Project Report - ECE 227

Gayatri Kharche
Electrical and Computer Engineering
University of California San Diego
A69027290

Atharv Sunil Biradar
Electrical and Computer Engineering
University of California San Diego
A69026845

Abstract

This project aims to analyze historical international football match data to gain insights into team performance, interactions, and network structures. By processing and transforming the dataset into node and edge tables, we can answer a variety of questions related to victories, losses, win/loss ratios, match interactions, and tournament competitiveness. Through rigorous data preprocessing, including the exclusion of draws and early historical matches, the study identifies 167 teams with substantial match histories for detailed analysis. Employing network analysis techniques and visualization tools, such as graph theory and community detection algorithms like Louvain, and utilizing platforms like NetworkX for graph construction and Gephi for advanced visualization, the research explores team interactions, community structures, and the evolution of competitive landscapes. Continent-specific visualizations provide deeper insights into regional football dynamics and historical rivalries. The findings aim to enrich understanding of international football's competitive dynamics and historical evolution, catering to analysts, historians, and enthusiasts interested in sports analytics.

1 Introduction

International football represents a global tapestry of history, competition, and cultural significance, making it an ideal domain for studying complex network dynamics. This project utilizes a vast dataset spanning more than a century of matches to deeply explore team performances, interactions, and community formations within this sport. By focusing exclusively on competitive fixtures and excluding non-representative matches, the analysis ensures that insights gleaned are pertinent and insightful for understanding the evolution of international football's competitive landscape.

The primary aim of this research is to uncover the underlying structures of the international football match network using advanced graph theoretical techniques and state-of-the-art visualization tools. This involves constructing a detailed representation of the network where teams are nodes and match outcomes are edges, allowing for the examination of how teams interact and compete over time. Beyond mapping out the topological properties and centrality measures of this network, the study delves into community structures across continents. This exploration sheds light on regional football dynamics, historical rivalries, and the broader global patterns that shape the sport.

Through meticulous data preprocessing and the application of sophisticated analytical methods, such as graph theory and community detection algorithms like Louvain, the research aims to offer actionable insights into various aspects of international football. This includes understanding team strategies, identifying pivotal moments in historical matches, and tracing the evolution of competitive behaviors. These insights not only contribute to scholarly discourse in sports analytics but also hold practical implications for sports management and strategy development.

Ultimately, this project endeavors to enrich our understanding of how international football has evolved as a global phenomenon. By dissecting the intricate network of match outcomes and team interactions, it aims to provide a comprehensive view of the sport's competitive dynamics across different eras and continents, benefiting analysts, historians, and enthusiasts alike in their exploration of sports data and its implications.

2 Background and Literature Review

The paper "Dynamic Network Analysis of the Euro2016 Final Preliminary Results[6]" explores the application of network analytics techniques to analyze the passing distributions of both Portugal and France in the Euro2016 Final. The study compares network analysis results with traditional performance indicators like passes and shooting attempts to explain Portugal's victory. The authors use a dataset from Opta Sports and employ tools such as the Gephi Consortium and NetworkX Python library for visualization and calculation of network metrics. The research builds on existing literature on network analysis in football, highlighting the need for more research to establish the relationship between network structure and team performance. The study presents preliminary findings on the passing dynamics of the Euro2016 Final, showcasing minor differences between the two teams and emphasizing the importance of considering substitute players and temporal changes in network structures for a comprehensive analysis.

This paper [9] explores the application of Network Analysis to evaluate the relationships within a professional football team, emphasizing that team performance is influenced by on-pitch interactions among players. The study's goal is to develop a model that assists management in making informed decisions about player assessment and acquisition. The methodology combines qualitative and quantitative approaches, with data gathered through direct observation and analyzed using Ucinet 6.4 software. The empirical investigation focuses on a UEFA Champions League match, demonstrating the potential of Network Analysis to evaluate team dynamics.The findings highlight the capacity of Network Analysis to provide a quantitative framework for assessing cooperation among team members, which can inform tactical development in football. This innovative approach offers a new tool for measuring team relationships, contributing valuable insights into the strategic management of football teams.In his blog post, Graham delves into the process of visualizing football match data using the network analysis tool Gephi. By leveraging Opta data from a Manchester City match, he meticulously imports details about players and their passing interactions into Gephi. This results in the creation of sophisticated network graphs that vividly illustrate the connections and passing patterns between players on the field. Utilizing metrics such as Eigenvector Centrality, Graham is able to pinpoint key players and crucial pass routes, offering an in-depth look at player influence and team dynamics. The blog underscores the significant potential of network graphs in football analytics, particularly for understanding the intricate web of interactions that underpin team performance. This method provides valuable insights that can be used for strategic planning and performance optimization in professional football.Graham's detailed walkthrough not only highlights the technical steps involved in creating these visualizations but also emphasizes the broader implications of this approach for enhancing tactical analysis and decision-making in football. The use of network analysis in this context represents a powerful tool for coaches, analysts, and management, enabling them to make more informed decisions based on a deeper understanding of player roles and team interactions.

The paper titled "Social Network Analysis in Team Sports [5]" authored by Filipe Manuel Clemente, Fernando Manuel Lourenço Martins, and Rui Sousa Mendes provides a comprehensive overview of applying Social Network Analysis (SNA) in the context of team sports. The authors delve into the definitions and concepts of SNA, observational tools for data collection in team sports, and present a case study on Argentina's network analysis during the FIFA World Cup 2014.The case study includes detailed methods such as sample selection, data collection, processing, variables studied, and statistical procedures employed. The results, discussion, and practical applications derived from the analysis offer valuable insights for understanding the network dynamics within a sports team. The authors acknowledge the support received from various institutions and individuals in conducting this research.Overall, this work serves as a significant contribution to the field of sports science, highlighting the importance of social network analysis in unraveling the intricate relationships and interactions within team sports.

3 Dataset and Data Description

The project mentioned in [3] employs an extensive dataset documenting international football results from 1872 to 2024, comprising over 47,000 matches. This dataset includes results from a diverse range of competitions, such as the FIFA World Cup, the FIFI Wild Cup, and regular friendly matches. It is specifically focused on men's full internationals and excludes Olympic Games matches or those involving a nation's B-team, U-23, or league select teams.The primary dataset, 'results.csv', provides detailed information about each match. The columns in this dataset include the date of the match, the names of the home and away teams, the full-time scores of both teams (including extra time but excluding penalty shootouts), the tournament name, the city and country where the match was played, and an indicator of whether the match was played at a neutral venue.In addition to 'results.csv', the dataset includes two supplementary files: 'shootouts.csv' and 'goalscorers.csv'. The 'shootouts.csv' file details matches that were decided by penalty shootouts, listing the date, home team, away team, the shootout winner, and the team that shot first. The 'goalscorers.csv' file records the individual goal scorers, including the match date, home and away teams, the team that scored, the player's name, and whether the goal was an own goal or a penalty. For consistency, the dataset uses the current names of teams for historical matches. For example, matches played by the 1882 Ireland team are recorded under Northern Ireland, the modern successor. Conversely, country names reflect their historical context at the time of the match. This approach ensures that home matches are accurately represented, even if the country's name has since changed, as indicated by the 'neutral' column.This rich dataset forms the foundation of our analysis, enabling us to explore international football team performance, interactions, and network structures. It serves as a valuable resource for football analysts, historians, and enthusiasts, providing deep insights into the dynamics of international football over a substantial historical period.

4 Methodology

4.1 Data Preprocessing

The data preprocessing for this project involved several key steps to transform the raw dataset into a structured format suitable for analysis. Initially, the dataset containing international football match results from 1872 to 2024 was loaded, and unnecessary columns were removed. Matches classified as "Friendly" were excluded to focus on more competitive fixtures. Additionally, matches before 1921 and those that ended in a draw were removed to ensure the analysis was relevant and meaningful.

Next, the dataset was segmented to identify teams with a substantial number of matches. Teams that played at least 100 matches were selected for further analysis, resulting in a list of 167 teams. Goals scored and conceded were aggregated for these teams, and their win and loss records were calculated. This comprehensive team statistics table included columns for matches played, goals for and against, wins, losses, win rate, and loss rate.

To analyze team interactions, match outcomes were determined, identifying the winning and losing teams for each match. These outcomes were filtered to include only the selected teams, and the resulting data was used to create an edge list representing team interactions. Each edge in the list indicated a match between two teams, with a weight reflecting the number of times the match-up occurred. The final nodes and edges tables, representing team statistics and interactions respectively, were saved as CSV files for subsequent network analysis.

4.2 Network Analysis and Visualization

The methodology for the network analysis and visualization in this project involved transforming raw match data into a structured network graph, identifying community structures, and visually representing these structures using both Python libraries and Gephi. Initially, the raw data was preprocessed to filter relevant matches and prepare necessary dataframes for nodes and edges. Matches categorized as 'Friendly' were excluded to focus on competitive games. Additionally, matches occurring after 1930 were included to ensure a more modern dataset, and only matches with decisive outcomes (non-tied results) were considered. This preprocessing step ensured that the data used in the analysis was both relevant and comprehensive.

4.2.1 Construction of Directed Graph

A directed graph was then constructed using the networkx library, with nodes representing teams and edges representing the matches. Each node was enriched with attributes such as matches played, goals scored, goals conceded, wins, losses, win rate, and loss rate. Directed edges were added from the losing team to the winning team, with the weight of each edge indicating the number of times the losing team lost to the winning team. This detailed representation provided a robust foundation for the subsequent network analysis.

4.2.2 Community Detection and Visualization

To identify community structures within the network, the directed graph was converted to an undirected graph, and the Louvain method was applied to detect modularity classes, representing the community clusters within the network. Each node was assigned a modularity class, and a color map was created to visually distinguish these classes. This step was crucial for understanding the underlying community structures and the relationships between different teams.

The graph was visualized using a spring layout in Python, where nodes were colored based on their modularity class and edges were colored according to the source node's color. This visualization highlighted community structures and relationships between teams, providing a clear and insightful representation of the network.

4.2.3 Advanced Visualization with Gephi

Additionally, the network data, including nodes and edges with their respective attributes and colors, was exported to CSV files and imported into Gephi for further visualization and analysis. Gephi was used to plot the graph of nodes and edges, offering advanced visualization features and tools for a more detailed and interactive exploration of the network.

4.2.4 Continent-wise Visualization

To enhance the analysis further, continent-wise visualizations were created by filtering the data based on geographical regions. Subgraphs were generated for each continent using Gephi's filtering capabilities. These continent-specific graphs maintained the modularity class color-coding, enabling a comparison of community structures, network densities, and team interactions within different continents. By comparing these subgraphs, differences in football dynamics across continents were observed, offering a nuanced understanding of the sport's global and regional intricacies. This granular approach provided deeper insights into regional rivalries and the competitive landscape within each continent, enriching the overall network analysis and visualization.

4.3 Additional Analysis

In addition to examining basic topological properties and centrality measures, this study delved deeper into the structural dynamics and community behavior of the international football match network.

4.3.1 Bridge Analysis and Triadic Closure

Utilizing the undirected graph representation, the analysis focused on identifying bridges, critical edges whose removal could potentially disrupt the network's connectivity. Bridges provide insights into key vulnerabilities and strategic points of influence within the international football landscape. Concurrently, triadic closure analysis assessed the network's tendency to form closed loops or triangles, indicative of higher clustering and cohesive substructures. This measure highlighted patterns of mutual engagement and competition dynamics among teams.

4.3.2 Strongly Connected Components

Since the initial graph is directed, the concept of strongly connected components (SCCs) was examined. SCCs are subsets of the graph where every node is reachable from every other node within the same subset. The number and sizes of SCCs indicate the robustness and flow of influence or competition dynamics within different subsets of teams.

4.3.3 Community Detection Algorithms

Several community detection algorithms were applied to uncover cohesive groups of teams based on their interactions and performances within the network. The Louvain method, renowned for its efficiency in detecting modularity-based communities, partitioned the graph into distinct clusters maximizing intra-cluster connections while minimizing inter-cluster links. Complementarily, the Label Propagation Algorithm (LPA) identified communities through iterative node label propagation, capturing subtle community structures and affiliations. Additionally, the application of the Kernighan-Lin algorithm offered insights into potential bipartite structures within the network, providing a nuanced understanding of community formations in specific contexts.

4.4 Topology and Centrality Measures

The study also explored key topological properties and centrality measures to characterize the international football match network comprehensively. Beyond the average degree and density, which quantify the overall connectivity and compactness of the network, centrality measures such as degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and PageRank centrality were computed for each node. These measures provided insights into the relative importance, influence, and positioning of teams within the network. Analysis of centrality measures identified pivotal teams that serve as key connectors (via high betweenness centrality) or central influencers (via high eigenvector centrality and PageRank) in the international football landscape.

These analyses collectively contributed to a nuanced understanding of the structural attributes, community dynamics, and influential nodes within the international football match network. By examining bridges, triadic closure, SCCs, and employing diverse community detection algorithms and centrality measures, this study advanced insights into the complex interactions and competitive dynamics shaping global football interactions.

5 Results

5.1 Graph Representation

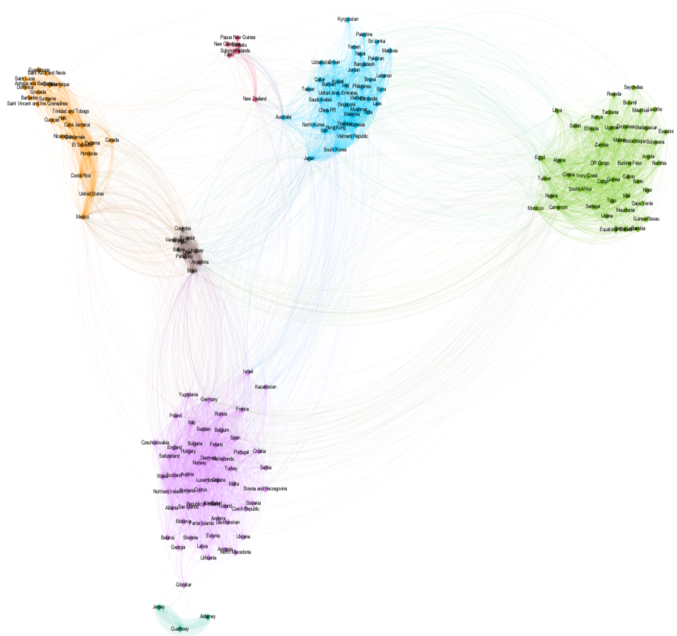

Figure 1: Directed Graph Construction using Gephi

Figure 1 illustrates the network divided into distinct clusters, each represented by a different color. Nodes within the same cluster exhibit higher connectivity compared to nodes in different clusters. This graph was created using Gephi, while Figure 2 shows a graph plotted using NetworkX.

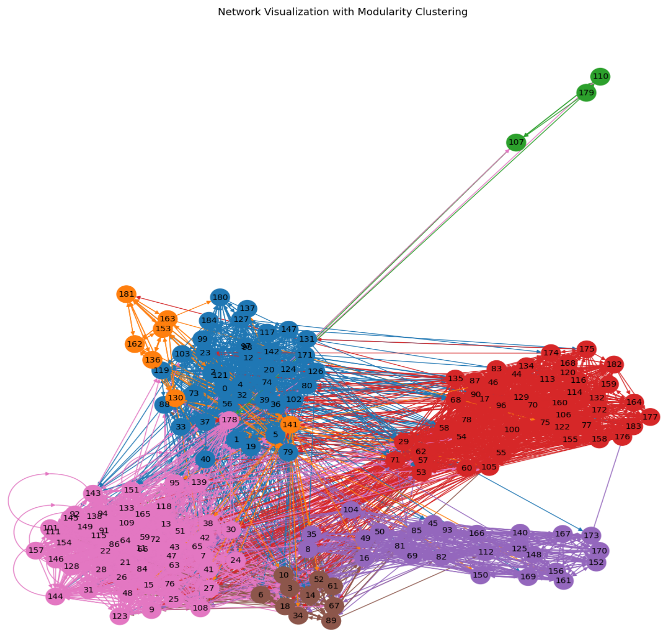

Figure 2: Network Visualization with Networkx Python library

5.1.1 Asia

In Asia, Japan and South Korea are among the most successful countries.

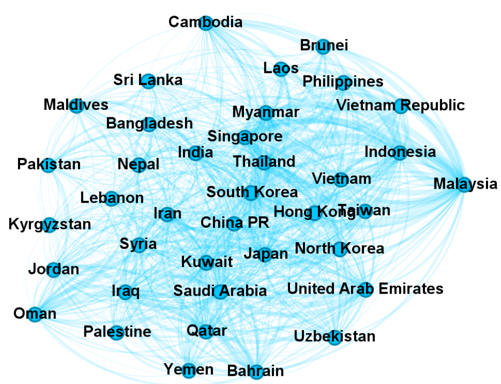

Figure 3: Continent-wise Subgraphs in Gephi - Asia

5.1.2 Africa

In Africa, Egypt leads with an in-degree of 50, followed by Morocco, Cameroon, Nigeria, and Ghana. These teams have won the Africa Cup of Nations the most times.

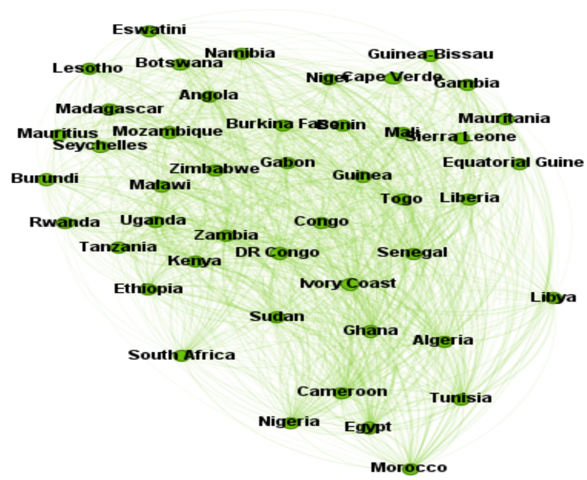

Figure 4: Continent-wise Subgraphs in Gephi - Africa

5.1.3 Europe

As observed, Spain, Germany, and France have the highest in-degrees, at 69, 67, and 66 respectively, reflecting their status as the most successful European national teams. Notably, the graph shows Yugoslavia and Czechoslovakia, two countries that no longer exist, positioned at the northern boundary. In contrast, the countries that emerged from their dissolution are located at the southern boundary. This logical arrangement is accurately depicted in the graph.

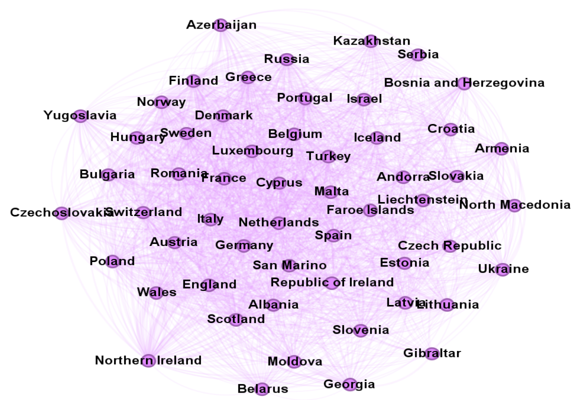

Figure 5: Continent-wise Subgraphs in Gephi - Europe

5.1.4 North America

In North America, Mexico has the highest in-degree of 47, followed by the United States with an in-degree of 41. Mexico is traditionally regarded as a strong team and has consistently reached the Round of 16 in every World Cup since 1994.

Figure 6: Continent-wise Subgraphs in Gephi - North America

5.1.5 Oceania

As expected, in Oceania, New Zealand stands out as the most successful country, followed by Jersey. The other countries in the region have struggled to win any matches.

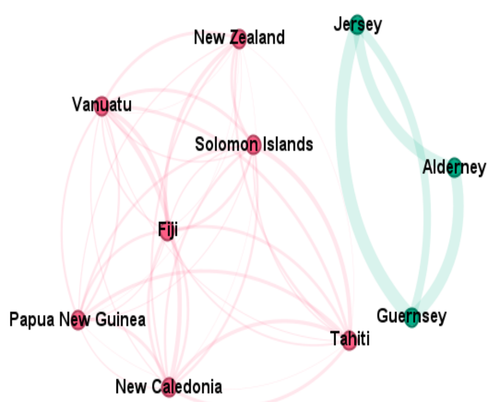

Figure 7: Continent-wise Subgraphs in Gephi - Oceania

5.1.6 South America

In South America, Brazil is clearly the most successful team with an in-degree of 59, followed by Argentina and Uruguay. This representation is logical, as these three teams are undeniably the most successful in the region, boasting the most wins and titles.

Figure 8: Continent-wise Subgraphs in Gephi - South America

5.2 Fundamental Network Characteristics

It is crucial to grasp the fundamental topological characteristics of our network. Our network comprises 185 nodes (representing countries) interconnected by 5,628 edges. With a network density of 0.165, around 16.5% of all potential node connections are established. The average clustering coefficient of 0.683 indicates a notable tendency for nodes to form clusters. Additionally, the network forms a single connected component, ensuring that all nodes are interconnected within a unified structure.

- Nodes: 185
- Edges: 5,628
- Network Density: 0.165 (approximately 16.5% of all possible connections between nodes are realized)
- Average Clustering Coefficient: 0.683 (indicating a high tendency for nodes to cluster together)
- Connected Components: Single (all nodes are part of a unified structure)

5.3 Degree Measures

To further comprehend the structure of our network, we examine various degree measures. The average degree of the network is 60.84, indicating that each node is connected to approximately 61 other nodes on average. The node with the highest degree is node 5, which has 121 connections, signifying its central role within the network. Regarding in-degree, node 30 stands out with 73 in-degrees, representing the highest number of incoming connections.

When considering edge weights, the average weighted degree is significantly higher at 210.69. Node 0 has the highest weighted degree of 527, emphasizing its prominence when edge weights are taken into account. Similarly, node 1 has the highest weighted in-degree, with a substantial value of 374 weighted in-degrees.

Additionally, we have included plots for the degree distribution (Figure9 and Figure 10) and the weighted degree distribution. These visualizations illustrate that most nodes have a degree between 25 and 95 and a weighted degree between 75 and 400, providing a clearer picture of the network's connectivity and the distribution of connection strengths.

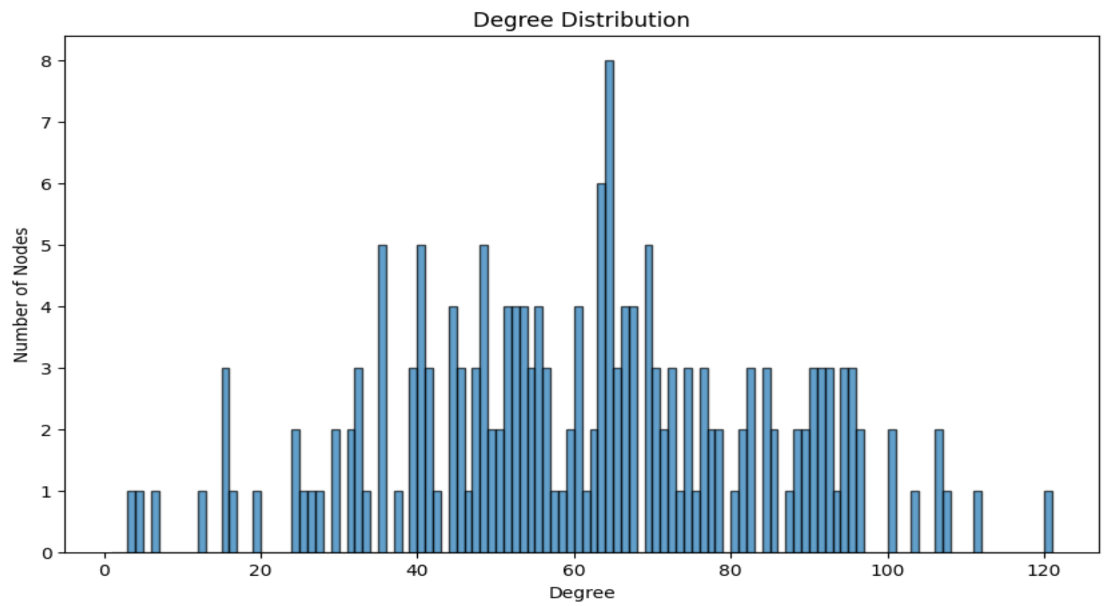

Figure 9: Degree Distribution

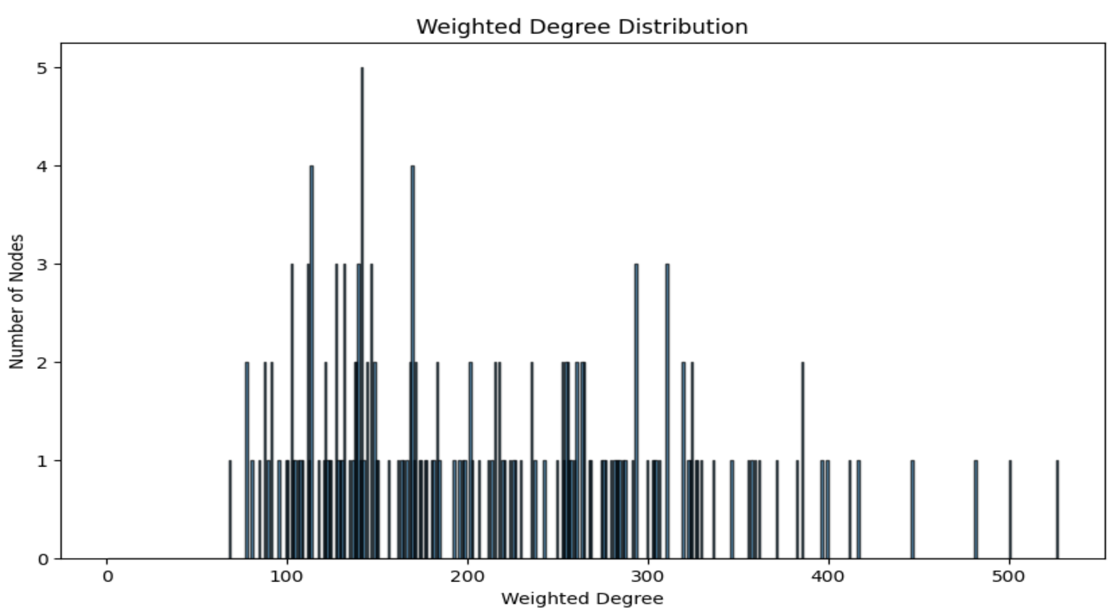

Figure 10: Weighted Degree Distribution

5.4 Centrality Measures

5.4.1 Degree Centrality

Degree Centrality measures how well-connected a node is within the network. For our analysis, it indicates the number of distinct matches won or lost against different opponents. The top 10 countries based on their Degree Centrality are led by Japan and South Korea. Notably, Japan has a higher in-degree than South Korea, reflecting more victories.

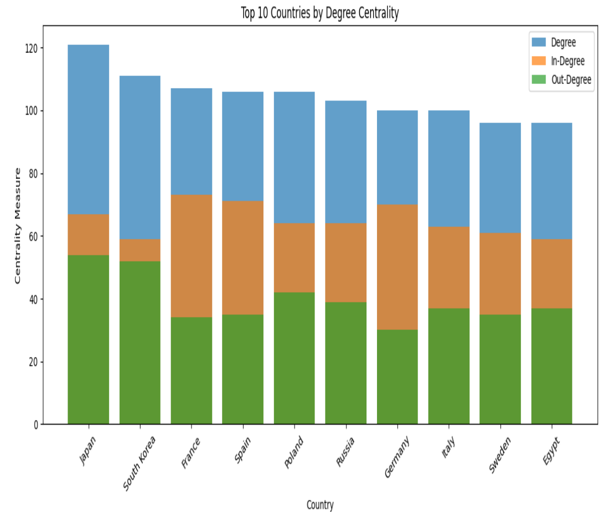

Figure 11: Degree Centrality

5.4.2 Betweenness Centrality

Betweenness Centrality highlights the importance of a node in connecting other nodes, essentially identifying countries that frequently lie on the shortest paths between others. The top 10 countries in this metric include South Korea, Mexico, Japan, and the United States. These countries are critical connectors, linking continents such as Asia and Africa to the football powerhouses in Europe and South America. Consequently, their removal would isolate many smaller countries from the major footballing nations.

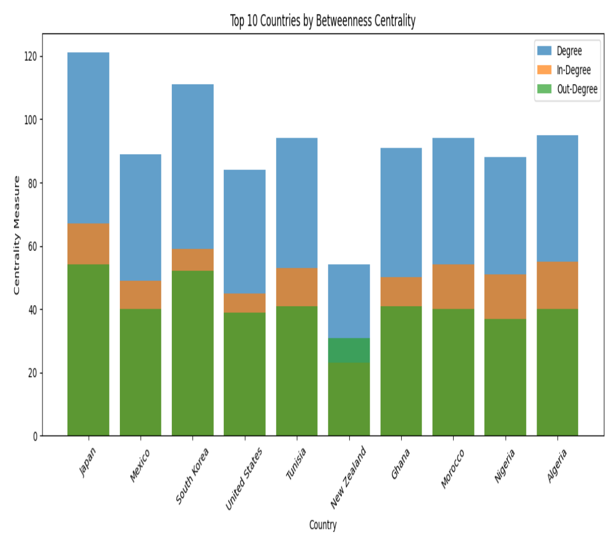

Figure 12: Betweenness Centrality

5.4.3 Closeness Centrality

Closeness Centrality measures how quickly a node can reach other nodes, indicating proximity to the network's center. The top 10 countries in this category are again dominated by those inside the primary football continents of Europe and South America, like Germany and Italy, emphasizing their central role in the global football network.

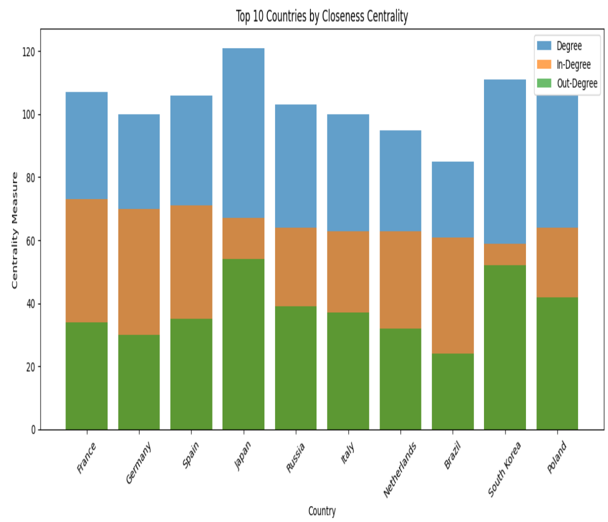

Figure 13: Closeness Centrality

5.4.4 Eigenvector Centrality

Eigenvector Centrality assesses a node's connections to other important nodes, providing a weighted degree centrality with a boost for connections to significant countries. This measure reveals that traditionally successful football nations such as Spain, France, Germany, and Italy lead, as they have numerous connections with other important national teams.

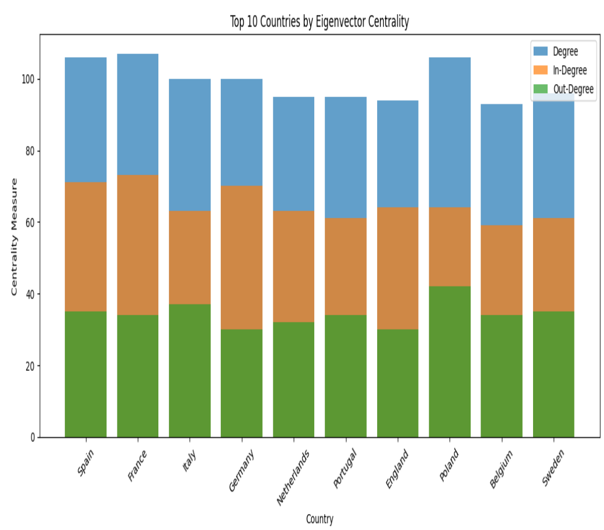

Figure 14: Eigenvector Centrality

5.4.5 Pagerank Centrality

PageRank Centrality, a variant of Eigenvector Centrality, considers the direction of links. It assigns scores based on the number of incoming links and the relative importance of the originating nodes. The top 10 countries by PageRank Centrality include Argentina, France, Germany, and England, aligning closely with the most successful national teams as recognized by football fans, highlighting their top performances over the century.

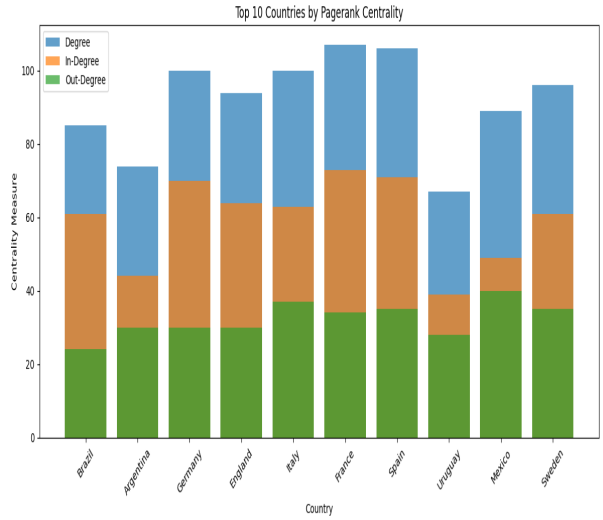

Figure 15: Pagerank Centrality

5.5 Network Analysis Report

The analysis of our network reveals several key topological properties and characteristics:

- **Network Size and Density:**
  - Nodes: 185
  - Edges: 5,628
  - Graph Density: 0.165 (approximately 16.5% of all possible connections between nodes are realized)
- **Clustering and Triadic Closure:**
  - Average Clustering Coefficient: 0.566
  - Triadic Closure: 0.580
- **Path Length and Diameter:**
  - The graph is not strongly connected, thus the average shortest path length and diameter cannot be calculated due to infinite path lengths in disconnected components.
- **Triangles and Bridges:**
  - Number of Triangles in the Graph: 37,861
  - No bridges found in the undirected graph.
- **Strongly Connected Components:**
  - Number of Strongly Connected Components: 3

In summary, the network exhibits characteristics of both cohesion and fragmentation. While it demonstrates significant local clustering and triadic closure, the presence of multiple strongly connected components suggests partitioning and potentially distinct communities within the network.

This analysis provides a foundational understanding of the network's structure and connectivity, essential for further exploration and interpretation of its dynamics and relationships.

6 Community Detection Analysis

6.1 Louvain Method

The Louvain Method, employed for community detection in our network, is known for optimizing modularity to identify cohesive groups of nodes. Applying this method to our dataset, we found that it identified a total of 7 distinct communities within the network. This algorithm is effective in revealing intricate community structures by maximizing internal connectivity and minimizing connections between different communities.

6.2 Label Propagation Algorithm (LPA)

The Label Propagation Algorithm (LPA) is a straightforward yet powerful approach for community detection. It propagates labels based on local connectivity patterns, effectively partitioning the network into communities where nodes share similar labels. In our analysis, the LPA identified 5 communities, demonstrating its capability to uncover clusters based on local information propagation.

6.3 Kernighan-Lin Algorithm

The Kernighan-Lin Algorithm, particularly suited for bipartite graphs, iteratively optimizes the partitioning of the graph into two balanced sets by minimizing the edge cut. When applied to our network, which was transformed into an undirected graph for compatibility, the Kernighan-Lin Algorithm resulted in 2 partitions. This method is advantageous for situations where a bipartite structure is suspected or explicitly defined.

6.4 Comparison of Results

- **Louvain Method:** Identified 7 communities.
- **Label Propagation Algorithm:** Identified 5 communities.
- **Kernighan-Lin Algorithm:** Identified 2 partitions.

The varying results from these algorithms provide distinct perspectives on the community structure within our network. The Louvain Method and Label Propagation Algorithm offer detailed insights into the modular divisions, whereas the Kernighan-Lin Algorithm emphasizes bipartite partitioning.

These findings are crucial for understanding the organizational structure and interconnections within the network, guiding further analyses and interpretations of community dynamics and relationships.

7 Threats to Validity

1. **Dataset Bias**

   Threat: Dataset may not represent all international matches or excludes certain types (e.g., friendlies).

   Mitigation: Rigorous criteria excluded non-competitive matches. Sensitivity analyses varied inclusion criteria.

2. **Exclusion Impact**

   Threat: Excluding draws and early matches limits insights into team performance over time.

   Mitigation: Sensitivity analyses considered including these matches for balanced interpretation.

3. **Data Quality**

   Threat: Variations in data quality across historical periods may introduce inaccuracies.

   Mitigation: Rigorous validation and cleaning procedures ensured data integrity.

4. **Algorithm Sensitivity**

   Threat: Choice of network algorithms may influence results.

   Mitigation: Multiple algorithms used for validation across different analytical approaches.

5. **Temporal Context**

   Threat: Changes in football regulations and geopolitical events may bias conclusions.

   Mitigation: Results contextualized within historical and regional dynamics for comprehensive analysis.

By addressing these validity threats through rigorous methodologies, comprehensive data preprocessing, and sensitivity analyses, this research aims to provide reliable insights into team interactions, competitive landscapes, and the evolution of international football dynamics. These mitigation strategies support the credibility and applicability of the study's findings, contributing valuable insights to sports analytics and management practices.

8 Conclusion and Future Work

In this study, we have undertaken a comprehensive exploration of international football through sophisticated network analysis techniques spanning over a century of competitive matches. Our analysis has uncovered significant insights into team performances, interactions, and the structural dynamics of global football networks.

Key findings highlight the enduring dominance of iconic teams like Brazil in South America and Germany in Europe, reflecting their historical successes and influence on international football. The network analysis revealed a highly interconnected structure characterized by substantial clustering and central nodes, underscoring the cohesive nature of global football interactions.

Centrality measures such as Degree, Betweenness, Closeness, Eigenvector, and PageRank provided deeper insights into pivotal countries and their roles in shaping global football dynamics. Continent-wise subgraph analyses illuminated regional football dynamics, showcasing historical rivalries and competitive landscapes across different continents.

Our study addresses various validity concerns through rigorous methodologies, ensuring the robustness and reliability of our findings despite challenges such as dataset biases and temporal variations. By integrating advanced visualization techniques and interdisciplinary perspectives, we offer valuable implications for sports analytics, management strategies, and fan engagement in international football.

Looking forward, future research could expand on temporal analyses to track the evolving dynamics of football across distinct historical periods. Predictive modeling using network characteristics may enhance our ability to forecast international match outcomes, while qualitative insights into player dynamics and coaching strategies could provide a holistic understanding of football competitiveness.

As football continues to evolve and transcend geographical boundaries, our study underscores the evolving roles of national teams and their community structures in shaping the global football landscape. By embracing complexity and leveraging cutting-edge methodologies, we aim to contribute to a deeper understanding of international football dynamics and its implications for sports enthusiasts and stakeholders worldwide.

9 Acknowledgements

We want to extend heartfelt thanks to Prof. Massimo Franceschetti for his consistent guidance and support throughout this project. His expertise in network analysis and thoughtful feedback played a crucial role in shaping our methodologies and understanding the findings of the international football match network. Prof. Franceschetti's mentorship not only helped us grasp complex concepts but also gave us confidence in tackling various challenges encountered during the research.

We are also grateful to TAs Sandeep Chintada and Onur Tepecelik for their dedicated assistance and valuable contributions. Their knowledge in data analysis and network visualization was instrumental in refining our methods and improving the clarity of the project. Their constructive feedback and availability during office hours were essential in overcoming obstacles and ensuring the project's thoroughness and quality.

Additionally, we referred to some GitHub repositories[8] [1] [2] and YouTube resources[7] [4], which have been properly cited in the reference section of our paper.

10 Tools Used

- **Pandas & Jupyter Notebook:** Used for data preprocessing and initial data exploration.
- **Gephi:** Employed for network representation, visualization, and advanced network analysis.

- **NetworkX:** Python library utilized for calculating various graph measures and network analysis.
- **GitHub Repository:** The project repository can be found at https://github.com/atharv2802/football_network_analysis_ece227.

References

[1] Peter Aldhous. Introduction to gephi for network visualization. https://paldhous.github.io/NICAR/2016/gephi.html, 2016.

[2] Babetas. Football network analysis.

[3] Filipe Manuel Clemente, Fernando Manuel Lourenço Martins, and Rui Sousa Mendes. Social network analysis applied to team sports analysis. 2015.

[4] Jennifer Golbeck. Gephi tutorials. https://www.youtube.com/playlist?list=PLk_jmmkw5S2BqnYBqF2VNPcszY93-ze49, 2024.

[5] Graham. Visualising a football match as a network graph using gephi, 2014. Accessed: 2024-06-13.

[6] Pavlina Kröckel, Alexander Piazza, and Kathrin Neuhofer. Dynamic network analysis of the euro2016 final: Preliminary results. In *2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW)*, pages 114–119, 2017.

[7] Natarajan Meghanathan. Gephi network analysis of soccer football network. https://www.youtube.com/watch?v=oZY5Xfp1yns.

[8] Mooseburger1. NFL Social Network Analysis. https://github.com/Mooseburger1/NFL-Social-Network-Analysis, 2017.

[9] Raffaele Trequattrini, Rosa Lombardi, and Mirella Battista. Network analysis and football team performance: a first application. *Team Performance Management*, 21, 03 2015.