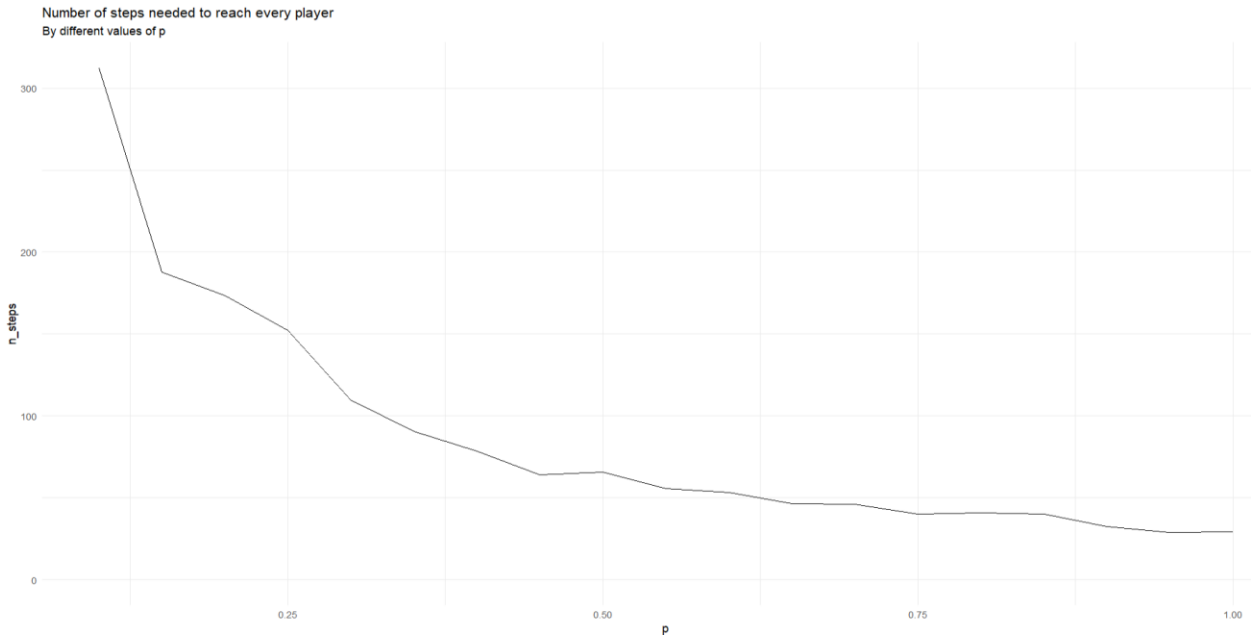
We see that for some seeds, it takes a long time before the diffusion process takes off. Who could these seeds be? And are some seeds more likely to get the rumor faster than others? We can then see how fast on average each actor gets some repercussions on his performance :



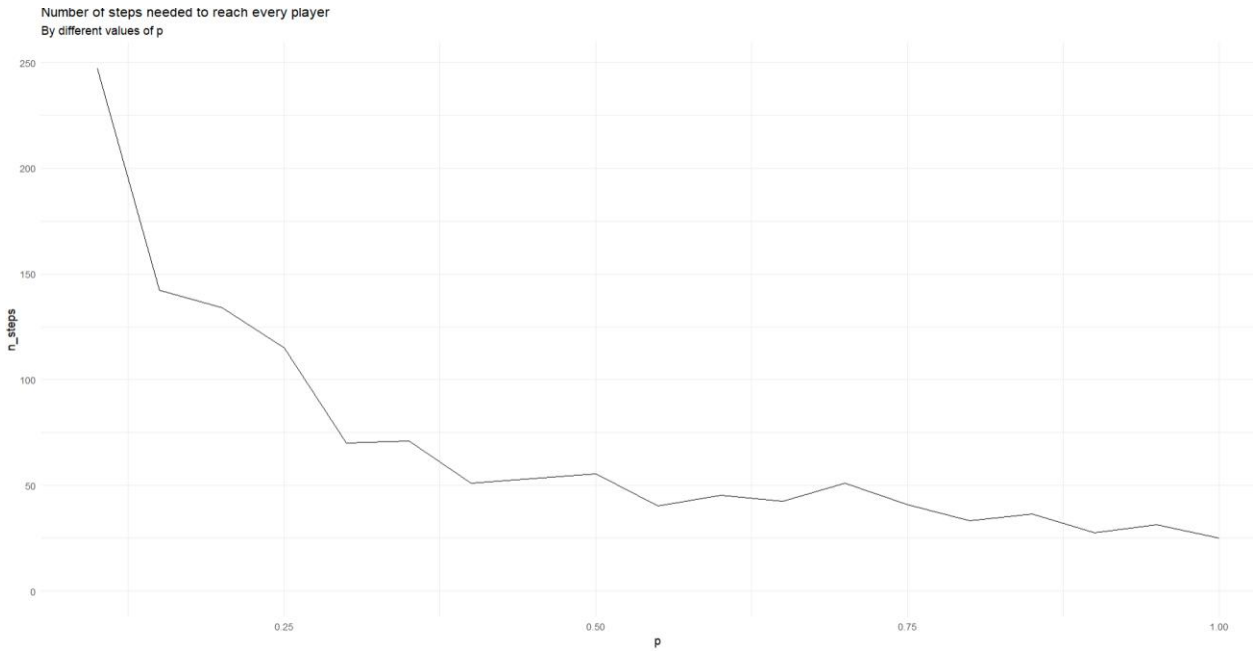
The players whose skills from such a training spread most rapidly to others are the central players. Given that these players can occupy multiple positions within the team, they are considered versatile. Therefore, it is advisable to select these players for the training opportunity.

We can also try different probabilities p, and see how quickly the rumor spreads on average. We also still assume that we don’t know where the rumor starts so we randomize the seed. Ideally, we should do many runs now, as there is quite some variance introduced by ‘freeing’ both factors.

Depending on the probability p chosen, the number of steps also called training to reached everyone fluctuates as follow :



Throughout the season, there are approximately 7 months * 4 weeks * 2 trainings per week, totaling 56 trainings or steps. We assume that the intensive training program occurs before the start of the season and that once the season ends, players who have not benefited from this knowledge dissemination will never do so. Therefore, it is in the club's best interest for all players to progress following this intensive training program. In this simulation where the selected player is random, we observe that the probability p must be greater than approximately 0.4 for it to be effective. What happens when we only consider players who have previously achieved the highest average scores to be starting positions?

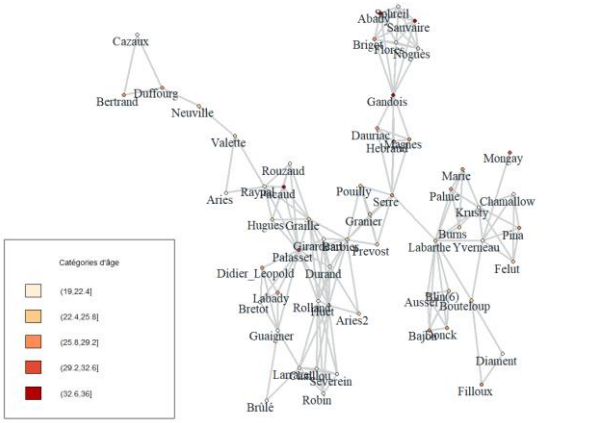


When considering only the players Barbier, Pouilly, Prevost, Granier, and Serre as starting positions, we observe a time efficiency improvement in reaching all players. To touch everyone in 56 weeks with these selected players as starting positions, the probability p needs to be approximately higher than 0,4.

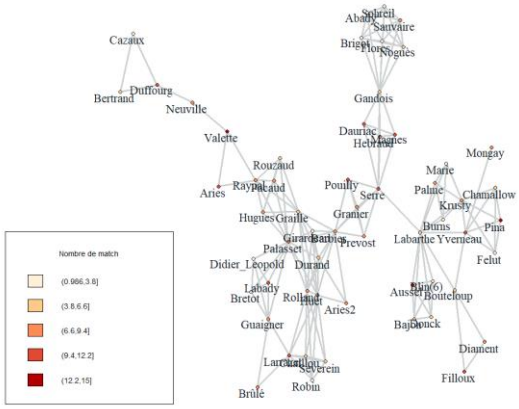
In conclusion, regardless of the transmission probability p, it remains more beneficial to send one of the central players to the network, as previously mentioned. However, if such an intensive program were not provided to clubs by the FFR (French Rugby Federation), for example, but instead incurred significant costs, we might question the cost-effectiveness of such an investment if the probability p were too low. To enhance this model, we would need to introduce some complexities into the transmission method. This could involve accounting for player absences from training sessions, as they would miss out on the effects of the intensive program. Additionally, incorporating a nonlinear effect, where the likelihood of receiving benefits increases with the number of teammates who have already benefited, could provide a more realistic representation of knowledge dissemination within the team.

Appendix

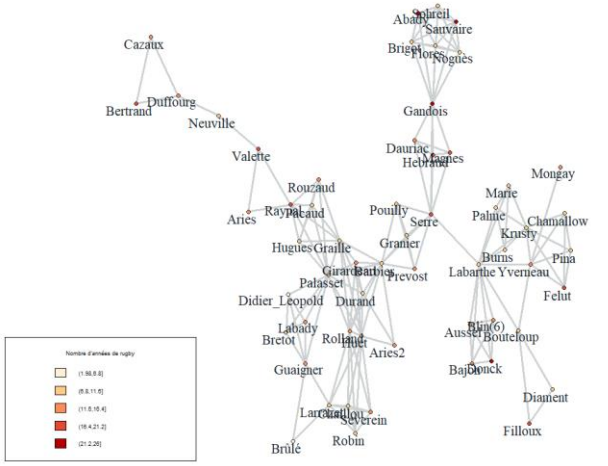
Graphique du réseau de rugby - Période 2 - selon l'âge



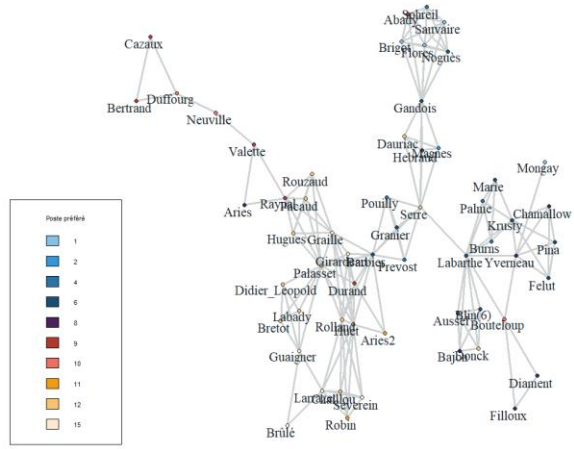
Graphique du réseau de rugby - Période 2 - selon le nombre de match



Graphique du réseau de rugby - Période 2 - selon le nombre d'année de rugby



Graphique du réseau de rugby - Période 2 - selon le poste préféré



Graphique du réseau de rugby - Période 2 - selon le poste de formation

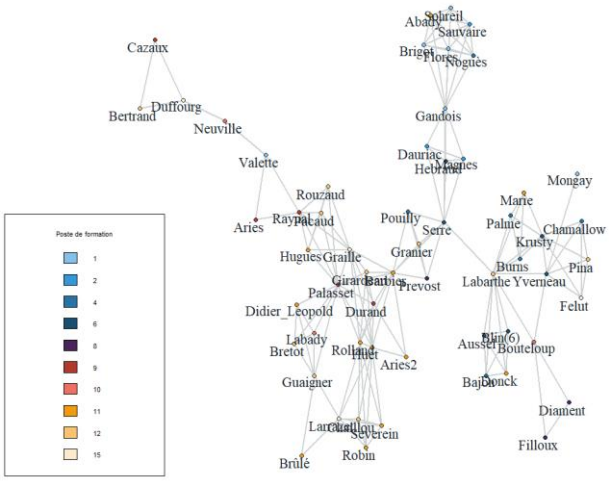


Image : positions in the team

For the attribute ‘preferred position’, positions 1 with 3, 4 with 5, 6 with 7, 12 with 13 and 11 with 14 were merged to get only 10 different positions.

For the attribute ‘preferred position reassembled’, positions (1,2,3), positions (4,5), positions (6,7,8), positions (9,10), positions (12,13) and positions (11,14,15) were merged to get only 6 different positions in this order from 1 to 6.

```
Network Logit Model

Coefficients:
             Estimate  Exp(b)     Pr(<=b) Pr(>=b) Pr(>=|b|)
(intercept) -3.066366 0.04659018 1       1       1
x1           1.622714 5.06682236 1       0       0

Goodness of Fit Statistics:

Null deviance: 2291.545 on 1653 degrees of freedom
Residual deviance: 948.8896 on 1651 degrees of freedom
Chi-Squared test of fit improvement:
         1342.655 on 2 degrees of freedom, p-value 0
AIC: 952.8896   BIC: 963.7103
Pseudo-R^2 Measures:
        (Dn-Dr)/(Dn-Dr+dfn): 0.4482008
        (Dn-Dr)/Dn: 0.585917
```

```
> mantel(rugby_adj, pref_diff,method = "pearson", permutations = 10000)

Mantel statistic based on Pearson's product-moment correlation

Call:
mantel(xdis = rugby_adj, ydis = pref_diff, method = "pearson",      permutations = 10000)

Mantel statistic r: 0.1865
      Significance: 9.999e-05

Upper quantiles of permutations (null model):
   90%    95%  97.5%    99%
0.0349 0.0449 0.0549 0.0666
Permutation: free
Number of permutations: 10000
```