

Graph-Based Modeling and Prediction of Match Outcomes in International Football Tournaments

Abstract

This project focuses on developing a graph-based approach for modeling and predicting match outcomes in international football tournaments. The proposed methodology incorporates graph theory, centrality scores, and random walk techniques to analyze the relationships between football clubs and national teams and generate predictions for match results. The model utilizes the FIFA club ranking scores as covariates to determine the influence of each team in the tournament. By assigning centrality scores to the participating teams, the model identifies the favorite team with the highest score. The random walk approach, utilizing normalized club ranking scores as transition probabilities, calculates the probabilities of each team winning a match. The research demonstrates the effectiveness of the model by achieving an accuracy score of 60.41% in predicting match outcomes in the FIFA World Cup 2022 and 54.90% in the Euro 2020 competition. These results highlight the model's ability to provide valuable insights into match outcomes and its potential to assist in betting strategies and tournament analysis. Moreover, the study suggests that combining the proposed model with statistical or advanced machine learning approaches can further enhance the accuracy of match predictions. The findings contribute to the field of sports analytics and provide a foundation for future research in the domain of football tournament prediction and analysis. To ensure transparency in the new proposed model, all relevant resources including the code, data, and illustrations can be accessed via the GitHub link provided: <https://github.com/Franosei/Network-Based-Method-for-Tournaments-Prediction>

Keywords: graph; eigenvector centrality; random walk; power method, club ranking; degree distribution

1 Introduction

Football tournaments are a common event for which statistics and machine learning algorithms are used to make predictions. To forecast the results of football games, a variety of strategies and methods have been employed, including statistical approaches, machine learning methods, AI techniques, and hybrid models (a combination of statistics and machine learning).

One of the earliest and most famous examples of using statistics for predicting football matches is the "Elo Rating System,". The Elo Rating System was adapted for use in football in the 1990s. The system uses a mathematical formula to calculate the relative strength of each team and then predicts the probability of each team winning a match. There are several other statistical models that can be used for football predictions. Dixon and Coles (1997) developed a statistical model to predict the outcome of football matches based on the number of goals scored by each team. The model took into account various factors, such as the strength of the teams and the location of the match. In the early 2000s, Ted Knutson proposed the Expected Goals (xG) which used statistical models to estimate the probability of each shot resulting in a goal. The model was based on shot location data and a variety of statistical techniques to estimate the probability of each shot resulting in a goal. Dyte and Clarke (2000), present a statistical model for simulating the outcomes of soccer matches in the World Cup using a Poisson distribution. Their model predicts the probability for each team to score a certain number of goals in a match. Karlis and Ntzoufras (2003), proposed the use of bivariate Poisson models for analyzing sports data. The bivariate model allows for the correlation between the scores of both teams to be taken into account, which can lead to more accurate predictions and insights of a match. Goddard and Asimakopoulos (2004), used a Bayesian approach to predict the outcome of football matches. They used a dataset of past football matches and took into account various factors, such as the strength of the teams, the location of the match, and the weather conditions. Mohr and Geyer (2005), used a Bayesian regularized Poisson regression model to predict the outcomes of soccer matches in the German Bundesliga. Rue and Salvesen (2007), used a Bayesian spatial model to predict the outcomes of soccer matches in the Norwegian Premier League. Their model accounted for the spatial structure of the league and included information on team strengths and home-field advantage. Karlis and Ntzoufras (2009), used regularized Poisson regression model to analyze soccer matches in the Greek Super League. Baio and Blangiardo in 2010 used a Bayesian hierarchical model to analyze soccer matches in the English Premier League. Andreas et al. (2015) demonstrate the effectiveness of

using regularized Poisson regression models for predicting the outcomes of international soccer tournaments. Boshnakov et al. (2017), demonstrate the effectiveness of using a bivariate Weibull count model for forecasting soccer scores. The model is able to capture the dependency between goals scored by different teams, which is an important factor in predicting match outcomes. Ley et al. (2019) compare different maximum likelihood approaches for ranking soccer teams based on their current strength. The paper compares the performance of different models, including the Bradley-Terry model, the Thurstone-Mosteller model, and the Logistic model, using data from international soccer matches. Machine learning algorithms and artificial intelligence have also been used to predict the outcome of football matches. Liu and Yao (2009), proposed a neural network-based approach for predicting the outcome of football matches, based on data from the Chinese Super League. Their study demonstrates that machine learning algorithms can predict football game results more accurately than more conventional techniques like the Poisson distribution and the Dixon-Coles model. Buluswar and Gutierrez (2010), demonstrate an artificial intelligence approach for predicting the outcome of football matches, based on data from the English Premier League. Karpathy (2016), uses a random forest model by using data from the English Premier League to predict the number of goals scored by each team in football matches. Petrillo and Eynard (2018), use a random forest model to predict the outcome of football matches in various European leagues. Groll and Korn (2018), discuss the use of random forest models for predicting the outcome of football matches. Vukicevic and Koprivica (2018), use several machine learning algorithms, including support vector machines and random forests, to predict the results of the 2018 FIFA World Cup. Abbasi and Sabri (2019), demonstrate the use of several machine learning algorithms, including decision tree and logistic regression, to predict the outcome of international football matches. Lin et al. (2020), use machine learning techniques, including logistic regression and k-nearest neighbors, to predict the results of the 2018 FIFA World Cup.

Hybrid models have become a popular approach for predicting the outcomes of football matches in recent years. In order to analyze enormous datasets and spot patterns and trends that might not be visible through conventional statistical analysis, AI approaches, such as machine learning algorithms, can be applied. Luengo et al. (2011), studied the use of Bayesian network classifiers to predict the outcomes of football matches with high accuracy. Malik and Ali (2018), use a hybrid approach that combines fuzzy logic and machine-learning techniques to predict the outcomes of football matches. Groll et al., present a hybrid random forest model to predict the outcomes of soccer matches in international tournaments. The model is a combination of a random forest model with Poisson ranking methods.

While statistical models, and AI techniques, can provide valuable insights into the dynamics of football tournaments and improve the accuracy of match predictions, there are also several potential disadvantages to these approaches. In the context of football tournaments, relevant data can be difficult to obtain or may not be publicly available. Football tournaments are complex and dynamic, and unexpected events such as injuries or changes in team strategies can significantly impact the outcome of matches. Previous Statistical and AI models for football predictions rely heavily on historical data, which may not always be a reliable indicator of future outcomes. AI and statistical models are often specific to a particular tournament or set of data, and may not be easily generalizable to other tournaments. This can limit the applicability and usefulness of the model beyond a specific context. Again, it can be challenging to interpret AI and statistical models, especially for non-experts. coaches, analysts, and other stakeholders may find it challenging to comprehend the assumptions behind the projections and come to wise conclusions as a result.

To address the potential problems involved in using statistics, and AI models in football predictions, a network-based method using eigenvector centrality and random walk techniques to identify the most important or influential teams in a tournament and estimate the probability of a team winning a match is developed. The model considers the quality and experience of the players in each team by considering the club ranking of each player as a key factor in determining the outcome of a match. The intuition behind this network-based approach is that a national team with the most international player pool in a tournament has a higher chance of winning the tournament, this is because players who have experience playing in top leagues and in high-pressure matches are more likely to perform well under pressure, which is important in a tournament setting. This model can be seen as a promising alternative to how national teams are ranked by FIFA during national competitions and again can offer helpful insights for coaches and analysts by taking into consideration the many connections and interactions between teams.

This report is organized into five sections. Section two introduces how eigenvector centrality is a useful measure for identifying the most influential teams in a football tournament. Section three describes how the random walk process can be used to estimate the probability of a team winning a match. In section four, we will investigate the UEFA Euro 2020 and 2022 world FIFA Men’s world cup with the new model, and section five lays a valid conclusion as well as captures future works to the study.

2 Graph Theory Approach for Ranking Football Teams

Graphs are found everywhere in our everyday lives, spanning a wide range of areas including social connections, communication networks, computer systems, the internet, and transportation. These graph structures are readily observable in our external interactions and experiences. At a deeper level, our brain cells and the intricate network of proteins within our bodies also form graph-like structures that are vital for our survival and cognitive functions. In the context of football/sport, a graph can be used to represent the relationships between teams.

A graph $G = (V, E)$ consists of two sets of information: a set of nodes $V = \{v_1, v_2, \dots, v_l\}$ and a set of vertices $E = \{e_1, e_2, \dots, e_n\}$ representing teams and the relationship between the teams respectively. In this framework, the nodes are partitioned into two disjoint sets $X = \{x_1, x_2, \dots, x_{l_1}\}$ (football clubs) and $Y = \{y_1, y_2, \dots, y_{l_2}\}$ (national teams) such that $X \cup Y = V$. We define the function

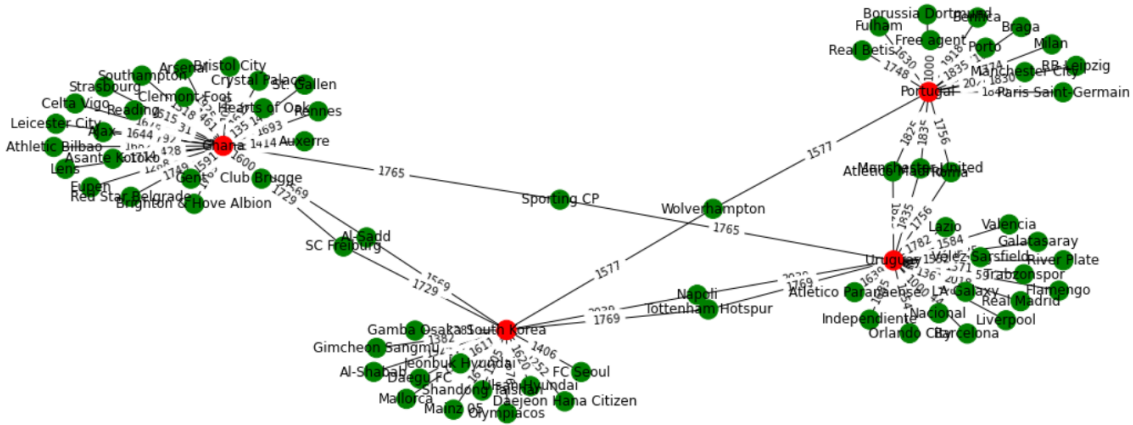


Figure 1: Graphical Representation of the Interconnections Among Group F Teams in the 2022 World Cup

$f : V \rightarrow \{0, 1\}$, where $f(v_i) = 0$ if $v_i \in X$ and $f(v_i) = 1$ if $v_i \in Y$, for $i = \{1, 2, \dots, l\}$, where l is the total number of nodes. Every edge in E has an endpoint in X and the other endpoint in Y . In Figure 1, the red node indicates a national team and the green represents football clubs. If a club x_i has a player who is representing their country y_i in a tournament, then there exists an edge between the two nodes.

The degree of a node (X or Y), denoted as $d(n_i)$ is the number of nodes adjacent to n_i . The degree in a football network ranges from 1 if a national team is adjacent to one club (all players representing a national team belong to a single club) to a maximum of the number of players allowed to represent a national team (maximum node degree for FIFA men’s world cup/Euro tournaments network is 26). Mathematically, the degree of a node n_i is computed as $d(n_i) = |\{n_j : (n_i, n_j) \text{ is an edge}\}|$, $i, j = 1, 2, 3, \dots, l$. The node degree is easy to compute but can be quite informative in football analysis. The node degree can be used to analyze the relationship between the diversity of football teams and their performance in tournaments. Keith et al. (2017), Maderer (2014), Garcia (2012), and Ignacio (2022) examine the relationship between diversity and team performance. A team with the highest node degree indicates a more diverse team. According to several articles authored by Keith et al (2013), Biehl (2017), and Jorge (2019), it has been proposed that a team with greater diversity performs better compared to a team with low diversity, but this advantage exists only up to a certain point. Once the node degree $d(n_i)$ surpasses a specific threshold, it starts to negatively impact the team’s performance. As illustrated in figure 1, Portugal, with a node degree of $d(\text{Portugal}) = 15$, outperformed Ghana

($d(\text{Ghana}) = 21$), Uruguay ($d(\text{Uruguay}) = 22$), and South Korea ($d(\text{South Korea}) = 17$) to secure the top position in Group H during the 2022 FIFA World Cup. Notably, South Korea attained the second position in the group, having the second lowest node degree among the teams.

Again, in football, node degree can also be employed to assess the level of connectedness between players within a passing network during a game. By analyzing the passes made between players, we can calculate the node degree for each player. The node degree of a player in this context refers to the number of passes they have made or received from other players. If a player has a high node degree, it signifies that they are more central within the passing network. This centrality implies that the player is actively involved in the team's passing game, frequently receiving and distributing the ball to their teammates. A player with a high node degree is often considered a key connector within the passing network, as they tend to be heavily relied upon for ball movement and playmaking. Their involvement in passing interactions indicates their importance and influence in orchestrating the team's offensive strategies. By examining node degrees in the passing network, coaches and analysts can gain insights into player dynamics, positional roles, and the overall effectiveness of the team's passing game. It helps identify players who contribute significantly to the team's passing success and highlights those who have a strong presence and impact in connecting different areas of the field. Once more, based on head-to-head outcomes, we can create a network of teams and analyze their competitive nature within a league. One way to assess a team's competitiveness is by determining its node degree within the network. Here node degree refers to the number of connections or interactions a team has had with other teams. In this context, if a team has a high node degree, it indicates that they are more central within the league network. This centrality suggests that the team has encountered and potentially triumphed over a larger number of other teams. In other words, a high node degree reflects a team's involvement in more matches and victories against various opponents. It serves as an indicator of the team's competitive standing within the league.

The variance of the node degree in football can be influenced by a number of aspects of the game's dynamics and structure. The variance of node degree is influenced by the number of teams competing in a league/tournaments, the overall number of games played, and the manner in which points are scored in those games. The variance of node degree is typically low if the games in a league/tournament are evenly split among all the teams and the scoring trends are known. This implies that most teams would interact with other teams on a comparable basis. There would be less variance because passes, goals, and other exchanges would be distributed fairly evenly across the teams. But on occasion, a small number of teams can have much more or less connections or interactions than the vast majority of teams. These clubs/teams may have played more games, faced more difficult opponents, or scored more or less goals. As a result, there is a larger disparity between the number of connections for these clubs and the other teams in the league, which raises the variation of node degree. Node degree variance, in general, represents the variety and distribution of connections or interactions between teams in a football league/tournaments. The variance of node degree in a graph can be mathematically expressed as:

$$S_D^2 = \frac{\sum_{i=1}^l (d(n_i) - \bar{d})^2}{l}, \quad (1)$$

where $\bar{d} = \frac{\sum_{i=1}^l d(n_i)}{l}$ denotes the average node degree. The degree distribution in football networks can frequently be described by a power law, with a few teams having significantly higher degrees than others. In these circumstances, the variance may be relatively high, indicating a significant fluctuation in the connectedness of various network nodes.

2.1 Degree Distribution

The node degree may not be enough to fully describe the connection of the network in a weighted network. The strength, severity, or relevance of the connections is often represented by these weights. We only consider the amount of connections or interactions a node has with other nodes, regardless of the intensity or weight of those connections, when calculating node degree in a weighted network. This may leave out important details regarding the actual impact or importance of the links. The strength of links in a weighted network might vary greatly. Strong connections carry a lot of weight, but weaker connections have less of an impact. It is possible to oversimplify the network's connectedness by only counting connections without taking their weights into account. In a weighted network, taking into account the weights attached to the connections becomes crucial to gaining a more thorough knowledge

of the network's connectivity. The significance of connections is captured by weighting the analysis and using metrics like weighted node degree or weighted degree centrality. By highlighting the varied strengths and relevance of connections, these metrics give a more complex picture of the network. The degree distribution is a crucial component of network research since it can reveal information about the composition and dynamics of a network. The degree distribution of an unweighted graph is a probability distribution that describes the proportion of nodes in the network that have a given degree. Degree distribution can be used in the context of a football tournament to forecast the likelihood that specific teams will perform well or poorly based on the relationships between teams. The degree of a team can represent the number of games played or the number of points earned. Mathematically, the degree distribution of a graph $G(V, E)$ is given as:

$$P(k) = (\text{number of nodes with degree } k) / l \quad (2)$$

where k is the degree and $P(k)$ is the probability that a randomly selected node in the network has degree k . In a weighted graph, as seen in figure 1, each edge has a corresponding weight or strength that describes how significant or intense the link is between the two nodes.

Assume we have a weighted network represented as an adjacency matrix, where each entry (i, j) in the matrix represents the weight of the connection between nodes i and j . $W(= [w_{ij}])$ is called a "weighted adjacency matrix" with $l \times l$ elements where l is the total number of nodes in the network, in which $w_{ij} = 0$ when there is no edge between nodes i and j but $w_{ij} = w$ when an edge exists, where w is a real number. The mathematical expression to compute the total weight, s , of node i in the weighted network can be written as follows:

$$s_i = \sum_j w_{ij} \quad (3)$$

In football tournament predictions, the degree distribution of a weighted network represents the distribution of the total weights of the edges connecting each node. The distribution of the total weights of all edges connected to a particular node makes up the degree distribution in a weighted network.

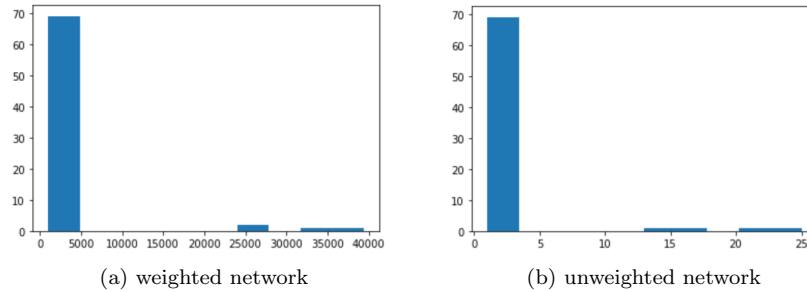


Figure 2: Degree distribution

Figure 7 shows the degree distribution of figure 1. The x-axis of the weighted network represents the edge weights while the x-axis of the unweighted network represents the node degrees. The y-axis represents the frequency or occurrence of each degree/weight value. It indicates how many nodes in the network have a specific degree/weight. The degree distribution of a network can follow different types of distributions depending on the characteristics of the network. From figure 7, the degree distribution follows power law distribution. It is a scale-free distribution used to explain networks where a small number of nodes have extremely high degrees while the majority of nodes have low degrees. Mathematically, the power law distribution can be expressed as:

$$P(k) \propto k^{-\alpha} \quad (4)$$

where $\alpha \geq 1$ is a scaling factor. Since many real-world networks, including football networks, share the power law distribution, it is crucial for football prediction. The power law distribution of the degree distribution suggests that there are relatively few "superstar" players with very high degrees (very well-known or influential), whereas the majority of players have low to moderate degrees in the context

of football prediction. In a football network, it's critical to recognize and evaluate the major players because they may have a disproportionate influence on how well their team performs. The creation of network-based prediction models that include centrality measures like degree, eigenvector centrality, and PageRank can also benefit from the power law distribution. PageRank is a technique that aids in determining the strength and significance of football teams in the context of football prediction. Teams are ranked according on how well they perform when compared to other strong teams, much as how web pages obtain votes of confidence from other significant pages. A team's rating value increases if it defeats a team with a high opponent ranking. To estimate the relative calibre of each team, the algorithm takes into account the entire network of match results. We can identify teams that have regularly outperformed formidable foes using PageRank, and we can utilise this knowledge to produce predictions about match results that are more precise. The complex relationships and interactions between players in the network are captured by the power law distribution, which might help in the construction of simulation models for better football prediction. Overall, the power law distribution is a crucial idea to take into account when making football predictions since it can shed light on the structure and dynamics of football networks and help design prediction models that are more precise. Consideration of a team's opponents' strength is one practical example of applying network analysis and power law distribution to football prediction. We can discover the relationships and interactions between teams by building a network of teams based on the results of their games, where each team is a node and the edges denote the games played. We can identify the existence of a few extremely powerful teams by using the power law distribution. A team's ability to play at a high level can be determined by how frequently they have encountered and overcome formidable opponents. A team's performance, on the other hand, might not be as dependable if it has primarily faced inferior competition and has had little contact with powerful rivals. Therefore, by combining network analysis and power law distribution, we can evaluate the strength of opponents a team encountered, revealing insightful information about their performance and assisting in more precise match prediction.

2.2 Eigenvector Centrality

Since the early 20th century, there have been many different and interesting centrality measures used in graph theory. Jacob Moreno developed the idea of centrality in social networks in the 1930s and coined the phrase "sociometric status" to refer to an individual's significance within a network. A related idea known as "status centrality" was proposed by psychologist Floyd Allport in the 1940s to define a person's place within a group. The idea of centrality in communication networks was first established by Alex Bavelas in his key study "A Mathematical Model for Group Structure" published in 1948. Bavelas employed an experiment in which participants were divided into groups of three and asked to use a communication network to solve a straightforward task. The findings demonstrated that participants with a more centralized communication network had a tendency to solve the issue more quickly and precisely than participants with a less centralized network. Linton Freeman proposed the notion of "eigenvector centrality" in the 1970s, which takes into account not just the number of connections a node has, but also the importance of the nodes to which it is connected. This measure has now become one of the most commonly used centrality measures in network analysis. Centrality measurements are now employed in a variety of applications ranging from social network research to the study of biological networks, transportation systems, and the Internet. They remain an important tool for comprehending the structure and dynamics of complex systems.

Eigenvector centrality is a measure of a node's importance in a network based on its connections to other significant nodes in the network. In the context of football prediction, eigenvector centrality can be used to determine the most influential teams in a tournament based on their past success, and the strength of the teams they've faced. Prior to moving on to the eigenvector centralities first, a quick review of eigenvectors.

The fundamental goal of eigenvector-based analyses is to reduce the number of raw variables to a manageable number of vectors, such as principal components or factors, that somehow capture or approximate the variability or information in the raw data. Selecting the appropriate number of components or factors to take into account is a crucial step in principal component and factor analysis. The majority of the time, numerous components are required, yet rarely one component may be sufficient. These analyses utilise eigenvalues and eigenvectors, the fundamental building blocks of multivariate statistical models. The principal components and factor analysis model operates as follows: Given a dataset X with dimensions $N \times p$ (representing the number of observations and variables, respec-

tively), it is transformed into a correlation matrix R of size $p \times p$. From this correlation matrix, two separate matrices are derived: eigenvalues Λ , ordered as $\lambda_1, \lambda_2, \dots, \lambda_p$, and eigenvectors \mathbf{W} , where each column consists of entries w_1, w_2, \dots, w_p . The transposed matrix of eigenvectors is denoted as \mathbf{W}' . Mathematically, the relationship between the correlation matrix and the eigenvectors is expressed as $R = \mathbf{W}\Lambda\mathbf{W}'$. By multiplying the eigenvectors with the corresponding variables of the dataset, we obtain a new score y_i for each observation i , given by $y_i = w_1X_{i1} + w_2X_{i2} + \dots + w_pX_{ip}$. The first eigenvector contains the weights that best transform the original variables into a single new score, explaining the maximum possible variance in the dataset X . Each variable is associated with a weight in the first eigenvector. Subsequent eigenvectors $2, 3, \dots, p$ include weights that create new variables, explaining the maximum remaining variance while ensuring that each newly created variable is uncorrelated with the previous ones. This framework allows us to extract essential information and patterns from the data.

In network analysis, eigenvectors and eigenvalues are used to measure the centrality of nodes in a network. With the help of eigenvectors and eigenvalues, significant nodes in a network can be located based on their connections to other significant nodes, and predictions about the behavior of the network as a whole can be made. In the context of football prediction, we can apply the concept of eigenvector centrality to understand the importance of teams in a football network. Just as in a social network, where some individuals are influential because they have connections with other influential individuals, in football, certain teams may gain significance by having ties with other influential teams rather than simply participating in numerous matches or having players coming from different nationalities. Mathematically, eigenvector centrality assigns a score x_i to each team $i \in V$ in a way that the score of a team is proportional to the collective score of its connected teams. Thus, we could like to find a vector $x \in \mathbb{R}^n$, such that

$$x_i = \beta \sum_{j \in V: (i,j) \in E} x_j = \beta \sum_{j \in V} A_{ji} x_j \quad (5)$$

where A_{ij} is 1 if there is a connection between teams i and j , and 0 otherwise and β is a scalar value. We want to find a vector x such that x_i represents the centrality score of team i . Equation 5 can be represented in a vector notion as $x = \beta A^T x$. The problem is equivalent to solving

$$Ax = \beta^{-1}x \quad (6)$$

There is a vector w_i such that $Aw_1 = \lambda_1 w_1$, where λ_1 is the largest eigenvalue of matrix A . This vector is what we call the eigenvector centrality vector; $(w_1)_i$ measures the centrality score of node i . Note that the eigenvector centrality vector solves equation 6 when $\beta = w_1^{-1}$. In the football context, this equation signifies that each team's centrality score is proportional to the combined scores of their connected teams. Football tournament predictions can benefit from using eigenvector centrality because it reveals which teams are more influential inside the competition. Eigenvector centrality can highlight the teams that play a key role and have a big impact on the overall result by taking connections and performance of teams into account in the tournament network. Let's consider a tournament with 4 teams: Team W, Team X, Team Y, and Team Z. We want to determine the eigenvector centrality scores for these teams and understand their significance in the tournament network. Assume after performing the computations, we obtain the eigenvector centrality scores for the teams, which are as follows: Team W: 0.501, Team X: 0.577, Team Y: 0.577, Team Z: 0.501. Based on the obtained results, it is apparent that Teams X and Y possess higher eigenvector centrality scores compared to Teams W and Z. This indicates that Teams X and Y hold greater significance within the tournament network and exert a more substantial influence on the ultimate rankings. As an example, the eigenvector centrality scores of Team X and Team Y will increase if they achieve success and secure victories against formidable opponents, symbolizing their impact on the network. Conversely, if Team W and Team Z continue to experience defeats, their centrality scores may decrease. This metric allows for a more thorough assessment of a team's effect because it considers both the quality of the opponents a team faced and the team's own performances. Here are examples where eigenvector centrality plays a significant role in football prediction:

- **Team Ranking:** Eigenvector centrality can be used to rank teams based on their total influence by building a network where teams are represented as nodes and the connections between them as edges. Taking into account both the strength of their rivals and their own performances, each

team’s eigenvector centrality ratings represent their significance within the network. Teams with higher eigenvector centrality scores are thought to be more influential and are probably more likely to succeed in upcoming games.

- In competitions featuring a group stage, eigenvector centrality can be used to forecast which teams would likely proceed to the knockout rounds. The eigenvector centrality scores can be used to determine which teams within each group have the most sway by looking at the relationships between them. Because of their advantageous network placements, these teams have a higher chance of winning and moving on to the next round.
- Upset Predictions: Eigenvector centrality can also be useful in predicting potential upsets in football tournaments. By considering the centrality scores of underdog teams and their connections to stronger teams, it is possible to identify situations where a lower-ranked team has a higher chance of defeating a higher-ranked team. This analysis can take into account past performance data, player statistics, and the strength of the opponents faced by each team.
- Analysis of Player Performance: Eigenvector centrality can be used to analyse player performance by taking into account the passing networks or interaction patterns that occur between players during games. It is feasible to identify players who play a key role in the team’s gameplay and greatly contribute to the team’s performance by building a network of passes between players and using eigenvector centrality. These players are more likely to have a greater effect on the team’s performance as a whole.

By analyzing the eigenvector centrality vector, we can identify the most influential teams in the football network, allowing us to make predictions about their impact on other teams or their overall performance. To compute the eigenvector centrality vector v_1 , we can utilize an iterative process known as the power method.

2.2.1 Power Method

In linear algebra, we have observed that repeatedly multiplying a vector by a matrix can significantly increase the contribution of the largest eigenvalue (denoted by λ_1). This principle forms the foundation for various algorithms designed to compute eigenvectors and eigenvalues. The power method is one of the simplest and most fundamental algorithms used for this purpose. The power approach can be used in the following situations to compute the eigenvector centrality vector in football prediction:

1. Choose an initial vector $\mathbf{x}^{(0)}$ of length n , where n is the number of teams in the football network.
2. Perform the following steps for a specified number of iterations or until convergence is reached:
 - 2a. Update the vector $\mathbf{x}^{(t)}$ at iteration t using the equation $\mathbf{x}^{(t)} = A\mathbf{x}^{(t-1)}$, where A is the adjacency matrix.
 - 2b. Normalize the vector $\mathbf{x}^{(t)}$ by dividing it by its largest element to prevent numerical instability and ensure that the magnitudes of the vector remain in a reasonable range.
3. Check for convergence by comparing the difference between $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(t-1)}$. If the change is below a certain threshold or after a fixed number of iterations, consider the process as converged.
4. The final normalized vector $\mathbf{x}^{(t)}$ represents the eigenvector centrality vector, where each element $x_i^{(t)}$ corresponds to the centrality score of the corresponding team/node i .

The power method effectively amplifies the influence of the dominant eigenvector of the adjacency matrix A and converges towards the eigenvector centrality vector, which provides insights into the relative importance of each team in the football network based on their connectivity.

Algorithm 1 Power Method for Teams Scoring During Tournaments

Require: Graph G , Initial node scores $\mathbf{s}^{(0)}$, Convergence threshold ϵ

Ensure: Dominant node scores $\mathbf{s}^{(t)}$

```
1:  $t \leftarrow 0$ 
2: repeat
3:    $\mathbf{s}^{(t+1)} \leftarrow AG\mathbf{s}^{(t)}$ 
4:    $\mathbf{s}^{(t+1)} \leftarrow \frac{\mathbf{s}^{(t+1)}}{\|\mathbf{s}^{(t+1)}\|}$ 
5:   if  $\|\mathbf{s}^{(t+1)} - \mathbf{s}^{(t)}\| < \epsilon$  then
6:     break
7:   end if
8:    $t \leftarrow t + 1$ 
9: until convergence
10: return  $\mathbf{s}^{(t+1)}$ 
```

In this algorithm, we start with a graph G representing the football network and the initial node scores $\mathbf{s}^{(0)}$. We set a convergence threshold ϵ to determine when the algorithm stops. The algorithm iteratively updates the node scores using the multiplication of the adjacency matrix A and the graph G with the current node scores $\mathbf{s}^{(t)}$. The updated node scores are then normalized to ensure they remain in a valid range. At each iteration, we check if the difference between the updated node scores and the previous iteration is below the convergence threshold. If it is, the algorithm exits the loop and returns the final dominant node scores $\mathbf{s}^{(t+1)}$. By performing these iterations and updates, the power method helps identify the key nodes in the football network that have a significant impact on the overall network dynamics and can be useful for football prediction tasks.

When calculating dominating eigenvectors and eigenvalues, the power method has a number of advantages over alternative approaches. The power approach is frequently chosen for the following main reasons:

- **Efficiency and Simplicity:** The power approach is computationally effective and comparatively easy to use. It calls for iterative matrix-vector multiplications, which are effectively carried out by optimised algorithms and libraries. This makes it a useful option for complex issues, such as examining football networks with a huge number of nodes and connections.
- **Dominant Eigenvalue and Eigenvector:** The power method is specifically made to calculate the dominant eigenvectors and its corresponding eigenvalues. The dominant eigenvector frequently contains essential details about the most important nodes or elements influencing the system's behaviour in numerous applications, including football prediction. The power technique offers a focused and efficient remedy by concentrating on the dominant eigenvector.
- **Convergence:** Iteratively, the power technique converges to the dominating eigenvector and eigenvalue. With each iteration, it starts with an initial estimate and gradually increases the accuracy. In spite of the fact that convergence is not guaranteed for all matrices, it frequently works well in practise, particularly for matrices having a dominating eigenvalue that is much larger than the rest.
- **Scalability:** The power method scales well and effectively handles both large and sparse matrices. This is especially beneficial when it comes to football predictions because there can be a large number of links between teams or players. With the power technique, complicated networks can be analysed without sacrificing performance or accuracy.

3 Rating of Football Matches through Graph-Based Analysis

In the world of football, it is essential to precisely gauge each team's strength and performance for a variety of reasons, including tournament predictions and team rankings. Traditional scoring methods frequently focus on straightforward data like wins, losses, and goal differentials, but they could ignore important elements that affect a team's success as a whole. To overcome these limitations, Graph-Based Analysis provides a powerful framework for analyzing football matches and capturing the intricate

relationships between teams. By representing football matches as a graph, with teams as nodes and match results/team ranking as weighted edges, we can leverage graph-based analysis techniques to derive more comprehensive and insightful ratings.

Imagine we have a dataset of football matches from a league/tournaments, including information on the teams involved and the outcomes of each match. We can construct a graph representation of this data, where each team is represented as a node in the graph. To create the edges in the graph, we can assign weights to them based on the match results or team rankings. For example, if Team A defeats Team B in a match, we can assign a weight to the edge connecting Team A and Team B to represent this result. The weight could be determined by the goal difference or other metrics that reflect the team's performance in that match. Once we have constructed the graph, we can apply various graph-based analysis techniques to derive more comprehensive and insightful ratings for the teams. Furthermore, by performing graph clustering algorithms (Distribution-based Clustering, Hierarchical Clustering, etc), we can identify groups or communities of teams with similar performance characteristics or playing styles. This can help us gain insights into the competitive landscape of the league/tournament and identify potential rivalries or strategic matchups. Football games are rated using graph-based analysis, which involves looking at the relevance and connection of teams within the graph while taking into account both their own performances and the strength of their rivals. We can find hidden dynamics and trends that affect team rankings using this method. Take tournament predictions as an illustration of how this approach might be used. We can calculate the likelihood of different teams winning a match by examining the graph structure and utilising graph algorithms. This allows us to model the progress of a tournament and anticipate the outcomes of individual matches, giving us valuable insights into the likely results and potential surprises.

In the realm of football analysis, random walks serve as valuable tools for algorithm design and probabilistic analysis, finding extensive applications. The concept entails working with a graph that represents the football network and initiating the process from a specific team. At each step, a neighboring team is chosen uniformly at random, and the process moves to that team. This procedure is repeated by selecting a random neighboring team from the current position and proceeding accordingly. The resulting sequence of teams chosen in this manner constitutes a random walk on the graph, capturing the dynamics of team transitions in a probabilistic manner. To initiate the discussion on random walks in football match prediction, it is essential to introduce fundamental definitions and conduct an initial analysis of random walks within a graph-based framework.

3.1 Basic Definitions

Consider a random walk on $G(V, E)$, the random walk starts from a specific node $v_0 \in V$. At each step t , if the walker is at node v_t , it moves to one of the neighboring nodes of v_t with a probability of $\frac{1}{d(v_t)}$, where $d(v)$ represents the degree (number of connections) of node v . The initial node v_0 can be fixed, or it can be chosen randomly from an initial distribution P_0 . We define the distribution of the random walk at step t as P_t , where $P_t(i)$ represents the probability that the walker is at node i at step t . In other words, $P_t(i)$ is equivalent to the probability:

$$P_t(i) = P_t(v_t = i) \quad (7)$$

The transition probabilities between nodes are denoted by the matrix $M = (p_{ij})_{i,j \in V}$, where $p_{ij} = \frac{1}{d(i)}$ if there is an edge between nodes i and j , and 0 otherwise. We can represent M in matrix form as DA , where A is the adjacency matrix of the graph $G(V, E)$ and D is a diagonal matrix with $(D)_{ii} = \frac{1}{d(i)}$. The evolution of the probabilities can be described by the equation:

$$P_{t+1} = M^T P_t \quad (8)$$

which implies $P_t = (M^T)^t P_0$. This equation allows us to compute the distribution of the random walk at any step t given the initial distribution P_0 . A stationary distribution P_0 is defined as the distribution that remains unchanged as the random walk progresses. In other words, if $P_1 = P_0$, we say that P_0 is a stationary or steady-state distribution for the graph $G(V, E)$. In this case, $P_t = P_0$ for all $t > 0$.

For any graph $G(V, E)$, the stationary distribution $\pi(i)$ is given by:

$$\pi(i) = \frac{d(i)}{2m}, \quad (9)$$

where $d(i)$ represents the degree of node i and m is the total number of edges in the graph.

3.2 Markov Chain

Random walks on graphs can be considered as specific instances of Markov chains, which are fundamental concepts in probability theory. To provide some context, let us introduce a few key findings from the general theory of Markov chains. A Markov chain X_0, X_1, X_2, \dots is a discrete-time stochastic process that operates within a set of states S and is characterized by a transition probability matrix P . At each time step t , the Markov chain is situated in a particular state $X_t \in S$. The state X_{t+1} at the next time step relies solely on the current state X_t and does not rely on the current time or any preceding states. Therefore, the random variable X_0, X_1, X_2, \dots can be fully described by specifying a distribution for the initial state X_0 and the transition probabilities p_{ij} , which represent the likelihood of transitioning from state i to state j . By setting the set of states S equal to the graph vertices V and the transition probability matrix P equal to the adjacency matrix M , the connection between random walks and Markov chains becomes evident.

A stationary distribution for a Markov chain is a probability distribution denoted by π over the states of the chain such that the equation $\pi = P^T \pi$ holds, where P is the transition probability matrix. A Markov chain is said to be irreducible if, for any pair of states i and j , there exists a sequence of transitions with positive probabilities that connects state i to state j , and vice versa.

The periodicity of a state i in a Markov chain refers to the maximum integer T for which there exist an initial distribution q_0 and a positive integer α such that the probability of being in state i at time t , denoted by $q_t(i)$, is zero if t does not belong to the arithmetic progression $\alpha + T \cdot k | k \geq 0$. On the other hand, an aperiodic Markov chain is one in which all states have a periodicity of 1, meaning that there are no such arithmetic progressions and the probabilities remain non-zero for all times.

Theorem 3.1 (Fundamental Theorem of Markov Chains). *Any finite, irreducible and aperiodic Markov Chain has the following*

1. *There is a unique stationary distribution π such that $\pi_i > 0$ for all states i*
2. *Let $N(i, t)$ be the number of times the Markov chain visits state i in t , then*

$$\lim_{t \rightarrow \infty} \frac{N(i, t)}{t} = \pi(i). \quad (10)$$

Applying the aforementioned theorem to a random walk on simple connected graphs, we can make the following observations. The periodicity of each state (vertex) in the graph $G(V, E)$, is determined by the greatest common divisor of the lengths of all closed walks within the graph. This means that the behavior of the random walk, specifically the recurrence or periodicity of visiting certain states, is influenced by the relationships between different paths in the graph. This information helps us understand how the random walk moves through the graph and how often it revisits certain states. Therefore, the random walk is aperiodic if and only if the graph $G(V, E)$ is not bipartite, meaning the set of vertices can be divided into two disjoint subsets without any edges existing between vertices within the same subset. If the graph $G(V, E)$ is both connected and non-bipartite, we can deduce that the distribution of the random variable v_t tends towards a unique stationary distribution, denoted as $\pi(i) = \frac{d(i)}{2m}$, as t approaches infinity, regardless of the initial distribution P_0 which is the probability distribution of the starting position of the random walk. Additionally, we observe that the expected number of steps between two visits to a vertex i is given by $\frac{2m}{d(i)}$. This result indicates that in a connected and non-bipartite graph, the random walk will eventually reach a state where the probabilities of being at different vertices stabilize. These probabilities are proportional to the degree of each vertex, reflecting the influence of the vertex's connectivity in determining its likelihood of being visited during the random walk. In our model, we may not observe a situation where the probabilities of being at different vertices stabilize over time. This is because our model is based on a connected bipartite graph. Due to the specific structure of our bipartite graph, the behavior of the random walk may differ, and we may not observe the convergence of probabilities to a stationary distribution.

An interesting characteristic of random walks on undirected graphs is their time-reversibility. This means that a random walk considered in reverse is also a random walk. This property can be verified by noting that $p_{ij}\pi(i) = p_{ji}\pi(j)$. The equation states that the product of the transition probability

p_{ij} and the stationary distribution $\pi(i)$ is equal to the product of the reverse transition probability p_{ji} and the stationary distribution $\pi(j)$. This equality implies that the probabilities of transitioning between vertices in both directions are balanced when considering the stationary distribution. The reasoning behind this result can be heuristically explained as follows: Time-reversibility implies that a stationary walk traverses each edge in both directions with the same frequency, on average, once every $\frac{1}{2m}$ times. By extension, the expected number of steps between two visits to a vertex i is also $\frac{2m}{d(i)}$.

3.3 Football Match Rating through Random Walk Analysis

Utilizing the information encoded in the weighted graph to assess team strengths, random walk simulation offers a means to predict match outcomes. By assigning probabilities based on the weights of the edges connecting teams, the random walk iteratively moves between the current team and its neighboring teams. This simulation captures the dynamics of the match and enables an estimation of the likelihood of one team emerging victorious over another. In the context of understanding how players' affiliations with football clubs and national teams influence game outcomes, we can utilize a random walk process. This approach allows us to simulate the dynamics of the game and gain insights into the probabilities of each team winning. By applying the random walk process, we can examine the connections between players, football clubs, and national teams. Players who are associated with both a specific football club and a national team create a link between the corresponding nodes representing the club and the national team. To assess the impact of players on game outcomes, we can simulate the random walk on this network. Each step of the random walk represents a transition from the current node (representing a team) to one of its connected nodes. The likelihood of moving to a particular connected node is determined by the strength of the connection, which reflects the influence of players' affiliations. By conducting the random walk simulation, we can observe how the walker moves from one team to another based on the strengths of the connections. This provides insights into how players' connections to football clubs and national teams can affect the probabilities of each team winning.

Imagine we have a network that includes 8 football clubs (A, B, C, D, E, F, G, H) and three national teams (X, Y, Z). To represent this network, we use a graph where each football club and national team is represented by a node. The weighted edges in the graph represent the relationships between national teams and football clubs, with the weights corresponding to the club rankings. A higher weight indicates a higher rank, implying that the corresponding national football club is considered more successful in the network. Conversely, a lower weight suggests a lower rank, indicating a comparatively weaker position for the team. It's important to note that the edges only exist between the national teams and the football clubs, representing their interactions. This means that we are focusing on the relationships and connections between these two groups of nodes within the network.

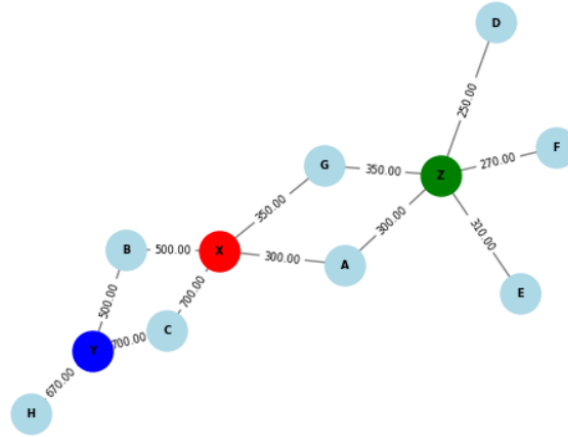


Figure 3: Football Clubs and National Teams Network with Club Ranking Score as Weights

Figure 3 represents the interconnectedness of the football clubs and national teams. The weights assigned to the edges indicate the strength or importance of the relationship between the corresponding nodes. These weight values provide insights into the relative strengths or associations between the

national teams and football clubs, contributing to a better understanding of the network structure and the dynamics within it. To further analyze the network, we compute the centrality score for each node. The importance of the weights becomes evident when considering their impact on the eigenvector centrality score calculation. By incorporating the weights into the computation, we take into account the strength of the relationships represented by the edges. In this case, the weights represent the prominence of the connections between national teams and football clubs. When calculating the eigenvector centrality scores, the weights play a crucial role in determining the influence and importance of each node. Nodes with higher weights in their connections will have a greater impact on the scores, indicating their stronger influence within the network. Thus, the weights provide valuable information for accurately assessing the relative significance and prominence of the national teams and football clubs in the overall structure of the graph. This combined analysis allows us to identify key players, influential clubs, and prominent national teams.

| Teams | Eigenvector centrality score |
|-------|------------------------------|
| X | 0.5015 |
| Y | 0.5069 |
| Z | 0.4012 |

Table 1: Evaluation of Centrality Scores for National Teams (Figure 3)

From table 1, the centrality score for each node is calculated by considering its connections to other nodes and the weights associated with those connections. Nodes that are connected to highly ranked clubs are likely to have higher centrality scores, indicating their proximity or influence in the network. Team Y having a higher centrality score (0.5069) compared to Node X (0.5015) suggests that Node Y is connected to clubs with higher rankings, contributing to its overall centrality. Similarly, Node Z having a lower centrality score (0.4012) indicates that its connections are predominantly with clubs of lower rankings.

To perform a random walk from national team X to any national team from figure 3, we can employ the concept of transition probabilities. At each step of the random walk, the walker moves to one of the neighboring nodes with a specific probability, determined by the weights assigned to the edges. The random walk process enables us to traverse the network and explore the likelihood of reaching any national team V starting from national team X . To perform the random walk from national team X to national team ξ , we can follow these steps:

1. Start at national team X .
2. Determine the neighboring nodes of X (football clubs connected to X).
3. Assign transition probabilities for each neighboring club based on the weights (club ranking score) of the edges. These probabilities indicate the likelihood of transitioning from the current national team X to each neighboring club.
4. Randomly select a neighboring club to move to based on the transition probabilities.
5. Calculate probability of strength for team X against team ξ based on eigenvector centrality.

The transition probabilities play a crucial role in determining the likelihood of moving from one node to another during the random walk. These probabilities capture the influence between national teams and football clubs, which is represented by the weights assigned to the edges connecting them. The random walk methodology provides us with a means to explore the dynamic behavior of the network and gain insights into the probabilities of reaching different nodes starting from a given point. It can be employed to analyze various aspects such as player movements between clubs and national teams, the impact of the centrality score on the success of the national team, and the likelihood of specific match outcomes based on the structure of the network. This analysis allows us to understand the dynamics and interactions within the network, shedding light on the relationships between the involved teams. To elucidate the probabilities associated with specific match outcomes using the given network structure and employing a random walk approach, we can employ mathematical notation to express these likelihoods. Let's define a weight function $w(e)$ that assigns a weight to each edge $e \in E$, which represents the club ranking score of the corresponding. To calculate the likelihood of a specific match

outcome between two nodes, such as a match between Team X and any other national team ξ , we can perform a random walk starting from Team X and compute the probability of reaching Team ξ within a certain number of steps. The initial distribution, P_0 , represents the starting point of the random walk and can be defined as $P_0(X) = 1$ (indicating that the walk starts from Team X) and $P_0(\xi) = 0$ for all other nodes ξ . The transition probabilities from team X to its neighboring nodes v , taking into account the edge weights, can be calculated as follows:

$$p(X, v) = \frac{w(e)}{\sum_{e \in V_X} w(e)} \quad (11)$$

over all edges e represents the set of all edges connected to node X and V_X is the set of all neighbours of X . To calculate the likelihood of team X winning against team ξ , we can perform the random walk until a stopping criterion is met or until a maximum number of steps is reached. The probability of team X winning can be estimated using the Markov chain analysis. Let $P(X)$ represent the probability of team X winning a match. In a Markov chain analysis, this probability can be estimated by solving the following system of equations:

$$P(X) = C_X \cdot p(X, u_0) \cdot p(u_0, u_1) \cdot p(u_1, u_2) \cdots p(u_n, \xi) \quad (12)$$

$$= C_X \cdot p(X, u_0) \cdot \prod_{k=1}^n p(u_{k-1}, u_k) \cdot p(u_n, \xi) \quad (13)$$

Here, C_X represents the centrality score of team X , which reflects its importance within the network. The transition probability $p(X, u_i)$ represents the likelihood of moving from team X to team u_i , where $u_0, u_1, u_2, \dots, u_n$ represent the intermediate teams in the sequence leading to team ξ . The centrality score acts as a weight that enhances or diminishes the probability based on the importance of team X within the network. It adjusts the likelihood by either increasing or decreasing it based on the importance of team X within the network structure. This estimation assumes that the probability of team X winning is equivalent to the probability of reaching team ξ within the random walk simulation. It suggests that the likelihood of team X winning against team ξ can be inferred from the probability of ending up at team ξ after the random walk, considering the network structure and the relative strengths represented by the edge weights. This approach considers the evolving dynamics of the network structure, incorporating the connections and strengths represented by the edge weights. As a result, we gain valuable insights into the potential results of matches involving various nodes within the network. Choosing the stopping criteria for a random walk does not rely on a specific mathematical theorem, as it varies based on the problem or application being considered. In this scenario, the random walk can be terminated when a particular condition or target node is reached. The determination of the stopping criteria is a design choice that should align with the objectives and requirements of the specific problem or application.

In general, $P_t(u \rightarrow v)$ may not be equal to $P_t(v \rightarrow u)$. The transition probabilities depend on the network structure, edge weights, and other factors that influence the movement between nodes. Mathematically, a random walk from node v to node u at time t can be represented as $P_t(u \rightarrow v) = P_t(u, v)$ and similarly, a random walk from node v to node u at time t can be represented as $P(v \rightarrow u) = P_t(v, u)$. Due to the asymmetric nature of the edge weights in a bipartite graph (i.e. the vertices are divided into two distinct sets or partitions), the probabilities of a random walk from u to v are generally not equal to the probabilities from v to u . In most cases, the probabilities of a random walk from node u to node v and from node v to node u are not equal, unless specifically stated or under certain conditions.

| Teams | X | Y | Z |
|-------|------|------|------|
| X | - | 0.48 | 0.61 |
| Y | 0.51 | - | 0.64 |
| Z | 0.38 | 0.35 | - |

Table 2: Estimating the probabilities of each team's victory in matches against one another

Table 2 displays the computed probabilities, which indicate the likelihood of each team winning their

matches against one another. The results reveal that national team Y has a higher probability of achieving victory in all their matches compared to the other teams. This observation aligns with table 2, where it is evident that team Y holds a greater influence. The increased probability of success for team Y can be attributed to several factors. Firstly, their players are associated with highly ranked football clubs, implying a higher level of skill and performance. Additionally, it is noteworthy that several players from other national teams are also affiliated with the same club as the players from team Z , further contributing to their strength and potential for success.

Algorithm 2 Random Walk with Eigenvector Centrality Scores for Match Rating

Require: Graph G , Start node $node1$, Target node $node2$

Ensure: Probability estimates of team strengths

```

1:  $probabilities \leftarrow$  dictionary of nodes in  $G$  with initial values of 0
2:  $probabilities[node1] \leftarrow 1.0$ 
3: for  $i \leftarrow 1$  to number of iterations do
4:    $currentNode \leftarrow node1$ 
5:   for  $j \leftarrow 1$  to maximum number of steps do
6:      $neighbors \leftarrow$  neighbors of  $currentNode$ 
7:     if  $neighbors$  is empty then
8:       break
9:     end if
10:     $weights \leftarrow$  list of edge weights from  $currentNode$  to  $neighbors$ 
11:     $nextNode \leftarrow$  randomly select node from  $neighbors$  with probabilities based on  $weights$ 
12:     $probabilities[nextNode] += 1.0$ 
13:     $currentNode \leftarrow nextNode$ 
14:  end for
15: end for
16:  $total \leftarrow$  sum of values in  $probabilities$ 
17:  $probabilities \leftarrow$  dictionary with values divided by  $total$ 
18:  $node1winningprobability \leftarrow probabilities[node1] \times degree\_centrality[node1] / (probabilities[node1] \times degree\_centrality[node1] + probabilities[node2] \times degree\_centrality[node2])$ 
19:  $node2winningprobability \leftarrow probabilities[node2] \times degree\_centrality[node2] / (probabilities[node1] \times degree\_centrality[node1] + probabilities[node2] \times degree\_centrality[node2])$ 

```

4 Application of the Enhanced Ranking and Rating Methodology to World Cup 2022 and Euro 2020 Tournaments

The focus of this section is to introduce a sophisticated ranking and rating system designed to evaluate the performance of football teams participating in the FIFA World Cup 2022 and UEFA Euro 2020 tournaments. The objective is to provide a comprehensive and unbiased assessment of team performance by considering various factors such as match outcomes and opponent strength. The implementation of this methodology aims to deliver a thorough and impartial evaluation of team performance during these prestigious competitions. The methodology goes beyond simple win-loss records and considers the quality of opponents faced by each team. The implementation of this sophisticated ranking and rating system enhances the accuracy and fairness of evaluating team performance in the FIFA World Cup 2022 and UEFA Euro 2020 competitions. It provides valuable insights into the strengths and weaknesses of each team and contributes to a deeper understanding of their overall performance throughout the tournaments.

4.1 Predictive Analysis of the UEFA EURO 2020

The UEFA Euro 2020 tournament stands as one of the most prestigious and highly anticipated international football competitions in Europe. Originally scheduled for 2020, it was postponed due to unforeseen circumstances and eventually took place in 2021. The tournament brought together top national teams from across Europe to compete for the coveted title of European champions. Figure 4 represents the Euro 2020 tournament network. The network consists of nodes that represent different entities involved in the tournament. To establish connections between the players' nationalities and

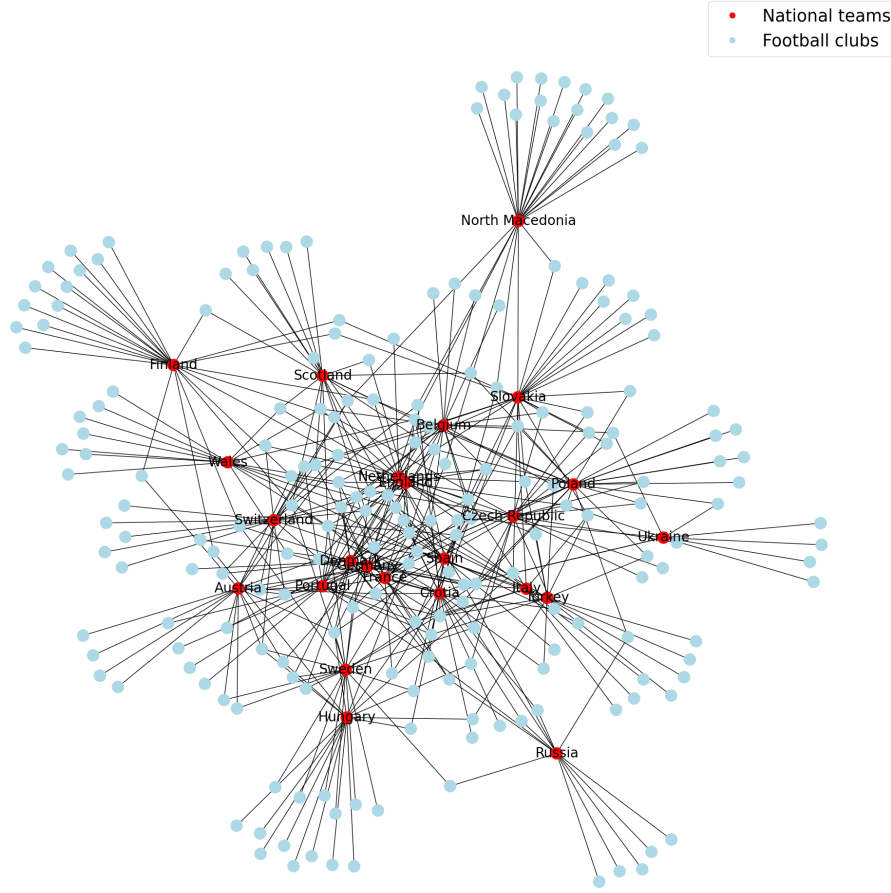


Figure 4: Euro 2020 tournament networks

their respective clubs, edges are formed in the network. An edge is created when a player represents a particular nationality and also plays for a specific club. This connection signifies the relationship between a player's national team and the club they are associated with. The network's edges also carry a weight associated with them, which is determined by the score of the football club in the FIFA club ranking. The higher the club's ranking, the stronger the connection or edge weight associated with that club in the network. The FIFA club ranking is a point-based method used to assess the performance and strength of football clubs around the world. It assigns scores to clubs based on their performance in various competitions over a specified period. Currently, Manchester City holds the top position in the FIFA club ranking with a score of 2070 points. They are followed closely by Bayern Munich, who have accumulated 1940 points, securing the second position. Real Madrid occupies the third spot in the ranking with a total of 1930 points. You can find the comprehensive ranking at the following <https://footballdatabase.com/>.

Figure 5 illustrates the degree distribution plot of the network. The x-axis corresponds to the degree of each node, representing the total weight of its edges. On the other hand, the y-axis represents the frequency of nodes with a particular degree. In a weighted graph, the degree of a node is determined by the cumulative sum of the weights of its edges. The degree distribution plot follows a power law distribution, indicating that the majority of players participating in the tournament are associated with lower-ranked teams compared to higher-ranked teams. This implies that there is a significant concentration of players from teams with lower overall rankings.

Higher centrality scores suggest stronger connections, greater influence, and more active participation in tournaments, indicating the potential for these teams to have a significant impact on match out-

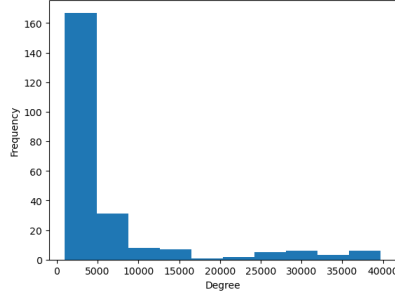


Figure 5: Degree distribution

























| Rank | | Country | Abbreviations | Centrality Score |
|------|---|-----------------|---------------|------------------|
| 1 |  | Spain | ESP | 0.193 |
| 2 |  | Denmark | DEN | 0.189 |
| 3 |  | England | ENG | 0.175 |
| 4 |  | Portugal | POR | 0.171 |
| 5 |  | Belgium | BEL | 0.159 |
| 6 |  | France | FRA | 0.158 |
| 7 |  | Netherlands | NED | 0.156 |
| 8 |  | Croatia | CRO | 0.148 |
| 9 |  | Poland | POL | 0.146 |
| 10 |  | Switzerland | SUI | 0.145 |
| 11 |  | Sweden | SWE | 0.142 |
| 12 |  | Germany | GER | 0.137 |
| 13 |  | Scotland | SCO | 0.132 |
| 14 |  | Slovakia | SVK | 0.111 |
| 15 |  | Italy | ITA | 0.110 |
| 16 |  | Wales | WAL | 0.110 |
| 17 |  | Austria | AUT | 0.104 |
| 18 |  | Turkey | TUR | 0.095 |
| 19 |  | Czech Republic | CZE | 0.086 |
| 20 |  | Finland | FIN | 0.068 |
| 21 |  | North Macedonia | NMD | 0.068 |
| 22 |  | Hungary | HUN | 0.051 |
| 23 |  | Russia | RUS | 0.051 |
| 24 |  | Ukraine | UKR | 0.048 |

Table 3: Estimating the eigenvector centrality score for all 24 teams based on 10,000 iterations and 0.01 tolerance

comes and the overall dynamics of the tournament network. Table 3 shows the centrality score for all the teams participating in the tournament. The centrality score was computed through 10,000 iterations with a tolerance of 0.01. The number of iterations in centrality score calculation (in this case, set to 10,000) is a parameter that affects the accuracy and precision of the computation. The power iteration method iteratively updates the centrality scores based on the adjacency matrix or the transition matrix of the network. The accuracy of the centrality scores improves with each iteration until it converges to a stable value. The tolerance value in centrality score calculation (set to 0.01 in this case) is a parameter that determines the convergence criterion for the iterative algorithm used to compute centrality measures. The tolerance value is used as a threshold to determine when the scores have converged sufficiently. If the difference in centrality scores between two consecutive iterations is smaller than 0.01, the algorithm assumes that the scores have converged and terminates the iteration process. We selected a convergence rate of 0.01 because the computation does not converge for values smaller than 0.01. Spain (0.193) has the highest centrality score among the given teams, indicating that they have a strong influence in the tournament network. This suggests that Spain is well-connected

with other teams, potentially indicating their participation in multiple tournaments or their ability to establish strong collaborations with other teams. Denmark holds the second-highest centrality score, signifying their notable involvement in the tournament network. Their strong connections with other teams indicate their active participation in multiple tournaments and their ability to influence the overall dynamics of the network. England possesses a relatively elevated centrality score, highlighting their significance and impact within the tournament network. Their strong connections with other teams indicate their involvement in multiple tournaments and their capability to influence match outcomes. The lower centrality score for teams in the tournament indicates their relatively lesser influence and connectivity within the network. These teams may have fewer interactions or connections with other teams, suggesting a lower impact on the overall dynamics of the tournament. Surprisingly, the winner of the tournament, Italy, was ranked 15th based on the eigenvector centrality score. At first glance, this might seem contradictory or counterintuitive. However, it's important to note that eigenvector centrality takes into account not only the direct connections of a team but also the connections of the teams it is connected to. Among the four semi-finalist teams, it is interesting to note that the centrality score assigned high values to three of them. However, it is noteworthy that neither of the two finalists, England and Italy, ranked among the top two teams with the highest centrality scores. This observation raises intriguing questions about the role of centrality in predicting tournament outcomes. While high centrality scores indicate strong connectivity and influence within the network, it does not guarantee a team's success in reaching the finals. Other factors, such as team strategies, individual player performances, match dynamics, and even elements of luck, can significantly impact a team's journey in the tournament.

The utilization of the random walk approach becomes essential in order to estimate the likelihood of different match outcomes. By applying this method, we can calculate the probabilities associated with each possible result, such as a win or loss for a particular team. Incorporating the random walk approach enhances our ability to make predictions based on opponent strength. This approach enables us to assess the potential outcomes of matches and provides valuable insights for betting, fantasy football, and strategic decision-making.

| | ESP | DEN | ENG | POR | BEL | FRA | NED | CRO | POL | SUI | SWE | GER | SCO | SVK | ITA | WAL | AUS | TUR | CZE | FIN | NMD | HUN | RUS | UKR |
|-----|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| ESP | - | 0.42 | 0.51 | 0.54 | 0.55 | 0.59 | 0.60 | 0.57 | 0.55 | 0.57 | 0.56 | 0.67 | 0.63 | 0.61 | 0.74 | 0.69 | 0.69 | 0.72 | 0.76 | 0.74 | 0.75 | 0.84 | 0.85 | 0.89 |
| DEN | 0.58 | - | 0.64 | 0.61 | 0.61 | 0.67 | 0.65 | 0.63 | 0.62 | 0.63 | 0.60 | 0.75 | 0.66 | 0.66 | 0.74 | 0.80 | 0.72 | 0.77 | 0.77 | 0.79 | 0.87 | 0.89 | 0.92 | |
| ENG | 0.49 | 0.36 | - | 0.5 | 0.50 | 0.56 | 0.54 | 0.53 | 0.51 | 0.53 | 0.53 | 0.64 | 0.58 | 0.58 | 0.73 | 0.67 | 0.67 | 0.68 | 0.69 | 0.73 | 0.72 | 0.82 | 0.85 | 0.87 |
| POR | 0.46 | 0.39 | 0.50 | - | 0.54 | 0.59 | 0.59 | 0.58 | 0.58 | 0.59 | 0.58 | 0.69 | 0.62 | 0.61 | 0.76 | 0.74 | 0.70 | 0.73 | 0.74 | 0.77 | 0.76 | 0.84 | 0.87 | 0.91 |
| BEL | 0.45 | 0.39 | 0.50 | 0.46 | - | 0.62 | 0.58 | 0.59 | 0.55 | 0.61 | 0.58 | 0.70 | 0.63 | 0.61 | 0.75 | 0.69 | 0.70 | 0.73 | 0.75 | 0.75 | 0.75 | 0.84 | 0.87 | 0.88 |
| FRA | 0.41 | 0.33 | 0.44 | 0.41 | 0.38 | - | 0.52 | 0.50 | 0.50 | 0.5 | 0.48 | 0.62 | 0.56 | 0.56 | 0.71 | 0.64 | 0.62 | 0.68 | 0.69 | 0.69 | 0.66 | 0.80 | 0.83 | 0.87 |
| NED | 0.40 | 0.36 | 0.46 | 0.41 | 0.42 | 0.48 | - | 0.51 | 0.51 | 0.54 | 0.53 | 0.64 | 0.60 | 0.58 | 0.70 | 0.64 | 0.65 | 0.68 | 0.69 | 0.69 | 0.71 | 0.79 | 0.83 | 0.87 |
| CRO | 0.43 | 0.37 | 0.47 | 0.42 | 0.41 | 0.50 | 0.49 | - | 0.55 | 0.54 | 0.57 | 0.68 | 0.61 | 0.61 | 0.73 | 0.69 | 0.65 | 0.70 | 0.72 | 0.72 | 0.69 | 0.82 | 0.85 | 0.89 |
| POL | 0.45 | 0.38 | 0.49 | 0.42 | 0.45 | 0.50 | 0.49 | 0.45 | - | 0.59 | 0.58 | 0.71 | 0.68 | 0.62 | 0.77 | 0.72 | 0.71 | 0.73 | 0.73 | 0.77 | 0.75 | 0.84 | 0.87 | 0.90 |
| SUI | 0.43 | 0.37 | 0.47 | 0.41 | 0.39 | 0.50 | 0.46 | 0.46 | 0.41 | - | 0.57 | 0.68 | 0.63 | 0.61 | 0.74 | 0.70 | 0.66 | 0.73 | 0.72 | 0.76 | 0.74 | 0.84 | 0.87 | 0.90 |
| SWE | 0.44 | 0.40 | 0.47 | 0.42 | 0.42 | 0.52 | 0.47 | 0.43 | 0.42 | 0.43 | - | 0.70 | 0.62 | 0.63 | 0.79 | 0.72 | 0.68 | 0.74 | 0.76 | 0.72 | 0.76 | 0.81 | 0.86 | 0.90 |
| GER | 0.33 | 0.25 | 0.36 | 0.31 | 0.30 | 0.38 | 0.36 | 0.32 | 0.29 | 0.32 | 0.30 | - | 0.48 | 0.47 | 0.63 | 0.56 | 0.52 | 0.59 | 0.61 | 0.63 | 0.63 | 0.71 | 0.77 | 0.83 |
| SCO | 0.37 | 0.34 | 0.42 | 0.38 | 0.37 | 0.44 | 0.40 | 0.39 | 0.32 | 0.37 | 0.38 | 0.52 | - | 0.58 | 0.66 | 0.64 | 0.66 | 0.70 | 0.72 | 0.68 | 0.72 | 0.82 | 0.84 | 0.89 |
| SVK | 0.39 | 0.34 | 0.42 | 0.39 | 0.39 | 0.44 | 0.42 | 0.39 | 0.38 | 0.39 | 0.37 | 0.53 | 0.42 | - | 0.75 | 0.73 | 0.67 | 0.7 | 0.68 | 0.69 | 0.69 | 0.81 | 0.86 | 0.89 |
| ITA | 0.25 | 0.20 | 0.27 | 0.24 | 0.25 | 0.29 | 0.30 | 0.27 | 0.23 | 0.26 | 0.21 | 0.37 | 0.26 | 0.25 | - | 0.55 | 0.53 | 0.52 | 0.55 | 0.58 | 0.58 | 0.71 | 0.75 | 0.79 |
| WAL | 0.31 | 0.26 | 0.33 | 0.26 | 0.31 | 0.36 | 0.36 | 0.31 | 0.28 | 0.30 | 0.28 | 0.48 | 0.34 | 0.27 | 0.45 | - | 0.59 | 0.64 | 0.68 | 0.66 | 0.65 | 0.78 | 0.80 | 0.86 |
| AUS | 0.31 | 0.28 | 0.33 | 0.30 | 0.30 | 0.38 | 0.35 | 0.35 | 0.29 | 0.34 | 0.32 | 0.44 | 0.34 | 0.33 | 0.47 | 0.41 | - | 0.67 | 0.65 | 0.61 | 0.64 | 0.74 | 0.83 | 0.87 |
| TUR | 0.28 | 0.23 | 0.32 | 0.27 | 0.27 | 0.32 | 0.32 | 0.30 | 0.27 | 0.27 | 0.26 | 0.41 | 0.30 | 0.30 | 0.48 | 0.36 | 0.33 | - | 0.62 | 0.64 | 0.75 | 0.8 | 0.85 | 0.85 |
| CZE | 0.24 | 0.23 | 0.31 | 0.26 | 0.25 | 0.31 | 0.31 | 0.28 | 0.27 | 0.28 | 0.24 | 0.39 | 0.28 | 0.32 | 0.45 | 0.32 | 0.35 | 0.38 | - | 0.6 | 0.6 | 0.74 | 0.79 | 0.82 |
| FIN | 0.26 | 0.23 | 0.27 | 0.23 | 0.25 | 0.31 | 0.31 | 0.72 | 0.23 | 0.24 | 0.28 | 0.37 | 0.32 | 0.31 | 0.42 | 0.34 | 0.39 | 0.38 | 0.4 | - | 0.68 | 0.74 | 0.84 | 0.86 |
| NMD | 0.26 | 0.21 | 0.28 | 0.24 | 0.25 | 0.34 | 0.29 | 0.31 | 0.25 | 0.26 | 0.24 | 0.37 | 0.28 | 0.31 | 0.42 | 0.35 | 0.36 | 0.36 | 0.4 | 0.32 | - | 0.82 | 0.88 | 0.89 |
| HUN | 0.16 | 0.13 | 0.18 | 0.16 | 0.16 | 0.20 | 0.21 | 0.18 | 0.16 | 0.16 | 0.19 | 0.29 | 0.18 | 0.19 | 0.29 | 0.22 | 0.26 | 0.25 | 0.26 | 0.26 | 0.18 | - | 0.73 | 0.78 |
| RUS | 0.15 | 0.11 | 0.15 | 0.13 | 0.13 | 0.17 | 0.17 | 0.15 | 0.13 | 0.13 | 0.14 | 0.23 | 0.16 | 0.14 | 0.25 | 0.20 | 0.17 | 0.2 | 0.21 | 0.16 | 0.12 | 0.27 | - | 0.73 |
| UKR | 0.11 | 0.08 | 0.13 | 0.09 | 0.12 | 0.13 | 0.13 | 0.11 | 0.10 | 0.10 | 0.10 | 0.17 | 0.11 | 0.11 | 0.21 | 0.14 | 0.13 | 0.15 | 0.18 | 0.14 | 0.11 | 0.22 | 0.27 | - |

Table 4: Estimating the Probabilities for each match outcome using the random walk approach

In Table 4, we utilize a random walk approach starting from a specific team, and perform it for a fixed number of iterations (1000). Each iteration of the random walk represents the number of times the random walk is repeated. The selection of the neighboring team is based on the weights assigned to the edges connecting them. In this case, the weights are determined by the normalized team club ranking scores of the teams. It is interesting to note that even though Spain has the highest centrality score, indicating its strong influence in the tournament, the probability of Spain winning against Denmark is in favor of Denmark with a probability of 0.58. This suggests that despite Spain's overall centrality, the dynamics of the random walk and the specific weights associated with the edges connecting Spain and Denmark contribute to this outcome. Likewise, we note a similar pattern in the case of Sweden and France, where the probability of Sweden winning against France is higher (0.52). These results highlight the significance of the random walk method and the assigned weights to the connections between teams in predicting the outcomes of matches. While the centrality scores offer valuable information about a team's overall significance in the network, there are additional factors to consider.

The specific connections between teams and their corresponding weights significantly influence the probabilities. By employing the random walk approach, we gain a comprehensive method to estimate these probabilities, enabling us to delve deeper into the possible results of football matches between teams in the tournament.

4.2 Analyzing Match Predictions for the FIFA World Cup 2022

The FIFA World Cup 2022 was scheduled to be the 22nd edition of the prestigious international football tournament. The 2022 World Cup marked a historic milestone as it was the first time the tournament was held in the Middle East. The tournament showcased the skills and talents of the world’s best football teams and players. It attracted immense global attention, with millions of fans eagerly supporting their favorite teams and players. The tournament served as a platform for nations to showcase their footballing prowess, unity, and national pride on the world stage. The FIFA World Cup 2022 captivated audiences worldwide, offering a blend of sporting excellence, cultural diversity, and passionate fan engagement. Historical data, team performance, player statistics, and various other relevant factors were utilized in previous models for predictive modeling of the World Cup 2022. These models employed statistical and machine learning techniques to analyze the available data and generate predictions regarding match outcomes and tournament results. By considering factors such as team form, past performance, player abilities, and tactical strategies, these models aimed to provide insights into the potential outcomes of the World Cup matches. The proposed model for predicting the World Cup outcomes introduces a novel approach that deviates from the conventional use of various covariates employed in previous models. Instead, it leverages the concepts of eigenvector centrality and random walk to rank the participating teams. In this model, the only covariate considered is the club ranking score associated with each football club that a player is affiliated with. Figure 6 illustrates the network representation of the FIFA World Cup 2022. In this network, the positioning of teams such as Costa Rica, Tunisia, and Qatar towards the end indicates that a significant number of their players do not have teammates from similar clubs participating in the tournament. This observation implies that these teams have a relatively lower level of connectivity and similarity in terms of the clubs their players belong to. It suggests that the players representing Costa Rica, Tunisia, and Qatar in the World Cup may come from diverse club backgrounds and may not have strong connections or shared experiences with each other compared to teams positioned closer to the center of the network. The network visualization offers valuable insights into the composition and connectivity patterns of teams participating in the World Cup. It provides a visual representation of the relationships between players and their respective clubs, highlighting the variations in team cohesion and shared experiences. By examining the positioning of teams within the network, we can gain a better understanding of the potential team dynamics and the level of familiarity among players representing different nations in the tournament.

The network representing the FIFA World Cup 2022 follows a power law distribution. This means that there is a significant variation in the number of connections or relationships that teams have, and the distribution of these connections follows a specific pattern. The power law distribution in the World Cup 2022 network suggests that there are a few teams with a high number of connections and a higher club ranking score. These teams are likely to have players from prestigious and successful clubs, indicating their strong presence and influence in the network. They may have a higher level of connectivity with other teams due to the shared club affiliations of their players. On the other hand, the majority of teams in the network have fewer connections and lower club ranking scores. These teams may have players from less prominent or lower-ranked clubs, resulting in relatively fewer connections and a lower level of influence in the network.

4.2.1 Assessing the Team Rankings in the FIFA World Cup 2022

In this section, we delve into the rankings of the teams participating in the FIFA World Cup 2022. The performance and positioning of each team play a crucial role in determining their success in the tournament. By analyzing various factors and utilizing a weighted graph representation, we aim to gain insights into the relative strengths and rankings of the teams. By exploring this ranking framework, we can better understand the competitive landscape of the World Cup and make informed predictions about team performance.

The centrality score of a team in the network indicates its level of influence and importance within the

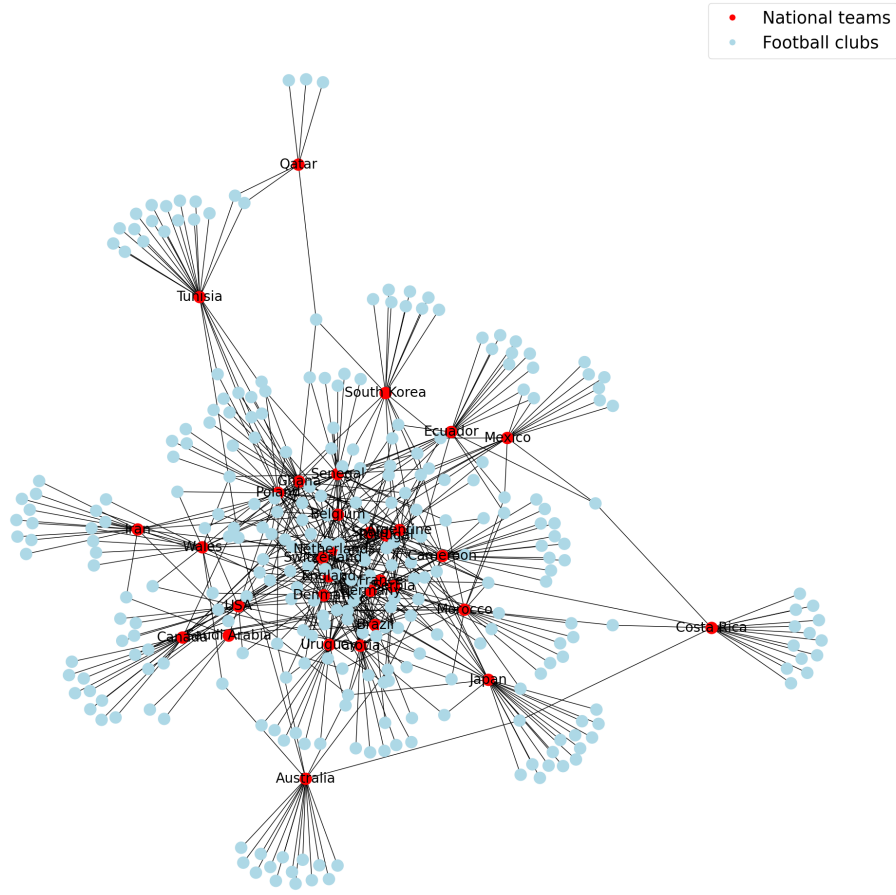


Figure 6: World cup 2022 tournament networks

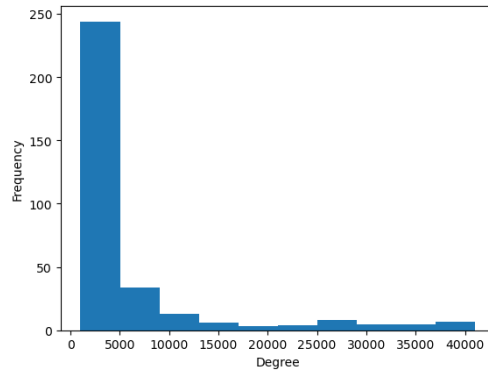


Figure 7: World cup 2022 tournament networks

context of the FIFA World Cup 2022. Teams with higher centrality scores are more well-connected and play a significant role in the overall dynamics of the tournament. France, Brazil, Argentina, and Croatia are identified as having the highest centrality scores in the network. This suggests that these teams have strong connections with other teams participating in the World Cup. These teams are considered key players in the tournament and are expected to have a significant impact on match outcomes. On the other hand, Tunisia, Saudi Arabia, and Qatar have the lowest centrality scores in the network. This indicates that these teams have fewer connections with other teams in the tournament.





| Rank | | Country | Centrality Score |
|------|---|--------------|------------------|
| 1 |  | France | 0.199 |
| 2 |  | Brazil | 0.180 |
| 3 |  | Croatia | 0.154 |
| 4 |  | Argentina | 0.154 |
| 5 |  | Uruguay | 0.154 |
| 6 |  | Denmark | 0.141 |
| 7 |  | Belgium | 0.141 |
| 8 |  | Germany | 0.139 |
| 9 |  | Spain | 0.137 |
| 10 |  | England | 0.126 |
| 11 |  | Portugal | 0.125 |
| 12 |  | Switzerland | 0.124 |
| 13 |  | Netherlands | 0.117 |
| 14 |  | Senegal | 0.113 |
| 15 |  | USA | 0.094 |
| 16 |  | Poland | 0.080 |
| 17 |  | Ghana | 0.077 |
| 18 |  | Morocco | 0.076 |
| 19 |  | Cameroon | 0.069 |
| 20 |  | Japan | 0.067 |
| 21 |  | Serbia | 0.053 |
| 22 |  | Wales | 0.050 |
| 23 |  | South Korea | 0.039 |
| 24 |  | Canada | 0.034 |
| 25 |  | Ecuador | 0.029 |
| 26 |  | Mexico | 0.026 |
| 27 |  | Iran | 0.021 |
| 28 |  | Costa Rica | 0.015 |
| 29 |  | Australia | 0.009 |
| 30 |  | Tunisia | 0.007 |
| 31 |  | Saudi Arabia | 0.003 |
| 32 |  | Qatar | 0.001 |

Table 5: Estimating the eigenvector centrality score for all 32 teams based on 10,000 iteration and 0.01 tolerance

These teams are considered to have relatively lower influence and may face challenges in competing against teams with higher centrality scores. Understanding the centrality scores of teams provides valuable insights into their positions within the World Cup network and helps gauge their potential performance.

4.2.2 Predictions and Ratings for the Round of 16 to the Final Matches of the 2022 World Cup

In this section, we present the predictions and ratings for the Round of 16 to the Final matches of the 2022 World Cup. These predictions are based on a random walk approach, which takes into account the centrality score in Table 5. By simulating a sequence of steps within the tournament network, the random walk algorithm allows us to estimate the probabilities of match outcomes and provide insights into the potential results of each match. Through this analysis, we aim to offer a glimpse into the exciting journey of the World Cup, providing fans and enthusiasts with a predictive perspective of the potential path to the coveted Final. In the prediction results, 3 out of the 4 (75%) teams predicted correctly reached the semi-finals of the FIFA 2022 tournament. In the quarter-finals, 5 out of 8 (62.5%) teams were predicted accurately, and in the round of 16, 12 out of the 16 (75%) teams were correctly predicted and 50% accuracy for the finals. The model successfully identified France as one



Figure 8: Bracket of the 2022 World Cup Tournament: Round of 16 to Final

of the tournament favorites, and they reached the final, which was ultimately decided through penalty shootouts. The model exhibits a minimum accuracy rate of 50% for each stage of the competition. This level of accuracy is noteworthy considering the inherent uncertainties and unpredictability of football matches. While no model can guarantee perfect predictions, achieving a minimum accuracy rate of 50% demonstrates the model's ability to capture key patterns and trends in the data. It provides valuable insights into the potential outcomes of the tournament and can serve as a useful tool for fans, analysts, and betting companies. By achieving a minimum accuracy threshold, the model establishes itself as a reliable and informative resource for assessing team performance and predicting match results. Its performance across multiple stages of the competition highlights its robustness and potential practical applications in evaluating and understanding the dynamics of the FIFA World Cup 2022. The predictions were based on the random walk approach and the eigenvector centrality score. One of the most surprising predicted matches was England versus Senegal. Although England had a higher centrality score, the random walk approach indicated that Senegal had the highest probability of winning with a probability of 0.59. This highlights the unpredictability of football matches and the limitations of any model. As George Box famously said, 'All models are wrong, but some are useful. Despite the inherent uncertainties, these match ratings provide a general understanding of the tournament dynamics and can assist betting companies in defining their odds. Furthermore, they offer valuable insights for team coaches to develop new tactics for each game.

4.2.3 Group Stage Predictions and Ratings for the 2022 World Cup

In this section, we present predictions and ratings for the Group Stages of the 2022 World Cup. Using our proposed model, we have generated predictions for the outcomes of matches and assessed the performance of each team in their respective groups. These predictions and ratings provide valuable insights into the potential outcomes and competitiveness of the Group Stage matches, offering a deeper understanding of the tournament dynamics.

In order to evaluate the performance of our model, it is essential to compare the predicted match outcomes with the actual results of the 2022 World Cup matches. The tournament consisted of a total of 68 matches, spanning from the group stages to the final. Among the 48 group stage matches, our proposed model accurately predicted 27 matches, resulting in a prediction accuracy of 58.33%. These statistics provide valuable insights into the performance of our model in terms of its ability to accurately forecast match results in the World Cup.

| Group A | | | | | |
|---------|------|------|------|------|------|
| | | | | | Rank |
| | - | 0.50 | 0.80 | 1.00 | 1 |
| | 0.50 | - | 0.87 | 1.00 | 2 |
| | 0.20 | 0.14 | - | 1.00 | 3 |
| | 0.00 | 0.00 | 0.00 | - | 4 |

| Group B | | | | | |
|---------|------|------|------|------|------|
| | | | | | Rank |
| | - | 0.49 | 0.68 | 0.85 | 1 |
| | 0.51 | - | 0.78 | 0.92 | 2 |
| | 0.32 | 0.78 | - | 0.85 | 3 |
| | 0.15 | 0.08 | 0.15 | - | 4 |

| Group C | | | | | |
|---------|------|-------|------|------|------|
| | | | | | Rank |
| | - | 0.65 | 0.90 | 1.00 | 1 |
| | 0.45 | - | 0.85 | 0.99 | 2 |
| | 0.10 | 0.015 | - | 0.97 | 3 |
| | 0.00 | 0.01 | 0.03 | - | 4 |

| Group D | | | | | |
|---------|------|------|------|------|------|
| | | | | | Rank |
| | - | 0.55 | 0.97 | 0.97 | 1 |
| | 0.45 | - | 0.96 | 0.97 | 2 |
| | 0.03 | 0.04 | - | 0.78 | 3 |
| | 0.03 | 0.08 | 0.22 | - | 4 |

| Group E | | | | | |
|---------|------|-------|------|------|------|
| | | | | | Rank |
| | - | 0.50 | 0.59 | 0.92 | 1 |
| | 0.50 | - | 0.61 | 0.92 | 2 |
| | 0.41 | 0.039 | - | 0.95 | 3 |
| | 0.08 | 0.08 | 0.05 | - | 4 |

| Group F | | | | | |
|---------|------|------|------|------|------|
| | | | | | Rank |
| | - | 0.65 | 0.77 | 0.99 | 1 |
| | 0.35 | - | 0.66 | 0.82 | 2 |
| | 0.23 | 0.34 | - | 0.78 | 3 |
| | 0.01 | 0.14 | 0.18 | - | 4 |

| Group G | | | | | |
|---------|------|------|------|------|------|
| | | | | | Rank |
| | - | 0.60 | 0.72 | 0.81 | 1 |
| | 0.40 | - | 0.56 | 0.78 | 2 |
| | 0.28 | 0.44 | - | 0.74 | 3 |
| | 0.19 | 0.22 | 0.26 | - | 4 |

| Group H | | | | | |
|---------|------|------|------|------|------|
| | | | | | Rank |
| | - | 0.68 | 0.69 | 0.88 | 1 |
| | 0.32 | - | 0.66 | 0.80 | 2 |
| | 0.31 | 0.34 | - | 0.80 | 3 |
| | 0.12 | 0.20 | 0.20 | - | 4 |

Table 6: Match Ratings of the Group Stage Matches in the 2022 FIFA World Cup

5 Conclusion and Future Works

In this study, we introduced a novel approach based on graph theory to rank and predict the outcomes of international football competitions. To construct the network representation, we represented football clubs and national teams as nodes, and established edges between them when a player was associated with a club and also represented their national team. This network framework allowed us to capture the relationships between clubs and national teams, considering the overlapping participation of players. In addition to player connections, we integrated the FIFA ranking score of each football club as an important factor in our analysis. The FIFA ranking score provides a quantitative measure of a club’s performance and standing in the football community. By incorporating this covariate, we aimed to assess the influence and reputation of each national team participating in the tournament. To determine the ranking and favoritism of teams, we assigned centrality scores to all the participating teams. Centrality scores serve as measures of prominence or importance within a network. In our context, the centrality score represented the significance or impact of a team based on its connections to other clubs and national teams through shared players. The team with the highest centrality score was deemed the tournament favorite, suggesting that it had the strongest connections and associations with other influential clubs and national teams. This approach allowed us to identify the teams that held a significant position within the network, implying a higher level of influence and potential success in the competition. By considering the centrality scores of all participating teams, we gained valuable insights into the relative influence and standing of each team within the network. This information played a crucial role in our prediction model, enabling us to make informed forecasts regarding match outcomes and overall tournament results.

To forecast match results, we employed a random walk approach, which leveraged the normalized the club ranking score as transition probabilities during the walk. By normalizing the club ranking scores, we ensured that they represented probabilities and could be used to guide the random walk process effectively. During the random walk, the model traversed the network representation by following

edges based on the assigned probabilities. Using this approach, we were able to compute the probabilities of each team winning a match. These probabilities indicated the likelihood of a team emerging victorious in a given match, taking into consideration its centrality score and the influence derived from its connections within the network.

Initially, our model was primarily focused on FIFA World Cup tournaments. However, recognizing the potential of the approach, we also explored its applicability to other tournaments beyond the scope of the FIFA World Cup. This expansion broadened the utility and versatility of our model, allowing for its adaptation and application to different football competitions.

In the Euro 2020 competition, we tested the model’s performance on 51 matches. The results showed a prediction accuracy of 54.90%. While the match predictions were not consistently accurate, the model successfully captured surprises such as Denmark advancing to the semi-finals. Subsequently, we applied the model to the FIFA World Cup 2022, which consisted of 68 matches from the group stages to the final. Among the 48 group stage matches, our model accurately predicted 27 matches, representing a prediction accuracy of 58.33%. In the knockout stage, out of the 16 matches, 10 were correctly predicted, resulting in a prediction accuracy of 62.50%. Overall, the model achieved a probability of 60.41% for correctly predicting match outcomes.

Although our model showed improvements in prediction accuracy, it is essential to consider the limitations of our approach. Match outcomes in football are influenced by numerous unpredictable factors, including individual player performances, team strategies, injuries, and even referee decisions. These factors may not have been fully accounted for in our analysis, potentially impacting the accuracy of our predictions. It is worth noting that our model’s predictions serve as a general understanding of the tournament dynamics and can be utilized by betting companies to define odds. Additionally, the insights provided by the model can assist team coaches in developing new tactics and strategies for each game. Further research and refinement of the model are necessary to enhance its prediction accuracy. Incorporating additional variables, such as player injuries and recent form of the national teams, could potentially improve the model’s ability to capture the intricacies of match outcomes.

Moreover, by combining the proposed model with either statistical models or advanced machine learning techniques, we can enhance the accuracy and predictive capabilities. The incorporation of statistical models allows us to consider additional factors and variables that may influence match outcomes, such as historical data, team statistics, and player performance metrics. Similarly, the integration of advanced machine learning algorithms enables us to leverage the power of computational models to analyze vast amounts of data and extract meaningful patterns. By combining the proposed graph-based model with statistical models or advanced machine learning techniques, we can harness the strengths of both approaches. The graph-based model provides a unique perspective by considering the network structure and centrality measures, while statistical models and machine learning algorithms offer the ability to incorporate a wider range of features and optimize the prediction process. This combination creates a robust and comprehensive framework that has the potential to significantly enhance the accuracy and reliability of match outcome predictions.

Appendix

A Convergence of the Power Method

Theorem A.1. (Perron-Frobenius). Let $A \in \mathbb{R}^{n \times n}$ be an irreducible nonnegative matrix with complex eigenvalues $\kappa_1, \dots, \kappa_n \in \mathbb{C}$. The Perron-Frobenius theorem states the following properties:

1. Among the eigenvalues with maximum modulus, there exists at least one eigenvalue μ of A that is real and positive. In other words, $\mu \in \mathbb{R}$ and $\mu \geq \max_{i=1}^n |\kappa_i| \geq 0$.
2. If A is symmetric and the support of A (denoted as $\text{Supp}(A)$) is not bipartite, then the eigenvalue with maximum modulus is unique, has multiplicity 1, and the corresponding eigenvector has nonnegative components.

Proof. We prove the statements separately

1. Let $\kappa_1, \kappa_2, \dots, \kappa_n$ be the complex eigenvalues of A . We assume without loss of generality that $|\kappa_1| \geq |\kappa_2| \geq \dots \geq |\kappa_n|$. Since A is an irreducible nonnegative matrix, it follows from the Perron-Frobenius theory that the spectral radius $\rho(A)$ is an eigenvalue of A and is positive. Therefore, $|\kappa_1| \leq \rho(A)$. Now, let \mathbf{v} be an eigenvector corresponding to κ_1 (the eigenvalue with maximum modulus). Since A is nonnegative, we can assume that \mathbf{v} has nonnegative components. If all components of \mathbf{v} are zero, then κ_1 is not an eigenvalue of A , which contradicts our assumption. Hence, there exists at least one component of \mathbf{v} that is positive. Now, consider the matrix power A^k . Since A is nonnegative and irreducible, it follows that A^k is positive (all entries are positive). Hence, there exists a positive entry in the vector $A^k \mathbf{v}$. This implies that κ_1^k is positive. Since κ_1 is a complex eigenvalue, κ_1^k can have positive real and imaginary parts. However, if κ_1^k has a nonzero imaginary part, then $|\kappa_1^k|$ cannot tend to infinity as k tends to infinity, which contradicts the fact that $|\kappa_1|^k$ becomes arbitrarily large. Therefore, κ_1^k must be real and positive. Since κ_1^k is a power of κ_1 , it follows that κ_1 itself is real and positive. Thus, property 1 is proven.

2. If A is symmetric, then all eigenvalues of A are real. Suppose there exist two distinct eigenvalues κ_1 and κ_2 with maximum modulus. Let \mathbf{v}_1 and \mathbf{v}_2 be the corresponding eigenvectors of κ_1 and κ_2 , respectively.

Since κ_1 and κ_2 are distinct, the eigenvectors \mathbf{v}_1 and \mathbf{v}_2 must be linearly independent. Moreover, for a symmetric matrix, eigenvectors corresponding to distinct eigenvalues are orthogonal. Therefore, \mathbf{v}_1 and \mathbf{v}_2 are orthogonal eigenvectors.

Now, consider the matrix $B = \mathbf{v}_1 \mathbf{v}_1^T$. Since \mathbf{v}_1 has nonnegative components, it follows that B is a nonnegative matrix. Moreover, B is symmetric because $\mathbf{v}_1 \mathbf{v}_1^T = (\mathbf{v}_1 \mathbf{v}_1^T)^T$.

Now, let $\mathbf{w} = A \mathbf{v}_2$. Since \mathbf{v}_2 is an eigenvector of A , we have $A \mathbf{v}_2 = \kappa_2 \mathbf{v}_2$. Therefore, $\mathbf{w} = \kappa_2 \mathbf{v}_2$. Consider the matrix product BA . We have $BA \mathbf{v}_2 = B(\kappa_2 \mathbf{v}_2) = \kappa_2 B \mathbf{v}_2$. On the other hand, $BA \mathbf{v}_2 = B(\kappa_2 \mathbf{v}_2) = \kappa_2 \mathbf{v}_1 (\mathbf{v}_1^T \mathbf{v}_2)$. Since \mathbf{v}_1 and \mathbf{v}_2 are orthogonal, we have $\mathbf{v}_1^T \mathbf{v}_2 = 0$. Therefore, $\kappa_2 \mathbf{v}_1 (\mathbf{v}_1^T \mathbf{v}_2) = \mathbf{0}$.

Since $BA \mathbf{v}_2 = \mathbf{0}$, it follows that $\kappa_2 B \mathbf{v}_2 = \mathbf{0}$. But since $\mathbf{w} = \kappa_2 \mathbf{v}_2$, we have $B \mathbf{w} = \mathbf{0}$. This implies that \mathbf{w} is orthogonal to \mathbf{v}_1 .

However, if \mathbf{w} is orthogonal to \mathbf{v}_1 , it implies that $\kappa_2 \mathbf{v}_2$ is orthogonal to \mathbf{v}_1 , which contradicts the fact that \mathbf{v}_1 and \mathbf{v}_2 are linearly independent. Therefore, the assumption that there exist two distinct eigenvalues with maximum modulus is false. Hence, there is only one eigenvalue with maximum modulus, and it has multiplicity 1. The corresponding eigenvector has nonnegative components.

□

Theorem A.2. Let A be a symmetric nonnegative matrix with eigenvalues $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_n$. Then the eigenvalue with the largest absolute magnitude is κ_1 , and $\kappa_1 \geq 0$. In other words, $|\kappa_i| \leq \kappa_1$ for all i .

Proof. We start by observing that A has n real eigenvalues because it is symmetric. Let $\mu \in \mathbb{R}$ and $w \in \mathbb{R} \setminus 0$ be any eigenpair of A , i.e., $Aw = \mu w$. Since $w^T w > 0$, we can write:

$$|\mu| w^T w = |\mu w^T w| = |w^T A w| \leq \sum_{i,j} |A_{1,j} w_i w_j| = \sum_{i,j} A_{1,j} |w_i| |w_j| = x^T A x$$

where x has components $|w_i|$. In the above inequality, we have used the triangle inequality: $|a + b| \leq |a| + |b|$. Rewriting the inequality, we have:

$$|\mu| \leq \frac{x^T A x}{w^T w} = \frac{x^T A x}{x^T x}$$

Let $x = \sum_i c_i v_i$, where v_i are the normalized eigenvectors of A and c_i is any constant. Substituting this into the equation, we get:

$$\frac{x^T A x}{x^T x} = \frac{\sum_j c_j v_j^T A \sum_i c_i v_i}{\sum_j c_j v_j^T \sum_i c_i v_i} = \frac{\sum_j c_j v_j^T \sum_i c_i \kappa_i v_i}{\sum_j c_j v_j^T \sum_i c_i v_i} = \frac{\sum_i c_i^2 \kappa_i}{\sum_i c_i^2} \leq \frac{\sum_i c_i^2 \kappa_1}{\sum_i c_i^2} = \kappa_1$$

It is worth noting that equality is possible only when $c_1 = 1$ and $c_i = 0$ for $i \neq 1$. In this case, we have $|\mu| \leq \frac{x^T A x}{x^T x} \leq \kappa_1$, which means that κ_1 is larger than the absolute value of any eigenvalue of A . \square

The eigenvalue κ_1 is also called the dominant eigenvalue of A . Note that, for the power method to converge, we require $|\kappa_i| < \kappa_1$ for $i \neq 1$.

B Estimating the Required Number of Iterations for the Power Method

The power method can be implemented in time complexity $O(m + n)$, which, assuming the network is connected, simplifies to $O(m)$. Here, m represents the number of iterations required to converge to an eigenvalue or eigenvector of interest, and n is the size of the matrix. However, a crucial question arises: how many iterations are necessary to achieve a small error, say less than some ϵ ?

Let us consider v_1 as the centrality eigenvector. We can measure the error using the formula:

$$\left\| \frac{x(t)}{c_1 \kappa_1^t} - v_1 \right\|$$

where $x(t) \in \mathbb{R}^n$ represents the eigenvector at time t during the iteration (evolves as the iteration progresses), v_i corresponds to the i -th normalized eigenvector of matrix A , κ_1 is the maximum eigenvalue and c_i are appropriate constant. $\|\cdot\|_2$ denotes the Euclidean norm of a vector, given by $\|\mathbf{y}\|_2 = \sqrt{\sum_{i=1}^n y_i^2}$. Since the eigenvectors of matrix A form a basis for \mathbb{R}^n , we can express the initial eigenvector of our iteration as $x(0) = \sum_i c_i v_i$. Then, at time t , we have:

$$x(t) = A^t x(0) = \sum_i c_i \kappa_i^t v_i = \kappa_1^t \sum_i c_i \left(\frac{\kappa_i}{\kappa_1} \right)^t v_i \quad (14)$$

Alternatively, this equation can be written as

$$\frac{x(t)}{c_1 \kappa_1^t} = v_1 + \frac{c_2}{c_1} \left(\frac{\kappa_2}{\kappa_1} \right)^t v_2 + \dots + \frac{c_n}{c_1} \left(\frac{\kappa_n}{\kappa_1} \right)^t v_n$$

Thus, the error at time t can be bounded above by:

$$\sum_{i \neq 1} \left| \frac{c_i}{c_1} \right| \left(\frac{|\kappa_i|}{\kappa_1} \right)^t \|v_i\| = \sum_{i \neq 1} \left| \frac{c_i}{c_1} \right| \left(\frac{|\kappa_i|}{\kappa_1} \right)^t \leq \left(\frac{\kappa'}{\kappa_1} \right)^t \sum_{i \neq 1} \left| \frac{c_i}{c_1} \right| \quad (15)$$

Therefore, $\kappa' = \max(|\kappa_2|, |\kappa_n|)$. In order to achieve an error of at most ϵ , we need:

$$t \geq \Omega \left(\frac{\log(1/\epsilon)}{\log(\kappa_1/\kappa')} \right)$$

Proof. Let κ_1 be the maximum eigenvalue of matrix A and κ' be the maximum magnitude of the remaining eigenvalues, i.e., $\kappa' = \max(|\kappa_2|, |\kappa_3|, \dots, |\kappa_n|)$. Using the properties of the power method, we know that the error at time t can be bounded above by:

$$\sum_{i \neq 1} \left| \frac{c_i}{c_1} \right| \left(\frac{|\kappa_i|}{\kappa_1} \right)^t \|v_i\| \leq \left(\frac{\kappa'}{\kappa_1} \right)^t \sum_{i \neq 1} \left| \frac{c_i}{c_1} \right|$$

To achieve an error of at most ϵ , we need the above expression to be less than or equal to ϵ . Therefore, we have:

$$\left(\frac{|\kappa'|}{\kappa_1}\right)^t \sum_{i \neq 1} \left| \frac{c_i}{c_1} \right| \leq \epsilon$$

Taking the logarithm on both sides, we get:

$$t \log \left(\frac{|\kappa'|}{\kappa_1} \right) + \log \left(\sum_{i \neq 1} \left| \frac{c_i}{c_1} \right| \right) \leq \log(\epsilon)$$

Simplifying the equation, we have:

$$t \geq \frac{\log(\epsilon)}{\log \left(\frac{|\kappa'|}{\kappa_1} \right)} - \frac{\log \left(\sum_{i \neq 1} \left| \frac{c_i}{c_1} \right| \right)}{\log \left(\frac{|\kappa'|}{\kappa_1} \right)}$$

Since κ' and κ_1 are fixed properties of the matrix A , and ϵ is a small positive number, the above expression provides a lower bound for the number of iterations t required to achieve an error of at most ϵ .

Therefore, we can conclude that:

$$t \geq \Omega \left(\frac{\log(1/\epsilon)}{\log(\kappa_1/\kappa')} \right)$$

□

In practice, the power method is iterated until the variation between $\frac{x(t+1)}{|x(t+1)|}$ and $\frac{x(t)}{|x(t)|}$ becomes negligibly small.

C How to Chose $x(0)$

The equation 14 indicates that the power method is ineffective if $x(0)$ is such that $c_1 = 0$, meaning that the initial vector $x(0)$ is orthogonal to v_1 . To prevent this issue, we deliberately choose $x(0)$ in a way that ensures all its components are positive. By applying the following theorem, we can guarantee that $v_1^T \cdot x(0) > 0$, thus confirming that $x(0)$ will not be orthogonal to v_1 .

Theorem C.1. *For a symmetric nonnegative matrix A with eigenvalues $\kappa_1 \geq \kappa_2 \geq \dots \geq \kappa_n$, it can be assumed without loss of generality that all components of the dominant eigenvector v_1 (associated with the dominant eigenvalue κ) are nonnegative.*

D Time-reversible Markov chains

We examine positive recurrent Markov chains $\{X_n : n \geq 0\}$, which, when in a steady-state (stationarity), exhibit the property that the reversed time process follows the same Markov chain (in terms of distribution). The key requirement for this property is that, for any pair of states i and j , the long-term rate at which the chain transitions from state i to state j is equal to the long-term rate at which it transitions from state j to state i : $\pi_i P_{i,j} = \pi_j P_{j,i}$, where π_i and π_j represent the steady-state probabilities of being in states i and j respectively, and $P_{i,j}$ and $P_{j,i}$ represent the transition probabilities from state i to state j and from state j to state i respectively.

D.1 Two-sided stationary extensions of Markov chains

Consider a positive recurrent Markov chain $X_n : n \in \mathbb{N}$ with a transition probability matrix P and a stationary distribution π . We are interested in extending this process to include negative time values as well, resulting in $X'_n : n \in \mathbb{Z}$. This two-sided extension allows us to assume that the chain started in the finite past. To achieve this extension, we begin by shifting the origin to time $k \leq 1$ and

extending the process k time units into the past. We define $X'_n(k) = X'_{n+k}$ for $-k \leq n < \infty$. It is important to note that $X'_n(k) : n \in \mathbb{N}$ has the same distribution as $X'_n : n \in \mathbb{N}$ due to stationarity. Furthermore, this extension on $-k \leq n \leq \infty$ remains stationary. As we let k approach infinity, the process $X'_n(k) : -k \leq n < \infty$ converges in distribution to a truly two-sided extension, which we denote as $X'_n : n \in \mathbb{Z}$. Importantly, this two-sided extension remains stationary. For each time $n \in \mathbb{Z}$, we have the property that $P(X'_n = j) = \pi_j$, indicating that the probability of observing state j at time n is equal to the stationary probability π_j . This extension of the Markov chain to include negative time values allows us to explore the behavior of the process in both the past and the future. It provides a comprehensive view of the system's dynamics and facilitates the analysis of various properties and behaviors through time.

D.2 Time-reversibility: Time-reversibility equations

Consider the two-sided extension $X'_n : n \in \mathbb{Z}$ of a positive recurrent Markov chain with transition matrix P and stationary distribution π . The Markov property states that "the future is independent of the past given the present state," which can be equivalently stated as "the past is independent of the future given the present state." This implies that the process $X_n^{(r)} = X'_{-n}$ for $n \in \mathbb{N}$, representing the process in reverse time, is also a stationary Markov chain. Remarkably, this reversed process has transition probabilities that can be computed in terms of π and P . Let $P(r) = (P_{i,j}(r))$ denote the time-reversed transition probabilities:

$$\begin{aligned} P_{i,j}(r) &= P(X'_1 = j | X'_0 = i) = P(X'_0 = j | X'_1 = i) \\ &= \frac{P(X'_1 = i | X'_0 = j)P(X'_0 = j)}{P(X'_1 = i)} \\ &= \frac{\pi_j}{\pi_i} P_{j,i} \end{aligned} \tag{16}$$

This expression shows that $P_{i,j}(r)$, the probability of transitioning from state i to state j in the reversed process, is related to π_i and π_j (the stationary probabilities) as well as $P_{j,i}$ (the original transition probability from state j to state i). By computing these time-reversed transition probabilities, we gain insights into the behavior of the reversed process and its relationship to the original Markov chain.

Definition D.1. A Markov chain is said to be time-reversible if it exhibits the same distribution in reverse time as it does in forward time. Specifically, a positive recurrent Markov chain with transition matrix P and stationary distribution π is considered time-reversible if the reverse-time stationary Markov chain $X_n^{(r)} : n \in \mathbb{N}$ has an identical distribution to the forward-time stationary Markov chain $X'_n : n \in \mathbb{N}$. This property can be expressed through the equation $P(r) = P$, where $P(r)$ denotes the transition probabilities of the reverse-time chain, and P represents the transition matrix of the original chain. Alternatively, time-reversibility can be characterized by the following set of equations:

$$\pi_i P_{i,j} = \pi_j P_{j,i} \tag{17}$$

These equations hold for all pairs of states i and j . Intuitively, these equations signify that the long-run rate at which the chain transitions from state i to state j is equal to the long-run rate at which it transitions from state j to state i , with the probabilities π_i and π_j serving as weights for each transition.

Proposition D.1. Let P be the transition matrix of an irreducible Markov chain. If there exists a probability solution π to the "time-reversibility" set of equations,

$$\pi_i P_{i,j} = \pi_j P_{j,i} \tag{18}$$

for all pairs of states i and j , then the Markov chain is positive recurrent and time-reversible. Furthermore, the solution π is a unique stationary distribution for the chain.

References

1. Jungnickel, D. (2014). *Graphs, Networks and Algorithms*. Springer Berlin, Heidelberg 978-3-642-32277-8
2. Moreno, L. J. (1946). *The Sociometric View of the Community*. The Journal of Educational Sociology, 19(9), 540–545. <https://doi.org/10.2307/2263771>
3. Bavelas, A. (1948). *A mathematical model for group structures*. Applied Anthropology, 7(3), 16–30
4. Bavelas, A. (1950). *Communication patterns in task-oriented groups*. The Journal of the Acoustical Society of America, 22(6), 725–730.
5. Bonacich, P. (1972). *Factoring and weighting approaches to status scores and clique identification*. Journal of Mathematical Sociology, 2(1), 113–120.
6. Barenbaum, N. B. (2000). *How social was personality? The Allports’ “connection” of social and personality psychology*. Journal of the History of the Behavioral Sciences, 36(4), 471–487..
7. Bonacich, P. (2007). *Some unique properties of eigenvector centrality*. Social Networks, 29(4), 555–564.
8. Borgatti, S. P. (2006). *Identifying sets of key players in a social network*. Computational & Mathematical Organization Theory, 12, 21–34.
9. Borgatti, S. P., & Everett, M. G. (2006). *A graph-theoretic perspective on centrality*. Social networks, 28(4), 466–484.
10. Faust, K. (1997). *Centrality in affiliation networks*. Social networks, 19(2), 157–191.
11. Fay, D., Kunegis, J., & Yoneki, E. (2013). *Centrality and mode detection in dynamic contact graphs; a joint diagonalisation approach*. In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (pp. 41–48).
12. Albers, P. C., & Vries, H. D. (2001). *Elo-rating as a tool in the sequential estimation of dominance strengths*. Animal Behaviour, 489–495.
13. Ólafsson, S. (1990). *Weighted matching in chess tournaments*. Journal of the Operational Research Society, 41(1), 17–24.
14. Dixon, M. J., & Coles, S. G. (1997). *Modelling association football scores and inefficiencies in the football betting market*. Journal of the Royal Statistical Society: Series C (Applied Statistics), 46(2), 265–280.
15. Dyte, D., & Clarke, S. R. (2000). *A ratings based Poisson model for World Cup soccer simulation*. Journal of the Operational Research Society, 51(8), 993–998.
16. Enos, B. (2012). *Practical Shooting, Beyond Fundamentals*. Loose Cannon.
17. Karlis, D., & Ntzoufras, I. (2009). *Bayesian modelling of football outcomes: using the Skellam’s distribution for the goal difference*. IMA Journal of Management Mathematics, 20(2), 133–145.
18. Groll, A., Ley, C., Schauburger, G., & Van Eetvelde, H. (2019). *A hybrid random forest to predict soccer matches in international tournaments*. Journal of Quantitative Analysis in Sports, 15(4), 271–287.
19. Groll, A., Hvattum, L. M., Ley, C., Popp, F., Schauburger, G., Van Eetvelde, H., & Zeileis, A. (2021). *Hybrid machine learning forecasts for the UEFA EURO 2020*. arXiv preprint arXiv:2106.05799.
20. Groll, A., Ley, C., Schauburger, G., Van Eetvelde, H., & Zeileis, A. (2019). *Hybrid machine learning forecasts for the FIFA Women’s World Cup 2019*. arXiv preprint arXiv:1906.01131.

21. Ley, C., Wiele, T. V. D., & Eetvelde, H. V. (2019). *Ranking soccer teams on the basis of their current strength: A comparison of maximum likelihood approaches*. Statistical Modelling, 19(1), 55-73.
22. Van der Wurp, H., & Groll, A. (2022). *Introducing LASSO-type penalisation to generalised joint regression modelling for count data*. AStA Advances in Statistical Analysis, 1-25.
23. Aldous, D., & Fill, J. (1995). *Reversible Markov chains and random walks on graphs*.
24. Chung, F., & Zhao, W. (2010). *PageRank and random walks on graphs*. Fete of Combinatorics and Computer Science, 43-62.
25. Razali, N., Mustapha, A., Yatim, F. A., & Ab Aziz, R. (2017, August). *Predicting football matches results using Bayesian networks for English Premier League (EPL)*. In Iop Conference Series: Materials Science and Engineering (Vol. 226, No. 1, p. 012099). IOP Publishing.
26. Byshevets, N., Denysova, L., Shynkaruk, O., Serhiyenko, K., Usychenko, V., Stepanenko, O., & Syvash, I. R. Y. N. A. (2019). *Using the methods of mathematical statistics in sports and educational research*.