UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Can effective cooperation between teammates enhance overall team performance?

Gender-Based Analysis of Football Passing Networks

Alba Cristina Martín Gonzálvez

Mestrado em Matemática Aplicada à Economia e Gestão

Trabalho de Projeto orientado por:

Professor Miguel Fragoso Constantino
Professora Marília Cristina de Sousa Antunes

2024

To myself, for my perseverance and discipline to end this project.

Resumo[1]

A principal motivação para este estudo foi perceber o que as redes de passes geradas num jogo de futebol nos dizem sobre as estratégias que as equipas seguem, os jogadores mais influentes e outras dinâmicas que podem ocorrer e que não podem ser observadas com outros tipos de análise. Além disso, é necessário perceber se as diferenças nas dinâmicas de jogo e nas estruturas de rede entre equipas masculinas e femininas podem explicar por que razão, apesar de o futebol feminino estar a ganhar popularidade, o futebol masculino continua a ser considerado mais emocionante e tem mais cobertura mediática e investimento financeiro.

Foram utilizados três métodos quantitativos diferentes para a análise: ANOVA unidirecional, modelos lineares mistos generalizados (GLMM) e métodos de agrupamento utilizando a medida de distância de Mahalanobis em vez da distância euclidiana. O estudo foi realizado considerando as duas últimas edições do Campeonato do Mundo de Futebol da FIFA para ambos os géneros, ou seja, as edições de 2018 e 2022 para as equipas masculinas e as edições de 2019 e 2023 para as equipas femininas. Estas análises permitiram-nos aprofundar a forma como as redes de passes podem influenciar o desempenho global das equipas ao longo das diferentes fases da competição, bem como observar as diferenças entre homens e mulheres.

Mais especificamente, o principal objetivo desta análise foi investigar se existe uma relação entre as métricas da rede de passes e o desempenho das equipas em competições internacionais, ou seja, se as equipas que chegam às fases finais do torneio apresentam padrões específicos em termos de número de passes, ligações entre jogadores e coesão em campo. Além disso, procurámos identificar as principais diferenças entre as equipas masculinas e femininas, tanto em termos de métricas de rede como de dinâmica de jogo, para ver como estes factores podem influenciar o sucesso competitivo de uma equipa.

As métricas utilizadas para os diferentes modelos estatísticos foram cinco: 1) número total de ligações ou conexões entre jogadores (ou seja, o número de passes que foram feitos durante o jogo); 2) diâmetro da rede, que mede a distância máxima entre dois jogadores na rede de passes, reflectindo o quão dispersos ou compactos os jogadores estão no campo; 3) densidade da rede, que mede quantas conexões possíveis entre jogadores foram realmente usadas, fornecendo informações sobre a coesão da equipa; 4) assortatividade, uma métrica que indica se os jogadores mais centrais na rede tendem a ligar-se a outros jogadores igualmente centrais; e 5) coeficiente de agrupamento geral, que mede o grau em que os jogadores tendem a formar pequenos grupos de passes, o que é indicativo de padrões de jogo mais fechados que não envolvem toda a equipa.

Adicionalmente, o desempenho da equipa foi medido em termos da fase máxima alcançada no torneio. Foram estabelecidas seis fases para análise: 1) Fase de grupos, 2) oitavos de final, 3) quartos de final, 4) meias-finais, 5) final para o terceiro lugar e 6) final. A ideia era determinar se as métricas de rede são ou não indicadores significativos do sucesso das equipas, ou seja, se ajudam a observar padrões claros que distinguem as equipas que chegam às fases mais avançadas do torneio.

Metodologia

O estudo abrangeu dados das edições mais recentes do Campeonato do Mundo de Futebol da FIFA, tanto para homens (2018 e 2022) como para mulheres (2019 e 2023). Foram analisados todos os jogos desde a fase de grupos até à final, tendo sido construídas redes de passes a partir dos dados de cada jogo. Antes de passar à análise dos modelos estatísticos, foi efectuada uma análise mais descritiva e comparativa, começando por estudar as métricas globais de cada equipa e, em seguida, analisando as estratégias de jogo das últimas finais através da visualização das redes de passes sobrepostas no campo com a posição média do onze inicial.

As redes foram depois comparadas com os modelos acima referidos. Em particular, foi utilizada uma ANOVA unidirecional para determinar se existiam diferenças significativas nos valores médios das métricas de rede entre as equipas que passaram às diferentes fases do torneio. Posteriormente, foram

[1] The summary was first written in Spanish and translated into Portugues with Deepl (https://www.deepl.com/pt-PT/translator)

aplicados modelos lineares generalizados mistos (GLMM) para controlar possíveis efeitos aleatórios e para avaliar se as métricas de rede continuavam a ser significativas após a contabilização de outros factores. Além disso, foram utilizados métodos de agrupamento para identificar padrões semelhantes entre as equipas que tiveram um desempenho semelhante no torneio, ou seja, quer tenham ou não passado da fase de grupos ou de fases mais avançadas.

Resultados

Uma das conclusões mais importantes deste estudo foi que as redes de passes nos campeonatos femininos mostraram uma relação clara entre o número total de ligações e a densidade da rede com o desempenho da equipa. As equipas que conseguiram passar a fase de grupos tendiam a ter redes mais densas e um maior número de ligações entre os jogadores. Isto sugere que a coesão da equipa e a capacidade de manter um elevado nível de interação entre os jogadores é crucial para o sucesso no futebol feminino.

Em contrapartida, nos campeonatos masculinos, não foi encontrada uma relação tão clara entre as métricas da rede e o desempenho da equipa. As equipas que passaram às fases finais do torneio não tinham necessariamente redes mais densas ou um maior número de ligações. Este facto levanta questões interessantes sobre as diferenças na dinâmica do jogo entre homens e mulheres. No futebol masculino, o sucesso pode depender mais da capacidade individual dos jogadores e da sua capacidade de fazer jogadas individuais, enquanto no futebol feminino o trabalho de equipa e a coesão parecem ser factores determinantes mais importantes.

Implicações do estudo

Este estudo oferece uma perspetiva interessante sobre a forma como os factores psicológicos e culturais de género podem influenciar a dinâmica e o desempenho das equipas no desporto. Em particular, sugere que as estratégias baseadas no trabalho de equipa e na coesão podem ser mais importantes no futebol feminino, enquanto os homens tendem a ser mais individualistas e competitivos.

Por outro lado, coloca-se a questão de saber se estas diferenças psicológicas e culturais explicam realmente as diferenças de rendimento e de perceção entre o futebol masculino e o feminino. Argumenta-se frequentemente que factores como a intensidade física e as diferenças tácticas também desempenham um papel, tornando o jogo mais rico e, por conseguinte, mais interessante e lucrativo. No entanto, um estudo mostrou que, sem informação sobre o género, a perceção da qualidade não varia entre os dois. Este facto corrobora a discussão deste estudo: as percepções são influenciadas pelos estereótipos de género. Por conseguinte, é importante desfazer estes estereótipos, uma vez que as diferenças mais significativas resultam de factores sociais e não tanto de factores biológicos.

Limitações e investigação futura

Embora este estudo esclareça as diferenças entre as redes de passes no futebol masculino e feminino, é importante referir algumas limitações. Em primeiro lugar, os dados utilizados para a análise estão incompletos. Não existem dados sobre os jogos dos quartos de final do Campeonato do Mundo Feminino de 2019, o que pode causar algum enviesamento e fazer com que a comparação não seja totalmente exacta. Além disso, nesse torneio havia menos equipas, 24, e só em 2023 é que se decidiu igualar o futebol feminino e masculino, com a participação de 32 equipas.

Por outro lado, para obter resultados mais fiáveis, seria preferível utilizar dados de ligas regulares ou ter dados históricos de mais Campeonatos do Mundo, uma vez que a evolução das equipas pode ser melhor analisada e podem ser obtidas métricas mais consistentes.

Em conclusão, este estudo levanta novas questões para investigação futura, a fim de fornecer provas científicas mais sólidas sobre a razão pela qual o futebol masculino gera mais interesse.

Palavras chave*:* futebol, análise de redes, One-Way ANOVA, GLMM, distância de Mahalanobis.

Abstract

In the current study, one-way ANOVA, generalized linear mixed models and clustering methods were applied to examine the passing networks for the latest four men's and women's FIFA World Cup championships. The aim of the study was to analyze and compare the overall network metrics across different stages of a competition and between male and female teams. Namely, we seek to find a) whether teams that reach the highest stages in competition present high values of total number of links and density and low values in clustering coefficient and diameter; b) the main differences (if any) between male and female games. The variables considered in this study are of two types: graph performance variables, which are the metrics derived from the passing network namely 1) total number of links, 2) diameter, 3) density, 4) assortativity, and 5) global clustering coefficient. Plus, the overall team performance variables understood as the maximum stage that each team reached in each of the tournaments considered. The variables of maximum stage are: 1) Group Stage, 2) Round of 16, 3) Quarterfinals, 4) Semi-finals, 5) 3rd Place Final, and 6) Final. While results from men's Worl Cups showed no evidence of significant differences between the maximum stage reached and the network metrics, results for women's Worl Cup tournaments revealed more interesting findings. Statistically significant differences were found between two network metrics - total number of links and density - and the maximum stage reached. Yet, there is still insufficient evidence to explains why men's soccer is perceived as more thrilling than women's.

Keywords: football, network analysis, One-Way ANOVA, GLMM, Mahalanobis distance.

Index

List of Acronyms and Abbreviations VIII
List of Tables IX
List of Graphs and Figures X
1. Introduction 1
1.1 Related works 1
1.2 Motivation 2
1.3 Structure 3
2. Methods 4
2.1 Sample 4
2.2 Data Collection and Formatting 4
2.3 Network Analysis 6
2.3.1 Graph and Network Theory 6
2.3.2 Metrics 7
2.3.3 Visualizations 13
2.4 Variables and Objective of this Study 15
2.5 Statistical Procedures 15
2.5.1 One-way ANOVA 16
2.5.2 Generalized Linear Mixed Model for a Binary response variable 17
2.5.3 Clustering by teams – Hierarchical clustering using the Mahalanobis distance 18
3 Results 21
3.1 Passing Network Analysis 24
3.2 Statistical Procedure Results 28
4. Discussion 35
5. Limitations 37
6. Conclusions 38
Bibliography 39
Annexes 43

List of Acronyms and Abbreviations

AIC - Akaike Information Criterion

ANN - Artificial Neural Network

ANOVA - Analysis of Variance

ASA - American Statistical Association

BIC - Bayesian Information Criterion

BLR - Binomial Logistic Regression

CMPN - Complex Multiplex Passing Network

FIFA - Fédération Internationale de Football Association

GLMM - Generalized Linear Mixed Model

GSPN - Goals Passing Network

HSD - Honestly Significant Difference

KNN - K - nearest Neighbors Algorithm

MWC - Men's World Cup

NB - Naive Bayes

RF - Random Forest

UEFA - Union of European Football Associations

WC - World Cup

WWC - Women's World Cup

List of Tables

Table 1 - Sample dataset variables with the corresponding model term 18

Table 2 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2018 (Male) 21

Table 3 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2022 (Male) 22

Table 4 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2019 (Fem) 23

Table 5 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2023 (Fem) 23

Table 6 - Argentina Microscale Metrics - WC 2022 Final 25

Table 7 - France Microscale Metrics - WC 2022 Final 26

Table 8 - Spain Women's Microscale Metrics - WC 2023 Final 27

Table 9 - England Women's Microscale Metrics - WC 2023 Final 28

Table 10 - ANOVA Table for Links - maxPhase WWC 2019 29

Table 11 - TukeyHSD test on Links – maxPhase WWC 2019 29

Table 12 - ANOVA Table for Density - maxPhase WWC 2019 & 2023 29

Table 13 - TukeyHSD test on Density – maxPhase WWC 2019 & 2023 30

Table 14 - Model fit measures and random effect variance MWC 2018 & 2022 31

Table 15 - Fixed effect predictors for model one (m1) - MWC 2018 & 2022 31

Table 16 - Fixed effect predictors for model four (m4) - MWC 2018 & 2022 31

Table 17 - Model fit measures and random effect variance WWC 2019 & 2023 31

Table 18 - Fixed effect predictors for model one (m1) - WWC 2019 & 2023 32

Table 19 - Fixed effect predictors for model four (m4) - WWC 2019 & 2023 32

Table 20 - Correlation matrix for MWC 2018 32

Table 21 - Correlation matrix for MWC 2022 32

Table 22 - Correlation matrix for WWC 2019 32

Table 23 - Correlation matrix for WWC 2023 32

Table 24 - Hierarchical clustering results MWC 2018 33

Table 25 - Hierarchical clustering results MWC 2022 33

Table 26 - Hierarchical clustering results WWC 2019 34

Table 27 – Hierarchical clustering WWC 2023 34

List of Graphs and Figures

Figure 1 – Example of the competitions.json file 5
Figure 2 – Example of matches. 5
Figure 3 - Example of events. 6
Figure 4 - Example of an undirected graph 6
Figure 5 - Example of a directed graph 6
Figure 6 - Adjacency matrix 6
Figure 7 - Graphic representation of the PageRank metric 10
Figure 8 - Graphic representation of the betweenness centrality 11
Figure 9 - Example of average position visualization 13
Figure 10 – Example of closeness between players visualization 14
Figure 11 - Example of offensive formation visualization 14
Figure 12 - Example of a dendrogram with the optimal number of 2 clusters 19
Figure 13 - Argentina Passing Network - WC 2022 Final 24
Figure 14 - France Passing Network - WC 2022 Final 26
Figure 15 - Spain Women's Passing Network - WC 2023 Final 27
Figure 16 - England Women's Passing Network - WC 2023 Final 28

1. Introduction

Football is not only the most popular sport in the world, with roughly 4 billion fans globally (Contributor, 2024), but also one of the most profitable. Many of the main soccer leagues are within the top 10 richest sport leagues in the world (Thakur, 2024). Furthermore, football's unpredictable outcomes and constant multidirectional passes add complexity to its tactical analysis, what makes football's success to rely on achieving the ideal team balance rather than just individual skills (Leela et al., 2024). This, combined with its social and financial relevance, may be pushing the arising trend of applying statistical analysis to the sport. The following lines will present a concise introduction of the origin and evolution of this trend.

When statistical analysis was not a widely used tool in the field of sports, decisions in team sports were often subjective, relying on the intuition and personal experience of decision-makers. It was not until the early 90s, that the American Statistical Association (ASA) started a section dedicated to sports statistics. However, decisions driven by data captured the media attention over a decade later, mainly because of the book, and the subsequent adapted film Moneyball, with Brad Pitt as starring character. The film is about the Oakland Athletics General Manager Billy Beane, who applied quantitative methods to build a winning team with players underrated by the market. (Ievoli et al., 2021).

The use of network analysis in statistics has gained popularity during the past decades, to a certain extent thanks to the publication of the book "*Who Shall Survive*" by the Romanian American psychiatrist, psychosociologist, and educator Jacob Levy Moreno (1889-1974) in 1934 and with the growth of the field of network science. Its application extends on multiple disciplines, including sociology, biology, computer science, health care (Sabot et al., 2017), political science (Ward et al., 2011), management and leadership (Kacanski & Lusher, 2017), and economics, just to name a few (Easley & Kleinberg, 2010). In football, network analyses provide a novel perspective on player behavior, identifying key players and measuring passing success. High connectivity metrics derived from the network graphs emphasize the importance of understanding player relationships and positions and identifying areas for improvement like cooperation, balance, and synergies within the team. In line with this idea, Beheshtian-Ardakani et al. (2023) pointed that "sports scientists use passing networks to (I) characterize a player's role in a team, (ii) find players with similar passing skills, and (iii) determine the importance of players". Recently, player and ball spatial data has been processed and saved for more dynamic visualizations and detailed analysis of team cohesion and tactics. Coaches can leverage this information to design training tasks and enhance decision-making. (Caicedo-Parada et al., 2020).

1.1 Related works

According to the literature, connectivity and centralization metrics derived from a soccer passing network can give a hint on how dependent the team is on few players: (I) a network with the highest density values (all players very well connected) leads to the best performances, and (ii) networks with high centralized tendencies (only few key players taking a role in the game) are associated with poor performances (Blundy & Harrison, 2006). To the same conclusion arrived Grund (2012) in his study over the English Premier League that studied how the team performance (goals scored) was associated with passing network metrics like density or centralization.

Other than the connectivity or distribution of the team, network analysis can be used for other purposes. There exists the possibility of examining the goals scoring passing networks (GSPN) in conjunction with contextual variables such as match result (winning, drawing, losing), and compare the GSPN during

group stages and knock out stages. The goal scoring pass used to build the network commenced when the goal-scoring team gained possession and was completed when the ball crossed the goal line. Mclean et al. (2018) examined this relationship and found that variations on the passing network, that is, team tactics, depend on the score of the game, but not on the stage of the tournament reached by the team. They also demonstrated how degree centralization (DC) metrics can be a method to describe important entities, not only key players in attack (Clemente, Martins, Wong, et al., 2015), but also prominent pitch zones when it comes to goal scoring.

Contrary to what previously was done by the literature, single-layer passing network analysis, Beheshtian-Ardakani et al. (2023) proposed a novel model called Complex Multiplex Passing Network (CMPN) model, in which each layer represents a specific type of pass between players. This distinction allows a detailed and deeper analysis at different scales. Given that passes are not aggregated in one layer, the information related to the distinct type of passes is not lost. The model proposed effectively predicts the outcomes of attacking plays in soccer with over 90% accuracy and its overall effectiveness is about 70%, which means that 70 out of 100 times, the model correctly identify and classify the attacking plays.

Similarly, a different perspective was given by Leela et al. (2024), emphasizing the importance of using three topological scales when analyzing soccer passing networks: microscale, mesoscale and macroscale. The microscale focuses on individual player roles, while the mesoscale examines patterns within motifs (small groups of players within the network) and their impact on team success. Finally, the macroscale studies the whole team network, revealing team performance and playing style. This distinction between scales allows a deeper understanding of player significance, network patterns, and team execution, likewise the one provided by the Multiplex Passing Network model mentioned above.

Machine learning models have been also examined under statistical network analysis. Concretely, Ievoli et al. in 2021 run four different techniques (Binomial Logistic Regression (BLR), Naive Bayes (NB), Artificial Neural Network (ANN), and Random Forest (RF)) to predict the probability of winning games by incorporating passing network indicators as explanatory variables. According to their results, the binomial logistic regression model performs the best and diameter and betweenness centralization variables are related to offensive actions and team performance.

The studies mentioned in the above paragraphs are only few examples of the vast possibilities in terms of methods, models, metrics and insights that could be derived from soccer data and concretely passing network analysis. The present study has followed the approach of Mclean et al. (2018)'with some variations, aiming to analyze and compare the overall network metrics across different stages of a competition and between male and female teams. Namely, we seek to find:

a) Whether teams that reach the highest stages in competition present high values of total number of links and density and low values in clustering coefficient and diameter.

b) The main differences (if any) between male and female games.

1.2 Motivation

As part of a summer course done a few years ago, I could be researcher for three weeks. The topic for the investigation was, indeed, the statistical analysis of a passing network: being coach for three weeks. The description that introduced the study mentioned how the complexity of a sport played by teams allows the utilization of networks to identify patterns. I felt curious, even though soccer is not my preferred sport. Thanks to this work I discovered how powerful and useful the networks are. It was a short experience, but it was enough to know that I wanted to continue the research I worked on during those three weeks.

When you deep dive into the football world, it is striking – or at least, from my standpoint as a woman – that there were very few studies on women's football. This, coupled with the ongoing debate about why female soccer generates less economic revenue has prompted me to focus my investigation on the main differences between genders: Is it a matter of real differences that is supported by data or is it a result of historical neglect of women's soccer? Do women, indeed, perform worse than men?

These questions, along with the curiosity on network analysis, made me want to embark the journey: *In search of the potential differences*.

1.3 Structure

The remainder of this report is structured as follows: Section 2 explains the methods used in the study, including the network analysis and the statistical procedures. Section 3 presents the results of applying these methods. Section 4 details a discussion highlighting the key insights obtained from the results. Section 5 presents the limitations of the study. Section 6 summarizes the main conclusions. Section 7 and 8 are the bibliography and the annexes respectively.

2. Methods

2.1 Sample

Four hundred eighty-eight matches were analyzed in this study. All the matches of all national teams that participated in the tournament were analyzed. One hundred and twenty-eight official matches from each competition:

- FIFA World Cup 2018
- Women's World Cup 2019
- FIFA World Cup 2022
- Women's World Cup 2023

For the network graph visualizations, redundant information was suppressed and only graphs from the final stages of the tournament were generated. That is, graphs related to the matches on Quarterfinals, Semi-finals, 3rd Place final and Final. However, for the metrics and for the statistical models all the phases were considered, that is, from the Group Stage to the Final Stage. Thus, a total of 488 adjacency matrices were generated on the basis of teammate interactions, but only 56 were converted into network graphs[2].

There was a total of 131,396 passes for both men's competitions and a total of 107,231 passes for the women's tournaments. However, passes categorized as *incomplete*, *out*, *unknown* or *pass offside* were removed from the dataset used to obtain the network graphs, the corresponding metrics and to carry out the statistical models. Thus, the final number of passes analyzed for men's competitions was 106,513 passes and 79,708 for women's.

2.2 Data Collection and Formatting

The data set used in this analysis is composed by all the spatiotemporal events occurred during the matches taking place in the competitions mentioned in the previous section. Accessibility to this data has been possible thanks to StatsBomb[3], which offers a public repository with information related to different football competitions and leagues all around the world.

The data is presented in JSON files with the following structure:

- *Competitions.json:* it saves information related to each competition with its corresponding season year.

[2] Dataset from Women's World Cup 2019 is smaller due to the smaller number of participant teams – 24 instead of 32 - and missing data from quarter-finals in StatsBombs source system. Hence, in this case only 104 matches were analyzed.
[3] Data can be accessed in this url: https://github.com/statsbomb/open-data

```
[ {
  "competition_id" : 9,
  "season_id" : 27,
  "country_name" : "Germany",
  "competition_name" : "1. Bundesliga",
  "competition_gender" : "male",
  "competition_youth" : false,
  "competition_international" : false,
  "season_name" : "2015/2016",
  "match_updated" : "2023-08-17T23:51:11.837478",
  "match_updated_360" : null,
  "match_available_360" : null,
  "match_available" : "2023-08-17T23:51:11.837478"
}, {
```


Figure 1 – Example of the competitions.json file

- *Matches:* this folder saves all the matches related to each competition and season. It is organized in subfolders which name is related to the *competition_id* of the competition they belong to. These subfolders contain the *.json* files named by the *season_id* of the season they belong to within that competition.


```
[ {
  "match_id" : 22949,
  "match_date" : "2019-06-12",
  "kick_off" : "18:00:00.000",
  "competition" : {
    "competition_id" : 72,
    "country_name" : "International",
    "competition_name" : "Women's World Cup"
  },
  "season" : {
    "season_id" : 30,
    "season_name" : "2019"
  },
  "home_team" : {
    "home_team_id" : 857,
    "home_team_name" : "Germany Women's",
    "home_team_gender" : "female",
    "home_team_group" : "B",
    "country" : {
      "id" : 85,
      "name" : "Germany"
    },
    "managers" : [ {
      "id" : 53,
      "name" : "Martina Voss-Tecklenburg",
      "nickname" : null,
      "dob" : "1967-12-22",
      "country" : {
        "id" : 85,
        "name" : "Germany"
      }
    } ]
  },
```


Figure 2 – Example of matches.

- *Events:* it saves all the events occurring during a match (passes, shots, fouls, goals, etc.). It contains the *.json* files and are named by the *match_id* they are related to. In the figure below it is shown an example of the structure of the event we used in our study, the passes.

```
}, {
  "id" : "c37b95fc-44a3-47b8-83cb-1d1389efb66c",
  "index" : 14,
  "period" : 1,
  "timestamp" : "00:00:05.459",
  "minute" : 0,
  "second" : 5,
  "type" : {
    "id" : 43,
    "name" : "Carry"
  },
  "possession" : 2,
  "possession_team" : {
    "id" : 776,
    "name" : "Denmark"
  },
  "play_pattern" : {
    "id" : 9,
    "name" : "From Kick Off"
  },
  "team" : {
    "id" : 776,
    "name" : "Denmark"
  },
  "player" : {
    "id" : 16190,
    "name" : "Rasmus Nissen Kristensen"
  },
  "position" : {
    "id" : 7,
    "name" : "Right Wing Back"
  },
  "location" : [ 86.5, 74.9 ],
  "duration" : 0.04,
  "related_events" : [ "7e3a43d6-07e6-4e98-b93e-c2073bb88431", "f53c3831-2601-4ccd-b0f5-5503e165abcd" ],
  "carry" : {
    "end_location" : [ 86.5, 74.9 ]
  }
} {
```


Figure 3 - Example of events.

For a detailed information of what else it is contained in the data available, in the Statsbomb github repository[4] there is a folder called *doc* in which there exists PDFs files explaining how data was tracked and saved.

2.3 Network Analysis

2.3.1 Graph and Network Theory

Graphs and Networks

In graph theory, a graph is defined as group of nodes and edges noted as G = (V, E), being V the set of nodes (also known as vertices), represented as an ordered set $V = \{v_1, v_2, \ldots, v_N\}$ and E the group of edges. Each edge $e \in E$ is represented as $e = (v_i, v_j)$ being v_i and v_j the connected nodes by the edge e.

There exist two types of graphs: undirected and directed shown in Figure 4 and Figure 5 respectively. When the graph is directed there is a distinction between the relationship between (v_i, v_j) and (v_j, v_i). This is because the edges could have opposite directions and even not exists one of the directions. However, for undirected graphs, there is not differentiation between (v_i, v_j) and (v_j, v_i) considering that there is no direction.

Likewise, nodes can have degrees. In the case of an undirected graph, the degree is the number of edges that affects the node, that is to say, the number of nodes connected to it. For directed graphs, the degree is the sum of the number of directed edges going to that particular node (inbound degree) in addition to the number of edges going out from the node (outbound degree).

[4] https://github.com/statsbomb/open-data/tree/master/doc

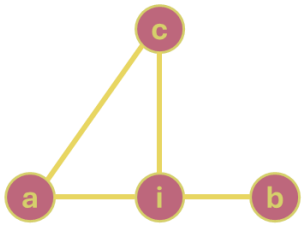

Figure 4 - Example of an undirected graph

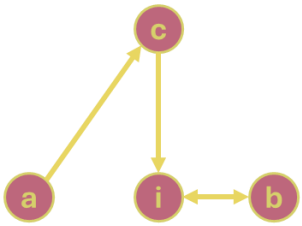

Figure 5 - Example of a directed graph

Edges can be weighted or not. The meaning of this weight depends on each context. In some cases, it could be the distance between the connected nodes, i.e. in a transportation network, kilometers between cities or the strength of the link, i.e. in a social network, frequency of communication or shared resources. When the edges of a graph are weighted, it is called a network.

Applying the theory to football's reality, it will be used graphs with weighted edges to represent the passing network. The nodes will represent the players and the edges the passes executed by them. Therefore, the edge (j_1, j_2) represents the pass between two players of the same team and depending on the nature of the edge (being directed or not) it will inform us whether it goes from player 1 to player 2 or the other way around, or just that the link between them exists. In this study, the networks are undirected graphs, and hence it will just inform about the existence of the pass. It has been decided to work with undirected graphs for two main reasons: 1) simplicity either in the visualizations and calculations and, 2) given the objective to focus on the global properties and interactions within the network, the specific details of which player passes to whom are less significant. Instead, the emphasis is on understanding the overall structure and distribution of passes. The weight of the edge will be the number of passes between two players, in the figure it will be seen thicker edges for those with more passes. Additionally, the size of the node is also weighted in terms of its degree. The more the player interacts with other players, the bigger the size of the node.

Please note that in this report, since we are considering undirected networks, the notation (*i*, *j*) to represent an edge between nodes *i* and *j* is equivalent to (*j*, *i*)

Adjacency Matrix

A graph can be represented matrix-wise with the Adjacency Matrix. In it are depicted the connections between a node (player) and an adjacency node (teammate). Given a graph G = (V, E), the adjacency matrix A will be a N x N matrix, being N the total number of nodes and each entry A_{ij} is an edge formed by (v_i, v_j). Consequently, $A_{ij} = 1$ when there exists a link between v_i and v_j and $A_{ij} = 0$ in the opposite case. Figure 6 shows an example: $A_{1,4} = 0$ because there are no links between node 1 and 4 and $A_{1,3} = 1$ because there exists a link between node 1 and 3.

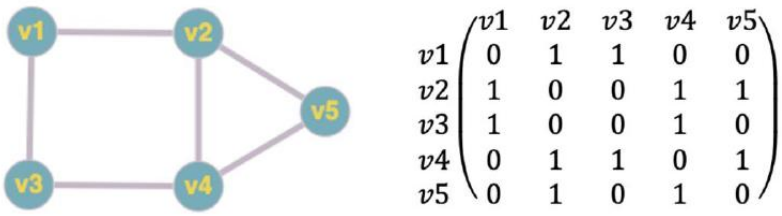

Figure 6 - Adjacency matrix

When the edges are weighted, each value of A_{ij} will be the specific weight associated with $e = (v_i, v_j)$. Similarly, when there is no link, the value is 0. Nevertheless, when weights are considered, the matrix is no longer an adjacency matrix since this one only contains 1s and 0s.

Knowing the adjacency matrix is crucial to calculate the eigenvectors of a graph an important metric to understand the importance of each node (player). In this study the matrices per se were not calculated due to the fact of using a function from python. However, it is important to understand the Math behind.

2.3.2 Metrics

Grouping the metrics into macro and microscale metrics, following Leela et al. (2024) distinction, allows a better understanding of where each metric puts the focus on and what are they relevant for. In this case, it was not selected any metric within the mesoscale level because subgraphs and motifs are not in the scope of this study. Likewise, the selection of the seven metrics to be described in the next lines is based on the popularity among the scientific community.

Macroscale metrics

The exhibition and playing style of soccer teams can be uncovered using these five key network measurements: total number of links, density, diameter, global clustering coefficient and assortativity

- **Total Links** refers to the absolute value of total passes conducted between teammates during a match. When the value is high, it means there was a high cooperation between players. However, since it is an absolute measure, the information it can provide is limited. Nevertheless, it can be useful to obtain a good idea which of the 32 teams of the competition are above and below the average Total Links index as an indicator of which teams cooperate more. Additionally, usually, this metric tends to be correlated with successful interaction that may result in long ball possessions, good performance and strong collective organization against the opponent team (Clemente et al., 2015).

$$\boldsymbol{L} = \frac{\mathbf{1}}{\mathbf{2}} \sum_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{N}} \sum_{\boldsymbol{i}\neq\boldsymbol{j}}^{\boldsymbol{N}} (\boldsymbol{i},\boldsymbol{j}) \tag{2.1}$$

With:

L being the network total number of links

$\sum_{\boldsymbol{i}\neq\boldsymbol{j}}^{\boldsymbol{N}} (\boldsymbol{i},\boldsymbol{j})$ is the total number of passes from player *i* to all its other teammates.

- **Density** is the relative value that measures the overall connection between players of the same team. It is defined as the ratio of the total links that present L to the maximum possible number of links**.** Due to the fact that the density metric is a ratio, values vary between 0 and 1. When the value is closer to 1 means the nodes are highly connected to each other. In an undirected graph it is calculated as follows:

$$\triangle = \frac{\boldsymbol{L}}{\frac{\boldsymbol{N}(\boldsymbol{N}-\mathbf{1})}{\mathbf{2}}} \tag{2.2}$$

Being:

Δ is the density of the graph

L the number of edges within the graph

N the number of nodes within the graph

N(N-1)/2 the maximum possible number of links in an undirected graph.

- **Diameter** represents the extension of a graph and the *geodesic distance* between the two most distant nodes. In graph theory, the *geodesic distance* is the length of the shortest path, in terms of number of edges (or steps) between two connected nodes.

$$\boldsymbol{d(i,j) = min\left\{\sum_{e \in P} w(e) \mid P\right\}} \tag{2.3}$$

Where:

P is the path from *i* to *j*

e represents an edge in the path *P*

w(e) represents the weight of the edge *e*.

Additionally, two nodes are *connected* if a sequence of nodes exists, and their links (*walks)* are adjacent. A *path* occurs when a walk is only formed by distinct nodes and lines. The diameter has a minimum of 1 when all nodes are directly connected with each other, and a maximum of n-1 (because a node is not able to pass to itself). Low values of a diameter reflect that each player could be connected to any other player within a low number of passes meaning the network is well connected and has an efficient ball distribution. Conversely, big diameters mean there are isolated players, and the team is dependent on a few key players. It is expressed as follows:

$$\boldsymbol{D = max_{i,j \in V}\{d(i,j)\}} \tag{2.4}$$

Where:

D is the diameter of the graph

V is the complete set of all the nodes

*d(i,j)*is the geodesic distance between node *i* and *j*.

- **Clustering Coefficient** metric was introduced by Watts and Strogatz in 1998. It measures how close a node and its neighbors in the graph are to become a complete subgraph, more precisely, a triangle (a subgraph of three nodes where each node is connected to every other node). With the Clustering Coefficient it can be found a graph that is a small-world network. That is, a network with a small average distance between its nodes but relatively large number of subgraphs. Thereby, the local clustering coefficient measures the degree of interconnectivity in the neighborhood of a node. High values mean that the node and its neighbors are close to become a complete subgraph. The formula for weighted graphs, is defined as the geometric average of the subgraph edge weights (Clustering — NetworkX 3.3 Documentation, s. f.):

$$\boldsymbol{C_v = \frac{1}{k_v(k_v - 1)} \sum_{vw} (\widehat{w}_{uv}\widehat{w}_{uw}\widehat{w}_{vw})^{\frac{1}{3}}} \tag{2.5}$$

Where:

C_v is the local clustering coefficient from node *v*

k_v is the number of neighbors of node *v* (its degree)

$k_v(k_v - 1)$ the maximum possible number of triangles that can be formed.

$\widehat{w}_{uv}$ is the edge weight normalized by the maximum weight in the network: $\widehat{w}_{uv} = w_{uv} / \max(w)$

We decided to use the same variant version of the global clustering coefficient implemented by Clemente et al. (2015) in their study, which is the network average of the local clustering coefficient. High values of it suggest poor overall cooperation and low interconnectivity since players tend to communicate more with few nodes than with all. Comparing this index with the average of the 32 teams, higher values than average indicate poor performance.

The formula implemented by the NetworkX package in python is defined as follows (Average_Clustering — NetworkX 3.3 Documentation, s. f.):

$$\boldsymbol{C = \frac{1}{n} \sum_{v \in G} c_v} \tag{2.6}$$

Where *n* is the number of nodes in *G*.

- **Assortativity** measures the tendency of the nodes (players) to connect with other nodes with a similar degree. A high positive value means that players tend to pass the ball to those players with a similar pass frequency. It helps to understand connectivity patterns and how the network is distributed in terms of cohesion. It is calculated through the standard Pearson correlation coefficient (Degree_Assortativity_Coefficient — NetworkX 3.3 Documentation, s. f.) (Newman et al., 2002):

$$\boldsymbol{r = \frac{\sum_{xy} xy(e_{xy} - a_x b_y)}{\sigma_a \sigma_b}} \tag{2.7}$$

Where:

e_{xy} represents the frequency of the links between players with a similar number of connections (passes).

xy is the product between the number of links *x* of one node (player) and the number of links *y* of another player.

a_x, b_y are the fraction of the links that start and end in players with degree *x* and *y* respectively. Since we work with undirected graphs, these values are the equal.

σ_a, σ_b are the standard deviations of the distributions a_x, b_yThis measure is focus on the similarity of the degree of the links (passing frequency) between players. It is more a measure of the homogeneity of connections between players with similar characteristics in terms of passes, that is, players within a similar "level".

Microscale metrics

The significance of each player's role and their immediate connections has been identified by: its degree, but in this case, we used two more elaborate metrics to gain better insights of the player's performance within the team. These are the PageRank and the Eigenvector centrality. Additionally, it has been also studied the betweenness centrality and closeness.

- **PageRank** metric helps identify core players. Basically, what helps to recognize whether a key player receives passes from other key players as well. It focuses on the relative importance of a player within the passing network. This importance of a player is based on the quality (and not just the quantity as was the case with the assortativity measure) of the connections he receives.

Although it can be condensed in a formula, the way it is computed is through an iterative algorithm. When the algorithm starts, all of the nodes have assigned the same PageRank value: 1/N. The package and formula used in the code was *NetworkX* from Python, *pagerank* function. The maximum number of iterations was 100. Its mathematical expression is the following:

$$\boldsymbol{PR_i = \frac{1 - p}{N} + p \sum_{j \in M(i)} \frac{PR_j}{L_J}} \tag{2.8}$$

Where:

PR_i is the PageRank value of player *i*

p is the damping factor, which represents the probability of a player passing the ball to another player following the current structure of the network. By default, in the network package, it is 0.85.

$1 - p$ is the probability of doing a random pass and not following the current structure of the network.

N is the total number of players.

Therefore, the first term of the formula distributes the probability of each node (player) to receive a pass from another player.

PR_j is the PageRank of player *j*

L_j is the total number of passes that player *j* pass the ball to other players.

$M_{(i)}$ is the set of players that pass the ball to player *i*,

The example below illustrates how the algorithm starts and calculate the metric in the first step:

Page Rank (k = 1)					
	A	B	C	D	E
Old	1/5	1/5	1/5	1/5	1/5
New	4/15	2/5	1/6	1/10	1/15

A: (1/3)*(1/5) + 1/5 = 4/15
B: 1/5 + 1/5 = 2/5
C: (1/3)*(1/5) + (1/2)*(1/5) = 5/30 = 1/6
D: (1/2)*(1/5) = 1/10
E: (1/3)*(1/5) = 1/15

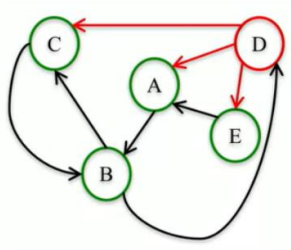

Figure 7 - Graphic representation of the PageRank metric

To calculate the PageRank value in the first iteration (k=1) for node A we must check how many nodes point to it. In this case, only two nodes point to node A (D and E). Since node A is influenced by the PageRank value of the nodes in touch with it, we need to see *how much* of that value will receive according to the current structure of the network. According to this, we can see with the red arrows from node D point also to nodes E and C apart from node A. Therefore, our node (node A) is going to receive *only* 1/3 of the PageRank value of D. Similarly, we do the same with the other node pointing to A, node E. In this case, node A is the only one connected to node E, therefore A is going to get the whole 1/5 from E. This is how the 4/15 displays in the A column of the table above is obtained.

Even though the metric was designed for directed graphs, the algorithm under the python package used does not check if the input graph is directed. It "will execute on undirected graphs by converting each oriented edge in the directed graph to two edges" (NetworkX, n.d.).

- **Betweenness centrality** represents the number of paths that pass through a given node. Said it differently, the periodicity by which a node is "used" to reach another node. This measure tries to find how relevant a player is so that the ball flows properly within the team. It helps to understand the consequences of changing a player during the match. High score of this index is associated with the most central vertices of the network and hence an important influence.

  Looking at the betweenness of all the players of a team, it seems reasonable to assume that the more homogeneous these values are, the more distributed the team is and all players seem to be equally relevant. However, when some players are above the team average value, sending off or changing a player may cause worse team performance.

  Below an illustration of what the betweenness centrality measures.

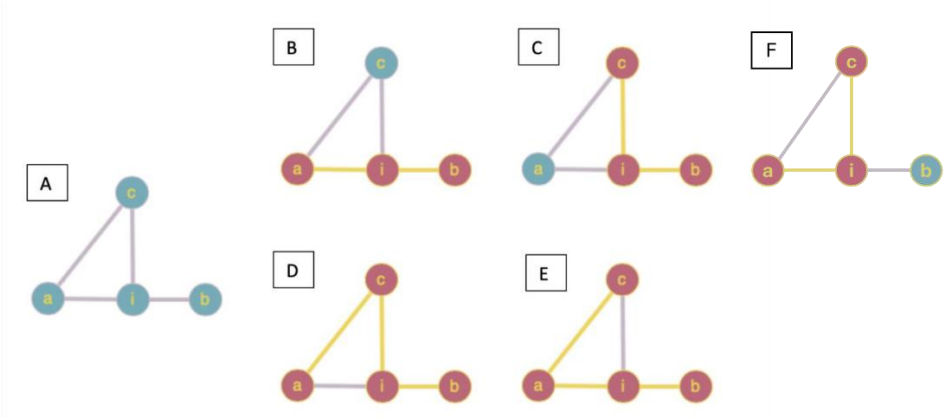

Figure 8 - Graphic representation of the betweenness centrality

Having our base network in figure A, we can see in figures B, C, D, E and F that there exist four different possible paths that pass-through node *i*. However, only the paths from figures B, C and F are considered by betweenness definition since they are the shortest ones (node *i* is connected by two edges in figures B, C and F against the three edges that connects node *i* in examples D and E).

Mathematically, it is defined as follows:

$$\boldsymbol{B_i} = \sum_{\boldsymbol{j \neq k \neq i}} \frac{\boldsymbol{n^i_{j,k}}}{\boldsymbol{g_{j,k}}} \tag{2.9}$$

Being:

B_i is the betweenness centrality metric of player *i.*

$n^i_{j,k}$ is the number of shortest geodesic paths between *j* and *k* passing through *i.*

$g_{j,k}$ is the total number of geodesic paths between *j* and *k.*

Therefore, in this example the shortest paths passing through *i* in the different figures are:

Graphs B, C and F

- Path $a \rightarrow i \rightarrow b$
- Path $b \rightarrow i \rightarrow c$
- Path $a \rightarrow i \rightarrow c$

Graph D and E

- Path $a \rightarrow c \rightarrow i \rightarrow b$
- Path $c \rightarrow a \rightarrow i \rightarrow b$

Since paths passing through *i* in graphs D and E are formed by three edges instead of two, those paths do not comply with the geodesic path definition and therefore, they are not considered in the betweenness centrality measure. And hence:

$$\boldsymbol{B_i = \frac{n^i_{a,b}}{g_{a,b}} + \frac{n^i_{a,c}}{g_{a,c}} + \frac{n^i_{b,c}}{g_{b,c}} = \frac{1}{1} + \frac{1}{1} + \frac{1}{1} = 3}$$

- **Closeness** measurement points how close are one node to another. In terms of football, it shows how well positioned a player is, that is, how easy is to reach that player with a pass. Thus, when higher the closeness value, the lower the distance to the teammates which means better connectivity. It can be calculated with the following form:

$$\boldsymbol{C_i = \frac{N-1}{\sum_{i \neq j} l_G(v_i, v_j)}} \tag{2.10}$$

Where:

C_i is the closeness centrality measure of player *i*.

N is the number of nodes of the graph.

$l_G(v_i, v_j)$ represents the length of the shortest path (in terms of number of edges) between nodes *i* and *j*.

- **Eigenvector** measures the number of edges of a node plus the number edges of the nodes connected to it. This is how it can be known the influence of a node in the entire network. Therefore, a network with strong links and relevant nodes will have high eigenvector values. This measure can be defined as follows:

$$\boldsymbol{v = \frac{1}{\lambda} A v} \tag{2.11}$$

Where:

v is the vector of the eigenvector centrality measure we want to find

A is the adjacency matrix of the network, where $A_{i,j}$ represents the weight of the pass from player *i* to player *j*. In an undirected graph, it is the total number of passes between player *i* and *j*.

λ as eigenvalue associated with *v*, and it is the dominant eigenvalue in the adjacency matrix.

2.3.3 Visualizations

The arrangement of a team can be viewed with respect to the connection between its players, leading to a network analysis of passes presented as undirected or directed graphs, with the weights as the numbers of passes between spatially embedded links (Leela et al., 2024). Since the structure of the network is constantly changing with the movement of the players as the game progresses, a challenge when plotting data related to the passes, is the node location. There exist different techniques to plot the network and each of them gives different information regarding the players and the games

Given the amount of data retrieved from a football match, not only in terms of events, but also tracking data of the players, drawing and analyzing the network formed from all the passes taking place during a match is only one way to obtain some insights about the team's behavior during the match. In this regard, some of the possibilities to plot the network are *average player position*, *closeness between players*, and *offensive formation*.

- **Average position**: on this technique nodes are plot based on the average location of each player when doing/receiving a pass. It helps to identify the strategy followed by the team during the match. If nodes and edges are weighted in terms of the number of passes, these visualizations can give insights also about the key players. In the picture below we can see how the game was mainly in the middle of the pitch with a strong relevance of Mariona Caldentey and Ona Batlle giving assistance to forward Jennifer Hermoso meaning the game of Spain was more on the left side of the pitch. Also, Oihane Hernández and Aitana Bonmati play a role, connecting both sides through Irene Hernández.

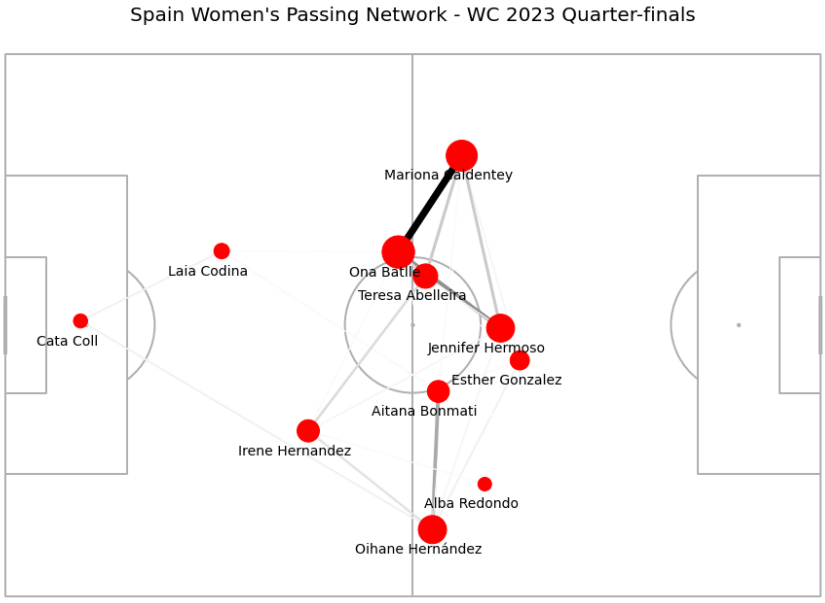

Figure 9 - Example of average position visualization

- **Closeness between players:** this method shows the relationship between subgroups of players. This graph is based on the pass's frequency: if some players do passes regularly during the game, this group of players will be displayed closer. Usually, the node size depends on the closeness coefficient. By contrast, this technique is unable to show clearly the strategy followed by the team because the player's location is unused and hence, the network plot is not on the pitch. Below a recycled graph as an example of how these visualizations resemble. The numbers represent the t-shirt number of each player, and the letters represent the end of the play: B =

shot on target, G = goal, F = shot off target, and S = no shot. The nodes with less frequency of passes seem to be node 1 and 8, which might be the goalkeeper and a substitute player. On the other hand, it seems the rest of the players played well connected with passes between all of them, although we cannot see the strategy followed.

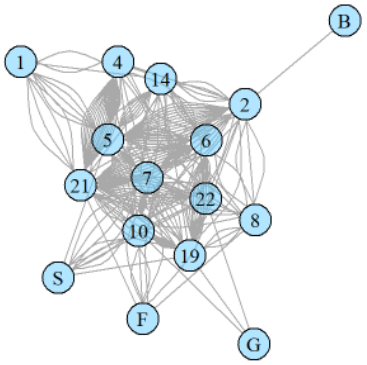

Figure 10 – Example of closeness between players visualization

- **Offensive formation:** this type of visualization is based on the movements of each player (when he or she is in possession of the ball), not on the events (passes, shots, goals, etc). This technique is known as *tracking*, and it helps to analyze the individual position of the player, relative to the team during the match. Nodes are plot based on the initial formation and the relative distance between the components of a team represents their positions when the team has possession. The size of the node can be represented by the clustering coefficient. The bigger the node the larger the number of motifs the player participates on. Herrero Candela (2022) in her project illustrates and describes this type of visualization as we can see in Figure 11. In these figures, the substitute players are also displayed and aligned with the replaced player. With this graph we can notice how the game occurs more backwards toward the sides with almost no relevance of the forward players.

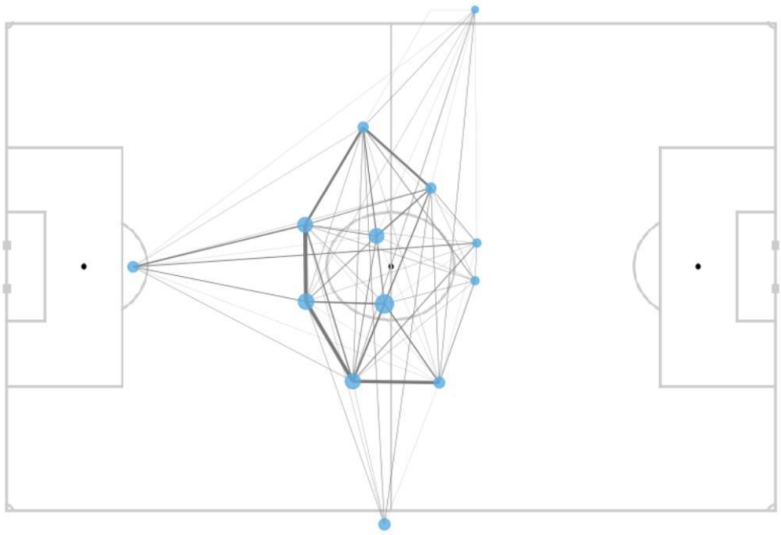

Figure 11 - Example of offensive formation visualization

In this section, most of the figures have been adapted from other studies, and the data source comes from a different dataset than the one used in the rest of the study. However, all the figures displayed in the upcoming sections are of own elaboration with data provided by StatsBomb.

To provide a simple and clear representation of the players on the pitch, and given the data available from the source system, the technique used to display the nodes of the networks is based on the average position of each player during the soccer game.

The open-source environment used for the code development was Spyder and the programming language Python 3.11. The different *.json* files regarding competitions, matches and events were loaded and transformed into individual data frames. The main packages and functions from python used for the data processing and visualization are: *json, pandas, matplotlib.pyplot* and *networkX*

2.4 Variables and Objective of this Study

The variables considered in this study are of two types: graph performance variables, which are the metrics derived from the passing network namely 1) total number of links, 2) diameter, 3) density, 4) assortativity, and 5) global clustering coefficient. Plus, the overall team performance variables understood as the maximum stage that each team reached in each of the tournaments considered. In this case we will call them categories. These can be: 1) Group Stage, 2) Round of 16, 3) Quarterfinals, 4) Semifinals, 5) 3rd Place Final, and 6) Final.

As it was advanced, the main objective of the present paper is to analyze and compare the differences on network graph variables between male and female teams that reached different stages during the competition.

a) Teams that achieve the highest stages in competition were expected to show high values of density and low values in clustering coefficient and diameter.

b) Main differences (if any) between male and female games.

2.5 Statistical Procedures

A preliminary analysis was conducted using basic descriptive statistics derived from the aggregated results of the macroscale metrics obtained from the national team's networks. Tables present the mean and the standard deviation of each metric per country. This initial analysis provides some hints about the dispersion and distribution of our data, which will thereafter be validated through the application of statistical procedures.

For this purpose and to continue with the objective of the study, three different techniques were tested. Firstly, and following the example of Clemente et al. (2015), we used the one-way ANOVA model to determine if the mean values of the aggregated network metrics varied among the different groups considered, in this case each stage of the competition. Initially, the model was executed for each metric and each season and then, joining the data from both seasons, assuming teams from same country in different competitions can be considered different teams.

Secondly, and with the aim of maximizing the utility of the available data, we opted to work with the raw data from each match using two distinct approaches. In this new context, we observed both *within-team* and *across-team* variability, as the same teams were measured throughout different phases of the competition, introducing more than one source of random variability. This led us to data highly correlated. Consequently, the methods used to account for these sources of variation were a Generalized Linear Mixed Model for a binary response variable, that models the probability of a binary event to occur - passing or not the Group Stage (variable named as *passed_GS*) controlling for the variability of the random effects – namely the teams.

Lastly, a clustering technique by teams was used, applying the Mahalanobis distance instead of the common Euclidean distance. The Mahalanobis distance measured the span between a point and a distribution, accounting for correlation between variables to assess the likelihood of grouping teams according to their characteristics.

All statistical analyses were performed using R packages: *lme4*, *emmeans*, *biotools* and *NbClust* and results are considered statistically significant at a significance level of $p < 0.05$.

2.5.1 One-way ANOVA

To explore real deviations between mean values of the network metrics in different stages of the competition it was implemented the One-way ANOVA model. This model is a particular case of the Linear Model. Therefore, it is assumed that the random errors ε_{ij} have the same properties as in the LR model (Antunes, 2022).

The model can be written as:

$$\boldsymbol{Y_{ij} = \mu_1 + \alpha_i + \varepsilon_{ij}, \quad \forall i = 1, \dots, k \quad , j = 1, \dots, n, (\alpha_1 = 0)} \tag{2.12}$$

Where:

Y_{ij} is the value of the specific network metric for the team *j* in the competition stage *i*. It is assumed to be independent.

In our case, $k = 5$ and $j = 1, \dots, 32$, $i = 1, \dots, 5$[5]

μ is the global mean

α_i is the effect of the competitions stage *i*.

$\varepsilon_{ij} \cap \mathcal{N}(0, \sigma^2)$ independent (these are i.i.d r.v.). It is the error associated to the *j*-th team of the competition stage *i*.

When ANOVA shows significant statistical differences between competition stages (p-value < 0.05), it was performed a Tukey's HSD post-hoc test to evaluate the stages where there is a significant difference.

Tukey's HSD post-hoc test

This test helps to identify which specific pairs present statistically significant differences after finding the general difference with the ANOVA test.

Let $\{W_i\}_{i=1}^{k}$ be independent random variables with normal distribution of the same parameters:

- $W_i \cap N(\mu_W, \sigma_W^2), \forall i = 1, \dots, k.$
- Let $R_W = \max_i W_i - \min_i W_i$ be the range of the total sample.
- Let S_W^2 be an estimator of the common variance, σ_W^2, such that $\frac{\nu S_W^2}{\sigma_W^2} \cap \mathcal{X}_\nu^2$
- Let S_W and R_W be independent r.v.

Then, the studentized range, $\frac{R_W}{S_W}$, follows a Tukey distribution, which depends on two parameters: *k* and *v* (Antunes, 2022).

[5] For 2019 Women's World Cup $j = 1, \dots, 24$ and $i = 1, \dots, 4$.

The formula to calculate the critical value of the TukeyHSD test is:

$$\boldsymbol{HSD = q\,x\,\sqrt{\frac{MS_{ERROR}}{n}}} \tag{2.13}$$

Where:

q is the Tukey distribution value for a given significance level and the error degrees of freedom.

MS_{ERROR} is the mean square of the residuals obtained in the ANOVA test.

n is the sample size per group.

2.5.2 Generalized Linear Mixed Model for a Binary response variable

Generalized linear mixed models - also known as generalized linear mixed effects models (GLMMs) assume that a response variable *y* follows a known parametric distribution $p(y|\mu)$ and that a parameter μ of this distribution (usually the mean) is related to the sum of so-called fixed effects $X\beta$ and random effects *Zu:*

$$\mathbf{y} \sim \mathbf{p}(\mathbf{y}|\boldsymbol{\mu})$$
$$\boldsymbol{\mu} = \mathbf{f}(\mathbf{X\beta} + \mathbf{Zu}) \tag{2.14}$$

In this case, *f()* is a link-function that ensures that $f(X\beta + Zu)$ is in the proper range, in the case of binary data, the mean must be between 0 and 1.

Therefore, GLMMs are an extension of linear mixed models to allow response variables from different distributions (hence the term *General*), such as in our case binary responses. Likewise, these models can be seen as an extension of generalized linear models to include both fixed and random effects (hence the term *Mixed*). According to this, the form of the model, in matrix notation, looks like follows:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Zu} + \boldsymbol{\varepsilon}^{6} \tag{2.15}$$

Where:

y is the response variable, and it is a *N* x 1 column vector.

$X\beta$ are the fixed effects, where:

X is a *N* x $(p+1)$ outcome variable matrix with *p* predictor variables

β is a *p* x1 column vector of the fixed-effects regression coefficients to be estimated.

Zu are the random effects, where:

Z is a *N* x *q* design matrix for the *q* random effects (the random complement to the fixed X)

u is *q* x 1 vector of the random effects (the random complement to the fixed β) and is assumed to follow a multivariate normal distribution

ε is a *N* x 1 column vector of the residuals, that part of the response variable *y* is not explained by the model $X\beta + Zu$

[6] (*Introduction To Generalized Linear Mixed Models*, s. f.)

What makes GLMMs so powerful is that they not only help to understand direct effects of the study variables, but also other random effects. This is what distinguishes them from a generalized linear model. Random effects can consist of, for instance, grouped (aka clustered) random effects with a potentially nested or crossed grouping structure (Sigrist, 2024).

For a better context, the table below displays a small sample of our dataset with each of the variables matching the corresponding model term.

Table 1 - Sample dataset variables with the corresponding model term

Zu (18 x 1)	*X* 18 x 6 (5 + intercept)					*y* (18 x 1)
Team	**Links**	**Density**	**Diameter**	**Assortativity**	**GlobalClusCoeff**	**passed_GS**
Argentina	70	0.769	2	-0.249	0.854	1
Argentina	68	0.747	2	-0.271	0.828	1
Argentina	62	0.795	2	-0.281	0.858	1
Argentina	69	0.758	2	-0.242	0.808	1
Argentina	86	0.717	2	-0.158	0.83	1
Argentina	76	0.633	3	-0.198	0.822	1
Argentina	74	0.705	2	-0.141	0.763	1
Argentina	78	0.65	3	-0.164	0.768	1
Argentina	89	0.654	2	-0.261	0.763	1
Argentina	71	0.592	3	-0.017	0.713	1
Argentina	77	0.566	3	-0.227	0.73	1
Australia	52	0.571	2	-0.347	0.671	0
Australia	66	0.725	2	-0.373	0.851	0
Australia	66	0.725	2	-0.293	0.807	0
Australia	63	0.6	3	-0.144	0.694	1
Australia	69	0.575	2	-0.309	0.752	1
Australia	66	0.55	3	0.201	0.576	1
Australia	74	0.617	2	-0.125	0.746	1

In our study, we want to evaluate the effect of the different network metrics on the group and the knock-out phases of the competition. That is why our binary response variable is *passed_GS* which reflects whether a team passed the Group Stage or not with 1 and 0 respectively. The different predictors are the measurements: *Links*, *Density*, *Diameter*, *Assortativity* and *GlobalClusCoeff* and we want to study how these metrics vary across the different teams, therefore, our random effect in this case is the categorical variable *Team*.

2.5.3 Clustering by teams – Hierarchical clustering using the Mahalanobis distance

An alternative approach for studying how network metrics differ at each competition stage was to cluster the teams based on their metric values. The expectation was that clusters would group teams that reached the same phase of the tournament, revealing some correlation between the metrics and the stages.

On top of that, to identify macro-level patterns for each competition, and since differences at group level - as teams - are more relevant than match differences, to perform the clustering, it was used the mean values of the aggregated network metrics.

The chosen clustering type was hierarchical clustering. This type of classification, contrary to other classification approaches, consists of a series of partitions in a hierarchical order, as it name states. It may run from a single cluster containing all individuals, to *n* clusters each containing a single individual – known as the *divisive* technique. Alternatively, it can proceed through a series of successive fusions of *n* individuals into groups – the *agglomerative* technique (Everitt et al., 2011). Since data are not partitioned into a particular number of clusters at a single step, the investigator needs to come up with

an "optimal" number of clusters so the classification can be done. Any of the techniques used for hierarchical clustering may be represented by a two-dimensional diagram called *dendrogram* (illustrated in Figure 12). In this context, the method used for hierarchical clustering was the method called Ward.D2, which is an agglomerative method that minimizes the increase in the total within-cluster sum variance. At each step the pair of clusters with minimum cluster distance are merged.

The rationale behind choosing the hierarchical clustering with Ward's method was based on the structure of our data – we did not know in advance the number of clusters, and we wanted the clusters as compacted as possible. Since Ward's minimizes the within-cluster variance, the resulting clusters are more homogeneous and well-differentiated.

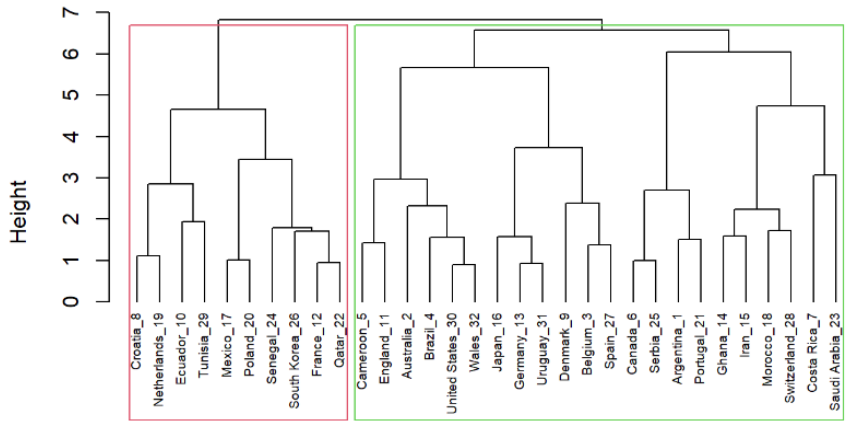

Figure 12 - Example of a dendrogram with the optimal number of 2 clusters

Furthermore, as mentioned previously, instead of employing the Euclidean distance, we used the distance calculation method introduced by P.C. Mahalanobis in 1936 (Wikipedia contributors, 2024).

The Mahalanobis distance is a measure between two data points in the space defined by relevant variables. More precisely, it is a measure of the distance between a point (vector) *P* and a distribution *D*, where the space between *P* and *D* is determined by the principal characteristics of the data. Since it accounts for unequal variances as well as correlations between variables, it will adequately evaluate the distance by assigning different weights or important factors to the variables of data points. This allows the distance to reflect how point *P* deviates from the center of the distribution *D* still considering the shape and orientation of *D*. Only when the variables are uncorrelated, the distance under a Mahalanobis distance metric is identical to that under the Euclidean distance metric. In addition, geometrically, a Mahalanobis distance metric can adjust the data points that are close to each other in the variable space so that the distance between similar data points is small (Xing et al., 2002). Thus, it can enhance the performance of clustering or classification algorithms. Such advantages can be used to perform special tasks on a given data set, if given a suitable Mahalanobis distance metric (Xiang et al., 2008).

Consequently, it is a multivariate equivalent of the Euclidean distance. According to Prabhakaran (n.d.), computationally, these two distances differ in the following three steps of how the Mahalanobis distance is obtained:

1. Transform the variables into uncorrelated ones
2. Scale the variables to make their variance equal to 1
3. Calculate de Euclidean distance.

The formula takes this form:

$$\boldsymbol{D^2 = (x - m)^T \cdot C^{-1} \cdot (x - m)} \tag{2.16}$$

Where,

D^2 is the square of the Mahalanobis distance.

$(x - m)$ is the distance of the vector from the mean where,

x is the vector of the observation (row in a dataset),

m is the vector of mean values of independent variables (mean of each column),

C^{-1} is the inverse covariance matrix of independent variables.

Taking a closer look, we can see the multivariate equivalent of a regular standardization: the distance of the vector from the mean is multiplied by the inverse of the covariance (or divided by the covariance matrix) ($z = (x - \mu) / \sigma$),

Using the inverse of the covariance, adjust the distance to account for the relationships between variables. This guarantees that the calculated distances reflect the correlation structure of the data, providing a precise measure of the distance between the points in the variable space.

3 Results

In this section we will deep dive into the results obtained from the methods explained. In part 3.1 of this chapter, it will be presented the passing networks graphs derived from the different matches that took place during the 2018 to 2023 World Cup competitions for both men's and women's teams. Given that match results are known, it will be easier to analyze team's strategies based on the network graphs and metrics. For simplicity, only some examples of passing networks figures will be discussed. This includes those belonging to the final stage of the most recent years, that is the Men's World Cup in 2022 and the Women's World Cup in 2023. The network graphs are accompanied by the corresponding microscale metrics table. The rest of the figures and tables can be found in the Annexes. Part 3.2 contains the results obtained in the ANOVA, GLMM, and hierarchical clustering approaches.

Nevertheless, before going through the network analysis, Tables 6 through 9 display the aggregated results of the macroscale metrics derived from the national team's networks across each of the four tournaments. Tables shows the mean and the standard deviation of each metric per each country. There is an additional column named *Index* in which is reflected the maximum stage of the competition reached by each team, being 6 the Final and 1 the Group Stage (number 4, semi-final, is missing because those that did not reach the Final played the 3rd Place Final, represented by the number 5). The tables present in descent order the final rank of the championship. The lower half displays the teams that did not pass the group phase.

In Table 2 can be noted that none of the best teams displays higher or lower values in any of the metrics. According to the theory, and if we didn't know the results of the competition, those national teams wouldn't have any chance to be in the semi-finals.

Table 2 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2018 (Male)[7]

	Index	Total Links		Density		Diameter		Assortativity		Global Clustering Coefficient	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Croatia	**6**	**64.857**	8.726	**0.696**	**0.056**	**2.143**	**0.378**	**-0.223**	**0.086**	**0.161**	**0.017**
France	**6**	**58.857**	**5.984**	**0.674**	**0.046**	**2.286**	**0.488**	**-0.238**	**0.121**	**0.163**	**0.046**
Belgium	**5**	**62.143**	**5.047**	**0.745**	**0.043**	**2.143**	**0.378**	**-0.243**	**0.082**	**0.134**	**0.043**
England	**5**	**67.571**	**5.503**	**0.746**	**0.070**	**2.286**	**0.488**	**-0.218**	**0.067**	**0.134**	**0.032**
Brazil	3	66.800	2.775	0.734	0.031	2.000	0.000	-0.202	0.142	0.136	0.017
Russia	3	68.600	6.427	0.712	0.072	2.200	0.447	-0.195	0.039	0.168	0.036
Sweden	3	56.000	4.528	0.615	0.050	2.200	0.447	-0.180	0.089	0.178	0.048
Uruguay	3	64.400	4.159	0.707	0.046	2.200	0.447	-0.197	0.102	0.171	0.029
Argentina	2	67.250	3.594	0.767	0.021	2.000	0.000	-0.261	0.018	0.141	0.029
Colombia	2	66.250	6.021	0.702	0.052	2.250	0.500	-0.242	0.016	0.141	0.026
Denmark	2	65.750	11.701	0.747	0.059	2.000	0.000	-0.152	0.065	0.159	0.022
Japan	2	61.750	2.872	0.705	0.030	2.500	0.577	-0.184	0.103	0.147	0.044
Mexico	2	65.500	5.447	0.720	0.060	2.000	0.000	-0.182	0.077	0.142	0.044
Portugal	2	63.500	3.697	0.698	0.041	2.500	0.577	-0.190	0.094	0.202	0.018
Spain	2	70.000	4.899	0.741	0.033	2.000	0.000	-0.197	0.068	0.122	0.037
Switzerland	2	60.000	5.477	0.739	0.049	2.250	0.500	-0.262	0.056	0.159	0.016
Australia	1	61.333	8.083	0.674	0.089	2.000	0.000	-0.338	0.041	0.095	0.022
Costa Rica	1	59.667	8.963	0.655	0.099	2.667	0.577	-0.060	0.181	0.195	0.041
Egypt	1	63.000	1.732	0.692	0.019	2.000	0.000	-0.269	0.022	0.155	0.033
Germany	1	68.667	4.163	0.754	0.046	2.000	0.000	-0.267	0.065	0.125	0.005
Iceland	1	58.333	2.887	0.641	0.032	2.667	0.577	-0.251	0.072	0.211	0.014
Iran	1	52.333	5.033	0.575	0.055	3.000	0.000	-0.136	0.033	0.203	0.013
Morocco	1	62.667	2.887	0.688	0.032	2.333	0.577	-0.187	0.133	0.151	0.039
Nigeria	1	59.000	4.000	0.684	0.076	2.667	0.577	-0.185	0.109	0.181	0.044
Panama	1	63.000	4.359	0.692	0.048	2.333	0.577	-0.222	0.066	0.172	0.021
Peru	1	66.000	5.568	0.725	0.061	2.000	0.000	-0.208	0.079	0.162	0.011
Poland	1	62.333	3.786	0.720	0.021	2.000	0.000	-0.189	0.053	0.131	0.048
Saudi Arabia	1	60.667	7.234	0.701	0.068	2.000	0.000	-0.246	0.092	0.151	0.010
Senegal	1	59.000	8.660	0.648	0.095	2.667	0.577	-0.209	0.055	0.153	0.008
Serbia	1	59.667	1.528	0.729	0.080	2.333	0.577	-0.156	0.091	0.189	0.065
South Korea	1	62.667	6.110	0.688	0.067	2.333	0.577	-0.199	0.089	0.171	0.034
Tunisia	1	64.667	5.132	0.710	0.056	2.333	0.577	-0.208	0.091	0.162	0.022

[7] Values in yellow represent the 3 highest values. Values in grey represent the 3 lowest values. Rows in bold are the semi-finalists and finalists.

For both male competitions, in general terms, we can observe that the teams that didn't pass the group phase have higher diameter values, and indicative of a less cohesive game. This, accompanied with lower values in assortativity, density and total links suggest teams less connected. Moreover, we can observe higher dispersion in the density and diameter metrics, which might indicate that these teams were more irregular in terms of connectivity and distribution of the network during the three matches of the group stage. Meaning that they couldn't follow a clear winning strategy or that they couldn't find a way to flow their game. Additionally, taking a look to the clustering coefficient, some of the highest values are found in those teams that performed worse. These values could indicate that teams tend to form triangles well connected but provoking a lower general connection of the whole team turning into a general worse performance.

Table 3 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2022 (Male)

	Index	Total Links		Density		Diameter		Assortativity		Global Clustering Coefficient	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Argentina	**6**	**78.714**	**6.473**	**0.645**	**0.055**	**2.571**	**0.535**	**-0.167**	**0.078**	**0.131**	**0.017**
France	**6**	**70.714**	**13.313**	**0.701**	**0.054**	**2.429**	**0.535**	**-0.180**	**0.094**	**0.139**	**0.048**
Croatia	**5**	**74.143**	**9.406**	**0.680**	**0.060**	**2.429**	**0.535**	**-0.127**	**0.172**	**0.116**	**0.028**
Morocco	**5**	**67.714**	**9.517**	**0.596**	**0.059**	**2.714**	**0.488**	**-0.189**	**0.082**	**0.123**	0.012
Brazil	3	77.800	3.701	0.648	0.031	2.200	0.447	-0.173	0.086	0.095	0.024
England	3	73.400	10.644	0.678	0.037	2.000	0.000	-0.185	0.104	0.110	0.015
Netherlands	3	72.800	9.960	0.672	0.032	2.600	0.548	-0.160	0.046	0.105	0.032
Portugal	3	78.600	2.408	0.655	0.020	2.400	0.548	-0.126	0.123	0.156	0.018
Australia	2	68.000	4.690	0.586	0.029	2.500	0.577	-0.094	0.213	0.123	0.030
Japan	2	70.750	10.243	0.607	0.063	2.250	0.500	-0.236	0.097	0.143	0.047
Poland	2	63.750	3.304	0.640	0.110	2.500	0.577	-0.205	0.058	0.119	0.017
Senegal	2	69.500	6.557	0.668	0.075	2.250	0.500	-0.221	0.071	0.166	0.061
South Korea	2	71.000	8.485	0.724	0.048	2.250	0.500	-0.248	0.025	0.157	0.021
Spain	2	80.750	5.315	0.676	0.066	2.000	0.000	-0.257	0.080	0.071	0.021
Switzerland	2	73.750	6.602	0.634	0.028	2.500	0.577	-0.238	0.050	0.140	0.043
United States	2	74.500	4.726	0.641	0.016	2.250	0.500	-0.139	0.211	0.107	0.024
Belgium	1	76.333	8.386	0.664	0.059	2.333	0.577	-0.241	0.024	0.087	0.006
Cameroon	1	70.333	8.145	0.669	0.036	2.000	0.000	-0.148	0.116	0.139	0.034
Canada	1	84.333	4.726	0.703	0.040	2.000	0.000	-0.201	0.027	0.152	0.045
Costa Rica	1	67.333	7.371	0.613	0.033	2.000	0.000	-0.194	0.005	0.192	0.053
Denmark	1	78.333	8.386	0.711	0.020	2.000	0.000	-0.286	0.023	0.119	0.043
Ecuador	1	63.000	7.000	0.627	0.022	3.000	0.000	-0.204	0.047	0.117	0.059
Germany	1	78.333	8.505	0.653	0.071	2.000	0.000	-0.249	0.042	0.128	0.045
Ghana	1	64.667	7.767	0.561	0.030	3.000	0.000	-0.239	0.044	0.132	0.014
Iran	1	70.333	6.110	0.565	0.082	2.667	0.577	-0.254	0.062	0.151	0.039
Mexico	1	67.667	2.309	0.651	0.096	2.333	0.577	-0.242	0.013	0.122	0.044
Qatar	1	68.667	8.386	0.690	0.075	2.333	0.577	-0.198	0.140	0.145	0.038
Saudi Arabia	1	67.000	9.849	0.607	0.044	2.667	0.577	-0.188	0.046	0.196	0.053
Serbia	1	81.333	1.528	0.678	0.013	2.000	0.000	-0.214	0.064	0.165	0.035
Tunisia	1	65.000	12.490	0.621	0.123	3.000	1.000	-0.121	0.218	0.125	0.027
Uruguay	1	75.667	4.933	0.658	0.009	2.000	0.000	-0.234	0.050	0.145	0.034
Wales	1	70.333	3.215	0.641	0.021	2.333	0.577	-0.136	0.189	0.115	0.035

Results from the women's competitions are quite aligned with the ones described for men's. In 2023 there seems to be slightly more variations, with teams in the upper half of the table presenting also high diameter values either in the mean and the standard deviation.

Table 4 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2019 (Fem)[8]

	Index	Total Links		Density		Diameter		Assortativity		Global Clustering Coefficient	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
Netherlands Women's	**6**	**63.143**	2.911	**0.727**	**0.072**	**2.571**	**0.535**	**-0.188**	**0.073**	**0.146**	**0.035**
United States Women's	**6**	**65.714**	**6.422**	**0.722**	**0.071**	**2.286**	**0.488**	**-0.161**	0.138	0.127	**0.017**
England Women's	**5**	**66.571**	**4.541**	**0.731**	**0.050**	2.000	0.000	**-0.225**	**0.075**	**0.131**	**0.033**
Sweden Women's	**5**	**61.714**	7.653	**0.677**	**0.065**	**2.571**	**0.535**	**-0.156**	**0.116**	**0.150**	**0.024**
Australia Women's	2	66.500	7.594	0.756	0.041	2.250	0.500	-0.214	0.037	0.138	0.021
Brazil Women's	2	68.500	6.137	0.725	0.039	2.000	0.000	-0.247	0.056	0.135	0.008
Cameroon Women's	2	57.750	6.238	0.661	0.088	2.250	0.500	-0.248	0.116	0.184	0.045
Canada Women's	2	65.750	6.292	0.782	0.061	2.000	0.000	-0.178	0.044	0.203	0.051
China PR Women's	2	63.500	7.000	0.725	0.078	2.000	0.000	-0.244	0.088	0.165	0.013
France Women's	2	60.200	5.586	0.682	0.039	2.000	0.000	-0.285	0.064	0.123	0.019
Germany Women's	2	72.600	4.561	0.798	0.050	2.000	0.000	-0.210	0.035	0.152	0.058
Italy Women's	2	64.200	5.805	0.705	0.064	2.000	0.000	-0.201	0.091	0.181	0.039
Japan Women's	2	62.250	4.500	0.738	0.021	2.500	0.577	-0.151	0.092	0.129	0.033
Nigeria Women's	2	57.750	6.752	0.634	0.074	2.500	0.577	-0.201	0.115	0.225	0.077
Norway Women's	2	61.400	4.615	0.674	0.051	2.600	0.548	-0.118	0.129	0.134	0.028
Spain Women's	2	70.250	5.439	0.772	0.060	2.250	0.500	-0.258	0.069	0.167	0.036
Argentina Women's	1	54.667	9.713	0.627	0.064	2.000	0.000	-0.153	0.216	0.177	0.019
Chile Women's	1	56.667	3.786	0.693	0.089	2.333	0.577	-0.173	0.121	0.213	0.058
Jamaica Women's	1	64.333	2.082	0.707	0.023	2.000	0.000	-0.213	0.076	0.240	0.083
Korea Republic Women's	1	64.000	7.211	0.703	0.079	2.000	0.000	-0.253	0.105	0.125	0.014
New Zealand Women's	1	58.333	6.807	0.641	0.075	2.667	0.577	-0.094	0.193	0.172	0.025
Scotland Women's	1	54.333	8.327	0.705	0.101	2.333	0.577	-0.249	0.059	0.161	0.065
South Africa Women's	1	57.000	3.464	0.626	0.038	2.000	0.000	-0.283	0.070	0.199	0.023
Thailand Women's	1	54.000	3.606	0.624	0.038	2.667	0.577	-0.277	0.066	0.206	0.031

Table 5 - Descriptive statistics (mean and standard deviation) of network performance per national team in WC 2023 (Fem)

	Index	Total Links		Density		Diameter		Assortativity		Global Clustering Coefficient	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD	Mean	SD
England Women's	**6**	**65.714**	**10.531**	0.752	**0.037**	**2.286**	**0.488**	**-0.172**	**0.110**	**0.128**	**0.061**
Spain Women's	**6**	76.286	**11.146**	**0.670**	**0.033**	2.000	0.000	**-0.230**	0.048	**0.106**	**0.023**
Australia Women's	**5**	**57.714**	**7.566**	**0.676**	**0.065**	**2.286**	**0.488**	**-0.185**	**0.104**	**0.131**	**0.032**
Sweden Women's	**5**	**64.857**	**10.715**	**0.628**	**0.046**	**2.429**	**0.535**	**-0.171**	**0.087**	0.104	**0.028**
Colombia Women's	3	55.800	6.017	0.674	0.073	2.600	0.548	-0.174	0.146	0.172	0.058
France Women's	3	63.200	4.382	0.770	0.109	2.400	0.894	-0.166	0.085	0.164	0.045
Japan Women's	3	69.400	8.562	0.689	0.100	2.000	0.000	-0.167	0.048	0.114	0.009
Netherlands Women's	3	75.600	9.607	0.720	0.052	2.200	0.447	-0.188	0.115	0.147	0.034
Denmark Women's	2	67.000	6.164	0.621	0.083	2.750	0.500	-0.255	0.047	0.110	0.043
Jamaica Women's	2	58.750	11.955	0.600	0.050	2.750	0.500	-0.063	0.076	0.216	0.084
Morocco Women's	2	55.000	7.118	0.589	0.117	2.500	0.577	-0.179	0.060	0.155	0.050
Nigeria Women's	2	63.250	8.098	0.667	0.038	2.000	0.000	-0.204	0.051	0.199	0.048
Norway Women's	2	67.500	9.469	0.641	0.051	2.250	0.500	-0.220	0.056	0.087	0.039
South Africa Women's	2	62.000	7.874	0.686	0.083	2.500	0.577	-0.177	0.088	0.169	0.012
Switzerland Women's	2	62.750	4.349	0.625	0.081	2.500	0.577	-0.211	0.050	0.147	0.056
United States Women's	2	67.250	9.251	0.710	0.184	2.750	0.957	-0.175	0.115	0.149	0.069
Argentina Women's	1	68.667	9.504	0.691	0.101	2.000	0.000	-0.270	0.051	0.165	0.070
Brazil Women's	1	72.333	8.505	0.656	0.027	2.667	0.577	-0.206	0.154	0.142	0.037
Canada Women's	1	76.333	5.508	0.609	0.021	2.333	0.577	-0.219	0.099	0.121	0.061
China PR Women's	1	61.333	6.028	0.611	0.025	3.000	0.000	-0.139	0.011	0.189	0.033
Costa Rica Women's	1	59.667	11.930	0.655	0.131	2.333	0.577	-0.221	0.078	0.151	0.082
Germany Women's	1	75.333	11.015	0.750	0.018	2.000	0.000	-0.189	0.116	0.085	0.030
Haiti Women's	1	61.000	11.533	0.642	0.114	2.667	0.577	-0.167	0.127	0.177	0.073
Italy Women's	1	74.000	10.149	0.703	0.018	2.000	0.000	-0.173	0.074	0.144	0.043
Korea Republic Women's	1	55.000	8.185	0.574	0.027	2.667	0.577	-0.276	0.079	0.117	0.038
New Zealand Women's	1	69.667	7.572	0.699	0.092	2.333	0.577	-0.216	0.151	0.144	0.082
Panama Women's	1	62.000	10.817	0.524	0.135	2.333	0.577	-0.294	0.101	0.175	0.035
Philippines Women's	1	55.000	1.000	0.506	0.075	3.000	0.000	-0.033	0.134	0.152	0.043
Portugal Women's	1	65.000	6.557	0.619	0.063	2.667	0.577	-0.157	0.171	0.128	0.022
Republic of Ireland Women's	1	59.000	5.292	0.591	0.044	2.333	0.577	-0.277	0.082	0.178	0.037
Vietnam Women's	1	59.000	5.196	0.514	0.041	2.667	0.577	-0.168	0.157	0.139	0.027
Zambia Women's	1	61.667	3.215	0.649	0.084	2.333	0.577	-0.198	0.065	0.203	0.040

Despite all these conclusions we have drawn, it does not seem to be a clear pattern determining what makes a team a winner or a loser, contrary to what the literature claims. We will need to explore this further with our models.

[8] Please remember that only 24 teams participated in the Women's World Cup 2019 and data from quarterfinal was missing.

3.1 Passing Network Analysis

The World Cup is a competition held every four years. Teams have limited time to prepare, meaning that the strategy employed by a team can change a lot during the competition. This lack of time to create a thoughtful strategy makes data analysis very important in this type of tournament. Coaches must find a suitable starting eleven for each type of match, so data analysis can help them identify both the weaknesses of their opponents and the players with the best characteristics to attack those weaknesses.

Hereunder, it will be described the game of each finalist in the last World Cup competitions. Please, note the following considerations: 1. edges are filtered and only those with a weight greater than 3 are displayed, and 2. passing network graphs and tables display only the starting lineup.

Final MWC 2022: Argentina vs. France

The match ended with the victory of Argentina, but France closely contested the match. Whilst Argentina dominated the first half with a result of 2-0 at halftime. France turned the game around in the final 10 minutes, prolonging the match for another half an hour. The extra time was a replica of the match: Argentina scored first, and once again, France managed to come back scoring two minutes before the end. Eventually, Argentina won on the penalties by 4-2.

Figure 13 displays the passing network of Argentina until the first change, that took place in the minute 63. By that moment, the team was still winning. We can see how the team's average location is quite advanced, indicating a more offensive strategy with a strong connection between the two defenders. It helps to explain why France had so many problems to score and could not do it until the last minutes of the game.

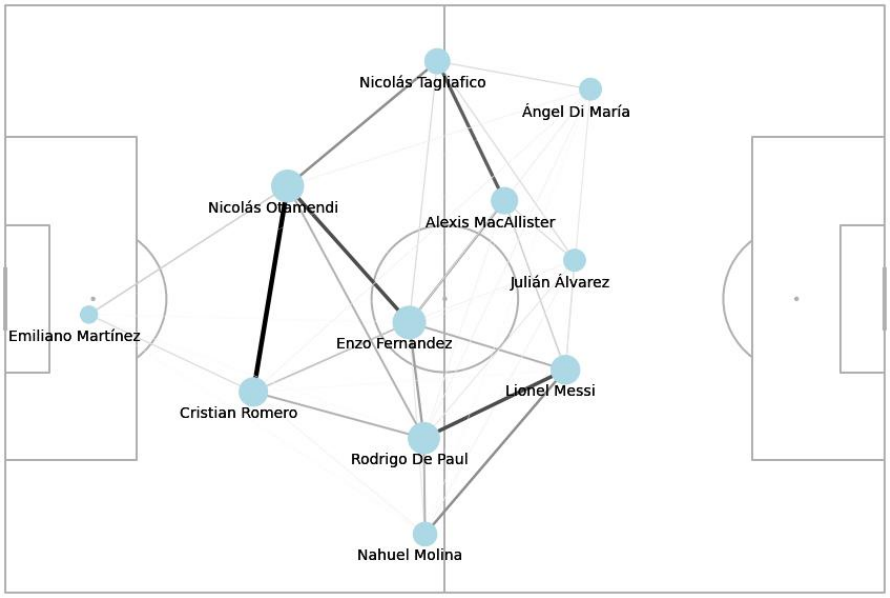

Figure 13 - Argentina Passing Network - WC 2022 Final

The size of the nodes seems to be similar between all the players, meaning that all of them do the same number of passes. Also, the network is well connected and distributed, a little bit sided to the right to assist Messi. Both, Otamendi and Romero seemed to distribute the game on both sides through Enzo Fernandez, key player positioned in the middle of the pitch.

The team was equally strong in both sides: with the MacAllister, Tagliafico and Otamendi on the left and Messi, de Paul and Molina on the right side. Forming a strong barrier in the back between Otamendi and Romero with Enzo Fernandez as the link.

Table 6, corroborates our suspicions from the network analysis: the most influential players are Enzo, Otamendi and De Paul – these are the ones with the highest PageRank values. However, it also provides additional relevant information that is not visible in the graph: the role of Di María. Di Maria presents the highest values in almost all measures, indicating that he is a key player in facilitating the ball's flow within the team. Additionally, and remembering the definitions of betweenness and closeness, these metrics help us to understand how well distributed and connected the team is. The higher the closeness values, the lower the distance between the players and hence the better the connectivity. Likewise, when the betweenness values are homogeneously distributed, it means all players are equally important, a characteristic we also noticed observing the network.

Table 6 - Argentina Microscale Metrics - WC 2022 Final[9]

Player Name	Eigenvector	Pagerank	Betweenness	Closeness
Alexis MacAllister	0.3178	0.0892	0.0206	**0.9091**
Enzo Fernandez	**0.3416**	**0.1282**	**0.0398**	**1.0**
Julián Alvarez	0.2853	0.0656	0.0178	0.8333
Lionel Messi	0.2984	0.1034	0.0028	0.8333
Nicolás Otamendi	0.3118	**0.1253**	**0.0326**	0.9091
Nicolás Tagliafico	0.1859	0.082	0.0028	0.6667
Rodrigo De Paul	**0.3222**	**0.1194**	0.0146	0.9091
Angel Di María	**0.3416**	0.0657	**0.0398**	**1.0**
Cristian Romero	0.3222	0.1014	0.0146	0.9091
Emiliano Martínez	0.2282	0.0452	0.0	0.7143
Nahuel Molina	0.3222	0.0745	0.0146	0.9091

Contrary to what we could see in the Argentinian network, France's play was installed at the back. There is a clear defense barrier formed by Koundé, Varane, Upamecano and Theo Hernández, who serves as link between the defenders and the main forward Mbappé. In line with the course of the match until the team managed to score. They had to defend the goal area from a very aggressive Argentina. Based on this, it seems reasonable that the defenders are represented with a slightly larger size.

Additionally, to assist the left side forward Mbappé, both Tchouameni and Rabiot, along with Theo Hernández, also serves as links to the player. We see a well-connected team, trying to find opportunities to came back to the game, but with the priority of defending.

[9] The highlighted cells are the three highest values for each metric.

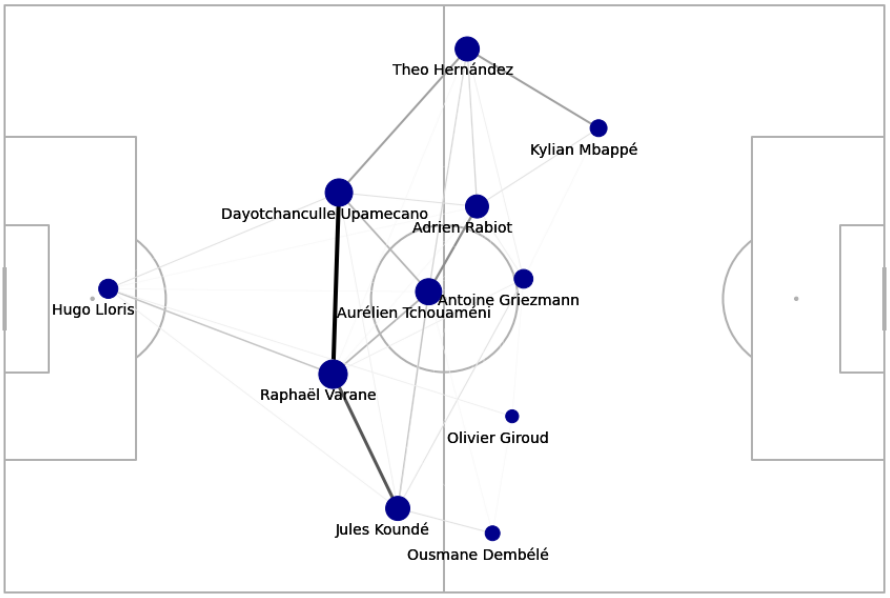

Figure 14 - France Passing Network - WC 2022 Final

The metrics, in this case, confirm our analysis from the network. Similarly, the betweenness values are also quite homogeneous indicating a well distributed network and the high values of closeness, in like manner as Argentina, means the team is well connected.

Table 7 - France Microscale Metrics - WC 2022 Final

Player Name	Eigenvector	Pagerank	Betweenness	Closeness
Adrien Rabiot	**0.3373**	0.0995	**0.0306**	**1.0**
Antoine Griezmann	**0.3373**	0.0711	**0.0306**	**1.0**
Aurélien Tchouaméni	**0.3373**	**0.1218**	**0.0306**	**1.0**
Dayotchanculle Upamecano	0.3148	**0.13**	0.0139	0.9091
Jules Koundé	0.3148	0.1064	0.0139	0.9091
Kylian Mbappé	0.2264	0.0586	0.0	0.7143
Ousmane Dembélé	0.2848	0.0492	0.0111	0.8333
Raphaël Varane	0.2915	**0.1408**	0.0028	0.8333
Theo Hernández	0.3373	0.1082	0.0306	1.0
Olivier Giroud	0.2195	0.042	0.0028	0.7143
Hugo Lloris	0.2848	0.0724	0.0111	0.8333

Final WWC 2023: Spain Women's vs. England Women's

Spain won its first-ever World Cup title in a match against the last UEFA Women's Euro champion in 2022, England. The winner could only score one goal, but it was more than enough to achieve the glory.

For this match, the Spanish coach Jorge Vilda replaced from the starting eleven Alexia Putellas by Salma Paralluelo, who was a key player against Sweden and The Netherlands in the quarters and semi-finals. By contrast, the English coach Sarina Wiegman kept the accustomed eleven with Alessia Russo and Lauran hemps leading the attacks.

There were few scoring opportunities for either team in the early minutes of the match, but it was not until the 29th minute that Olga Carmona, with her left foot, struck the winning goal in a play assisted by Mariona Caldentey. England fought for the title but didn't find the way to overcome the Spanish wall and their game didn't flow as they wish. Even the penalty stopped by the goalkeeper Earps was not enough for the English players to turn the game around.

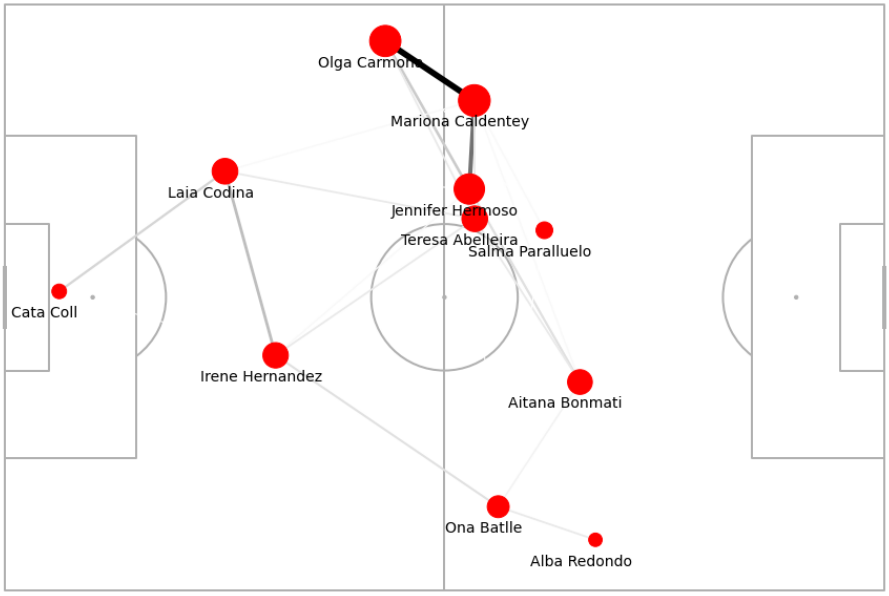

Figure 15 - Spain Women's Passing Network - WC 2023 Final

Figure 15 clearly shows the attacking strategy of the Spanish team, as well as their dominant position in the opponent's half before the first change. Most of the plays take place on the left-hand side, with the most relevant players Olga, Caldentey, and Jenni Hermoso. Additionally, Irene Hernández and Laia Codina starred some plays in the back positions. Aitana, Ona and Alba were more advance, but they barely could find opportunities to play the ball and score. Finally, the winning shot came from the nodes were most of the links were occurring.

Table 8 - Spain Women's Microscale Metrics - WC 2023 Final

Player Name	Eigenvector	Pagerank	Betweenness	Closeness
Aitana Bonmati	0.2987	0.0949	0.0113	0.8333
Alba Redondo	0.2324	0.0455	0.0	0.7143
Irene Hernandez	**0.3507**	0.0918	**0.0435**	**1.0**
Jennifer Hermoso	**0.3507**	**0.1226**	**0.0435**	**1.0**
Mariona Caldentey	**0.3268**	**0.141**	**0.0243**	**0.9091**
Olga Carmona	0.2948	**0.116**	0.0156	0.8333
Ona Batlle	0.2912	0.0819	0.023	0.8333
Salma Paralluelo	0.3259	0.0584	0.0224	0.9091
Teresa Abelleira	0.3268	0.1028	0.0243	0.9091
Cata Coll	0.2242	0.0447	0.0069	0.7143
Laia Codina	0.2629	0.1005	0.0074	0.7692

Table 8 reveals also the importance of the defender Irene Hernández in the match. A key player to stop the English team. Focusing now on the distribution and connectivity metrics, we can observe that the values are slightly more spread out than the values of the men's team, meaning that the strategy followed was not very well connected and distributed. Only a few players are connected to each other, as we can see from observing the two distinct sides in the network.

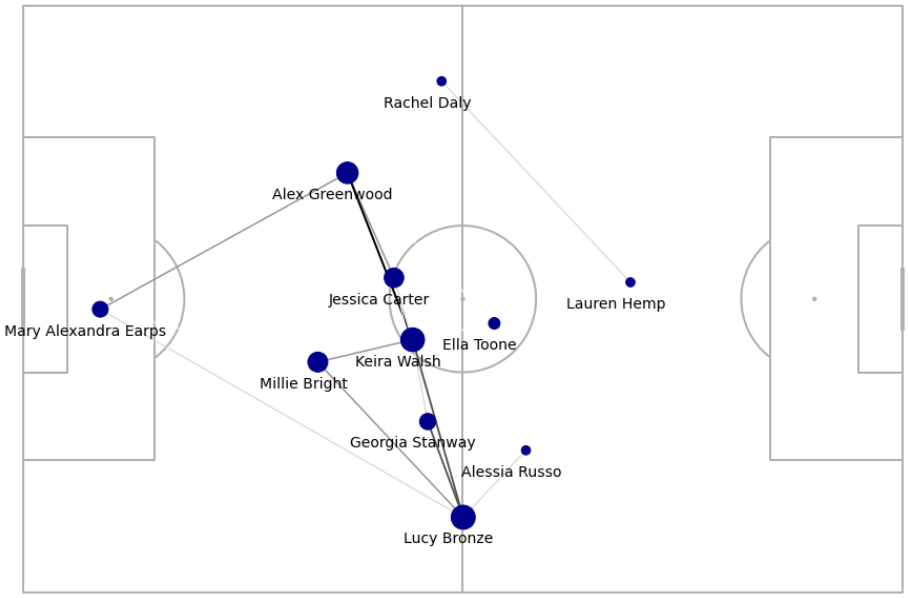

Figure 16 - England Women's Passing Network - WC 2023 Final

In the English case, both Figure 16 and Table 9 display the same information. There are three main players: Alex Greenwood, Keira Walsh, and Lucy Bronze. Whilst there are some players barely playing a role in the match (Toone, Hemp and Daly), the team seems to be connected and equally distributed in the middle and the right-hand side. That might explain why Bonmati couldn't find a way to reach the goal area on that side and most of the plays of the Spanish team had to be on the left side, where they found a break-through.

Table 9 - England Women's Microscale Metrics - WC 2023 Final

Player Name	Eigenvector	Pagerank	Betweenness	Closeness
Alessia Russo	0.2963	0.0683	0.0028	0.9091
Alex Greenwood	**0.3191**	**0.1084**	**0.0128**	**1.0**
Keira Walsh	**0.3191**	**0.1267**	**0.0128**	**1.0**
Lucy Bronze	**0.3191**	**0.1463**	**0.0128**	**1.0**
Ella Toone	0.2963	0.0687	0.0028	0.9091
Georgia Stanway	0.3191	0.0865	0.0128	1.0
Jessica Carter	0.2932	0.0923	0.0063	0.9091
Lauren Hemp	0.3191	0.0924	0.0128	1.0
Mary Alexandra Earps	0.236	0.0701	0.0	0.7692
Millie Bright	0.3191	0.0863	0.0128	1.0
Rachel Daly	0.268	0.054	0.0	0.8333

3.2 Statistical Procedure Results

The way teammates connect to each other during a match is key when it comes to define team's playing style. Thus, the present study explored the network differences among national teams at every stage over the course of the FIFA World Cups from 2018 to 2023. Concretely, the study addresses these two aspects:

a) Teams that achieved the highest stages in competition were expected to show high values of density and low values in clustering coefficient and diameter.

b) Main differences (if any) between male and female games.

Sections 3.2.1 through 3.2.3 present the results of the distinct statistical approaches. Only the significant results will be described, the left results can be found in the Annexes.

3.2.1 One – Way ANOVA Results

ANOVA tests, followed by Tukey's HSD post-hoc tests, were performed to determine if significant differences existed between the maximum competition stage achieved and the network characteristics of the teams. Accordingly, the null hypothesis states that the mean values of each network measures under analysis are equal across all the stages of the competition. Only when the p-value (*Pr (> F) column*) is less than the significance level of 0.05, we reject the null hypothesis and conclude that some phases of the competitions have different means.

In the case of men's World Cup championships, no statistically significant differences were found (at a significant level of $p < 0.05$) between the maximum stage reached in the tournament and any of the dependent variables based on the network metrics studied. By contrast, for women, results seemed more promising. Significant differences were found between the maximum phase reached and two dependent variables: total number of links and network density. Links presented statistical differences when only data for 2019 was considered, while density showed differences when the ANOVA model was run joining both seasons, assuming that teams from same country in different competitions can be considered different teams.

To pinpoint which phases are provoking the null hypothesis rejection, a pairwise comparison is necessary, therefore, a TukeyHSD post-hoc test was conducted in every case. Statistically significant differences in the means were only observed between the Round of 16 and the Group Stage phases. Similarly, there is insufficient evidence to suggest differences between any knock-out phase. Table 10, Table 11, Table 12 and Table 13 summarize these results.

Table 10 - ANOVA Table for Links - maxPhase WWC 2019

Source	D.F	SS	MS	F(obs)	Pr (>F)
maxPhase	3	213.1	71.05	3.747	0.0276
Residuals	20	379.3	18.96		
Total	23	592.4			

Table 11 - TukeyHSD test on Links – maxPhase WWC 2019

	diff	lwr	upr	p adj
Final - 3rd Place Final	0.2857	-11.9030	12.4744	0.9999
Regular Season-3rd Place Final	-6.2262	-15.8622	3.4098	0.2986
Round of 16 - 3rd Place Final	0.0780	-9.2313	9.3873	1.0000
Group Stage - Final	-6.5119	-16.1479	3.1241	0.2630
Round of 16 - Final	-0.2077	-9.5170	9.1015	0.9999
Round of 16 - Group Stage	6.3042	0.7408	11.8675	0.0228

Table 12 - ANOVA Table for Density - maxPhase WWC 2019 & 2023

Source	D.F	SS	MS	F(obs)	Pr (>F)
maxPhase	4	0.0487	0.01217	3.446	0.0144
Residuals	51	0.1802	0.00353		
Total	55	0.2289			

Table 13 - TukeyHSD test on Density – maxPhase WWC 2019 & 2023

	diff	lwr	upr	p adj
Final-3rd Place Final	0.0396	-0.0792	0.1585	0.8786
Group Stage-3rd Place Final	-0.0399	-0.1307	0.0509	0.7262
Quarter-finals-3rd Place Final	0.0351	-0.0838	0.1539	0.9186
Round of 16-3rd Place Final	0.0114	-0.0807	0.1034	0.9967
Group Stage-Final	-0.0796	-0.1703	0.0112	0.1117
Quarter-finals-Final	-0.0046	-0.1234	0.1143	1.0000
Round of 16-Final	-0.0283	-0.1203	0.0638	0.9070
Quarter-finals-Group Stage	0.0750	-0.0158	0.1658	0.1504
Round of 16-Group Stage	0.0513	0.0004	0.1022	0.0474
Round of 16-Quarter-finals	-0.0237	-0.1158	0.0684	0.9490

3.2.2 Generalized Linear Mixed Model for a Binary response variable Results

Given the results of the ANOVA tests for women's competitions – significant differences between the mean values from the group stage and the round of 16 phase – we aimed to further research this by modelling the probability of passing the group stage (our dependent variable in this case) based on network metrics (fixed effects) while controlling for team differences (random effects). For that purpose, we built five different models: from the most complex one (accounting for the five independent variables) to the simplest (with only two variables).

To describe the results, it will be presented the most comprehensive model and the one with the lowest AIC and BIC[10] among the remaining four models[11]. On top of that, these output considerations will be examined:

- **Random effects variance** → how much the predicted values vary due to this effect. The higher this value, the more the variability between the teams influencing the binary response variable: *passed_GS* (passing the group stage of the competition or not). Meaning the differences between the teams are important for the model.
- **Fixed effects predictors** and the corresponding p-value to understand the statistical significance of its impact.

For both genders, model 4 outperforms the most complex one. We will now analyze the results for each case in detail.

Comprehensive model (*m1*) formula:

$$logit\big(\mathbb{P}(passed_{GS} = 1)\big) = \beta_0 + \beta_1 \cdot Links + \beta_2 \cdot Diameter + \beta_3 \cdot Density + \beta_4 \cdot Assortativity + \beta_5 \cdot GlobalClusCoeff + u_{Team}$$

Secondary model (*m4*) formula:

$$logit\big(\mathbb{P}(passed_{GS} = 1)\big) = \beta_0 + \beta_1 \cdot Density + \beta_4 \cdot Assortativity + u_{Team}$$

Starting with the men's World Cups of 2018 and 2022 together, results are displayed below.

[10] Acronyms that stands for Akaike Information Criterion and Bayesian Information Criterion respectively
[11] The complete RMarkdown script with the GLLM results can be found in the annexes.

Table 14 exhibits lower AIC and BIC values for model 4, denoting a better fit, which means that fewer predictor variables adjust better the model. The random effect variance is quite large for both, slightly greater for m4. As we advanced before, great variability in the random effect implies that differences between the teams are important for the model.

Table 14 - Model fit measures and random effect variance MWC 2018 & 2022

	AIC	BIC	r.e variance
m1	207.6	232.4	32.21
m4	202.1	216.2	32.42

According to the results from Table 15 and Table 16, none of the predictors values are statistically significant for a level of $p < 0.05$, the closest in both models is the assortativity measure, but at a significance level of $p < 0.1$, meaning that this metric could be associated with the probability of passing the group stage. Nevertheless, these results won't be considered for the discussion in chapter 4.

Table 15 - Fixed effect predictors for model one (m1) - MWC 2018 & 2022

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.378	1.272	-0.297	0.766
Links	-0.025	0.293	-0.084	0.933
Diameter	-0.020	0.326	-0.061	0.952
Density	0.184	0.305	0.603	0.547
Assortativity	0.521	0.313	1.668	0.095
GlobalClusCoeff	0.205	0.319	0.644	0.519

Table 16 - Fixed effect predictors for model four (m4) - MWC 2018 & 2022

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.344	1.230	-0.280	0.779
Density	0.234	0.263	0.888	0.374
Assortativity	0.540	0.293	1.845	0.065

Continuing with women's World Cups 2019 and 2023:

Table 17 - Model fit measures and random effect variance WWC 2019 & 2023

	AIC	BIC	r.e variance
m1	162	186.2	41.83
m4	160.2	174	35.94

Table 17 does not show very different values in terms of model fit measures. Regarding the random effect variance, in model 1 the team effect seems to have more influence in the dependent variable.

Either model 1 or model 4 shows that density is a significant predictor of passing the group stage, being slightly larger the significance for model 4. Resembling results from men's data, the total number of links is statistically significant for a level of $p < 0.1$. These findings are generally consistent with the ANOVA results.

Table 18 - Fixed effect predictors for model one (m1) - WWC 2019 & 2023

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.053	1.522	0.692	0.489
Links	-0.803	0.432	-1.858	0.063
Diameter	-0.028	0.419	-0.066	0.947
Density	1.333	0.492	2.707	0.007
Assortativity	0.293	0.339	0.866	0.387
GlobalClusCoeff	-0.171	0.354	-0.483	0.629

Table 19 - Fixed effect predictors for model four (m4) - WWC 2019 & 2023

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.971	1.377	0.706	0.480
Density	1.103	0.400	2.757	0.006
Assortativity	0.199	0.313	0.636	0.525

3.2.3 Clustering by teams (using Mahalanobis distance) Results

The reason of choosing the Mahalanobis distance is displayed in Table 20 through Table 23. These tables are the correlation matrices for the four men's and women's World Cups tournaments under analysis. In them we can observe, highlighted in yellow, the highest correlation values.

Table 20 - Correlation matrix for MWC 2018

	Links	Density	Diameter	Assortativity	ClusCoeff
Links	1.000				
Density	0.800	1.000			
Diameter	-0.659	-0.689	1.000		
Assortativity	-0.215	-0.257	0.454	1.000	
ClusCoeff	-0.537	-0.541	0.698	0.547	1.000

Table 21 - Correlation matrix for MWC 2022

	Links	Density	Diameter	Assortativity	ClusCoeff
Links	1.000				
Density	0.522	1.000			
Diameter	-0.636	-0.584	1.000		
Assortativity	-0.151	-0.124	0.242	1.000	
ClusCoeff	-0.174	-0.120	-0.035	-0.042	1.000

Table 22 - Correlation matrix for WWC 2019

	Links	Density	Diameter	Assortativity	ClusCoeff
Links	1.000				
Density	0.846	1.000			
Diameter	-0.387	-0.335	1.000		
Assortativity	-0.024	-0.030	0.463	1.000	
ClusCoeff	-0.392	-0.350	0.017	-0.061	1.000

Table 23 - Correlation matrix for WWC 2023

	Links	Density	Diameter	Assortativity	ClusCoeff
Links	1.000				
Density	0.516	1.000			
Diameter	-0.545	-0.523	1.000		
Assortativity	-0.218	-0.046	0.432	1.000	
ClusCoeff	-0.462	-0.156	0.250	0.212	1.000

By clustering teams characterized on their network metrics, we expected to see teams grouped more accordingly to their ranking position in the championship. However, the results were far from this expectation. The cluster method applied - hierarchical clustering - does not reveal a clear pattern showing that network metrics influence the stage of the competition achieved in any of the seasons.

The complete scripts with the results from each model can be found in the annexes. Below only the cluster classification from the hierarchical clustering considering the Mahalanobis distance.

Table 24 - Hierarchical clustering results MWC 2018

Team	maxPhase	Cluster
Argentina	Round of 16	1
Australia	Group Stage	1
Brazil	Quarter-finals	1
Colombia	Round of 16	1
Croatia	Final	1
Denmark	Round of 16	1
Egypt	Group Stage	1
England	3rd Place Final	1
France	Final	1
Germany	Group Stage	1
Iceland	Group Stage	1
Mexico	Round of 16	1
Morocco	Group Stage	1
Panama	Group Stage	1
Peru	Group Stage	1
Poland	Group Stage	1
Portugal	Round of 16	1
Russia	Quarter-finals	1
Saudi Arabia	Group Stage	1
South Korea	Group Stage	1
Spain	Round of 16	1
Sweden	Quarter-finals	1
Tunisia	Group Stage	1
Uruguay	Quarter-finals	1
Belgium	3rd Place Final	2
Costa Rica	Group Stage	2
Iran	Group Stage	2
Japan	Round of 16	2
Nigeria	Group Stage	2
Senegal	Group Stage	2
Serbia	Group Stage	2
Switzerland	Round of 16	2

Table 25 - Hierarchical clustering results MWC 2022

Team	maxPhase	Cluster
Argentina	Final	1
Australia	Round of 16	1
Belgium	Group Stage	1
Brazil	Quarter-finals	1
Cameroon	Group Stage	1
Canada	Group Stage	1
Costa Rica	Group Stage	1
Denmark	Group Stage	1
England	Quarter-finals	1
Germany	Group Stage	1
Ghana	Group Stage	1
Iran	Group Stage	1
Japan	Round of 16	1
Morocco	3rd Place Final	1
Portugal	Quarter-finals	1
Saudi Arabia	Group Stage	1
Serbia	Group Stage	1
Spain	Round of 16	1
Switzerland	Round of 16	1
United States	Round of 16	1
Uruguay	Group Stage	1
Wales	Group Stage	1
Croatia	3rd Place Final	2
Ecuador	Group Stage	2
France	Final	2
Mexico	Group Stage	2
Netherlands	Quarter-finals	2
Poland	Round of 16	2
Qatar	Group Stage	2
Senegal	Round of 16	2
South Korea	Round of 16	2
Tunisia	Group Stage	2

Table 26 - Hierarchical clustering results WWC 2019

Team	maxPhase	Cluster
Argentina Women's	Group Stage	1
Brazil Women's	Round of 16	1
Cameroon Women's	Round of 16	1
China PR Women's	Round of 16	1
England Women's	3rd Place Final	1
France Women's	Round of 16	1
Italy Women's	Round of 16	1
Korea Republic Women's	Group Stage	1
South Africa Women's	Group Stage	1
Australia Women's	Round of 16	2
Canada Women's	Round of 16	2
Chile Women's	Group Stage	2
Germany Women's	Round of 16	2
Jamaica Women's	Group Stage	2
Japan Women's	Round of 16	2
Netherlands Women's	Final	2
New Zealand Women's	Group Stage	2
Nigeria Women's	Round of 16	2
Norway Women's	Round of 16	2
Scotland Women's	Group Stage	2
Spain Women's	Round of 16	2
Sweden Women's	3rd Place Final	2
Thailand Women's	Group Stage	2
United States Women's	Final	2

Table 27 – Hierarchical clustering WWC 2023

Team	maxPhase	Cluster
Argentina Women's	Group Stage	1
Canada Women's	Group Stage	1
Germany Women's	Group Stage	1
Italy Women's	Group Stage	1
Japan Women's	Quarter-finals	1
Netherlands Women's	Quarter-finals	1
New Zealand Women's	Group Stage	1
Nigeria Women's	Round of 16	1
Panama Women's	Group Stage	1
Republic of Ireland Women	Group Stage	1
Spain Women's	Final	1
Zambia Women's	Group Stage	1
Australia Women's	3rd Place Final	2
Brazil Women's	Group Stage	2
China PR Women's	Group Stage	2
Colombia Women's	Quarter-finals	2
Costa Rica Women's	Group Stage	2
Denmark Women's	Round of 16	2
England Women's	Final	2
France Women's	Quarter-finals	2
Haiti Women's	Group Stage	2
Jamaica Women's	Round of 16	2
Korea Republic Women's	Group Stage	2
Morocco Women's	Round of 16	2
Norway Women's	Round of 16	2
Philippines Women's	Group Stage	2
Portugal Women's	Group Stage	2
South Africa Women's	Round of 16	2
Sweden Women's	3rd Place Final	2
Switzerland Women's	Round of 16	2
United States Women's	Round of 16	2
Vietnam Women's	Group Stage	2

4. Discussion

Human survival and evolutionary success were shaped thanks to cooperation among the individuals. The desire to understand the extent to which people's behaviors and relationships influence others and outcomes, usually translated into performance, innovation, social cohesion, information diffusion, etc. (Nunes & Abreu, 2020), rose the question posed in the title of this project: *Can effective cooperation between teammates enhance overall team performance?* Johan Norberg introduced his book *Open. The Story of Human Progress* claiming the following:

> "*Homo sapiens* is a cooperative species […].
>
> Man is a trader by nature. We constantly exchange know-how, favors and goods with others, so that we can accomplish more than we would if we were limited to our own talents and experiences. […]. The sharing of knowledge and goods made it possible for humans to survive and prosper in inhospitable climates all over the planet. This gave rise to science, which is built on the exchange, criticism, comparison and accumulation of knowledge, and to technology, which is the application of science to solve practical problems.
>
> We observe the benefits cooperation and mobility have given us when it's suddenly shut down. The World Bank has calculated that the greatest economic damage from epidemics like swine flu, SARS or the new coronavirus do not come from mortality, morbidity, treatment and associated loss of production, but from increased fear of associating with others. Up to 90 per cent of the damage comes from aversion behavior, which shut down places of production, transportation, harbors and airports".

Even in today's business landscape, there exists also the desire to understand the extent to which the dynamic interactions of project people across all the phases of a project lifecycle, influence a project's outcome (Nunes & Abreu, 2020). According to literature, the presence of networks of collaboration has a positive impact into business organizations when efficiently distributed across functions, geographies and technical expertise domains, helping them gain competitive advantages, foster innovation and improve performance. The network factor is often a key predictor of high organizational performance. In such cases, organizations are typically characterized by broader and diverse problem-solving networks fueled with positive energy (Workday, 2018).

After reading these lines, one might directly jump to the conclusion that the better the cooperation between team members, the better the performance. However, are team sport dynamics the same as the complex dynamics of common social structures? Can we measure cooperation's impact in a meaningful way when it comes to soccer data?

The truth is that, despite the popularity of graph theory application in studies from distinct disciplines, the network analysis science applied to sports is still very recent and there is no strong evidence of a concrete network metric being a definitive predictor of team success. Even less about women's teams.

Since 2004, there have been few studies analyzing different World Cups along the years: Scoulding, James, & Taylor (2004) analyzed the FIFA World Cup 2002 and Clemente (2012) the FIFA World Cup taking place in 2010. Both studies concluded that the passing strategy did not discriminate successful teams from the unsuccessful ones. However, Clemente, Martins, Kalamaras, Wong and Mendes, in 2017 considered FIFA World Cup 2014 data and the insights were different. They did find statistical evidence on the assumption that successful teams have the highest levels of network density, total links, and clustering coefficient. Suggesting that "the ability to increase the connection between the teammates may result in excellent overall team performance in a competition".

The present study analyzed data available from the last four competitions since 2018. That is FIFA World Cup 2018 and 2022 for men and FIFA World Cup 2019 and 2023 for women. In our context, none of the men's tournaments shed some light on the matter and results were in line with Scoulding et al. (2004) and Clemente (2012): passing strategy did not discriminate successful teams from the unsuccessful ones. Nevertheless, and amazingly enough, the results are different in the case of women's teams and more in line with the study by Clemente et al. (2017) study: the better the overall team connectivity, the better team performance in a competition. More precisely, ball's possession and team connectivity seem decisive for passing the group phase. This prompts a new question: *could psychological or emotional factors be influencing performance more strongly in the female game?*

Shin and Park (2023) concluded that women benefit more from supportive networks than men but that they are also more vulnerable to a lack of supportive relationships. This suffering might be due to high expectations and devotion toward social relationships (Davis and Greenstein, 2009). *Could this dynamic extend to soccer team interactions?* The social burden of being the caregiver – always looking after others - together with the beloved guilt may lead to a less individualistic playing style, one that emphasizes unity over ego. In line with Shin and Park's conclusions, since women are more susceptible to the absence of supportive links, facing a team with a strong and well-connected network may cause their opponents to fall out of the game due to the psychological pressure this cohesion creates. This might not be the case for men's games, in which the playing style is more individually competitive, allowing each player's ability to outperform others to mitigate psychological factors.

Are these findings strong enough to explain differences in revenue, investment or coverage between men's and women's soccer? Why does the perception persist that men's soccer is more thrilling than women's?

Besides the team cohesion discussed above, there could be underneath factors not captured by the network that make male soccer more appealing than female one. It is broadly assumed that absolute differences in quality of athletic performance are the driving force for these differences (Gomez-Gonzalez et al., 2023). On top of that, some fans could argue that they find the physical intensity and speed of play men's football matches more exciting, along with the differing tactical and technical approaches. Nevertheless, the experiment carried out by Gomez-Gonzalez et al. (2023) tested these beliefs. More than half a thousand participants watched short clips of elite female and male soccer players. There were a control and a treatment group. In the former the videos were unmodified, while in the latter the videos had the gender blurred. Surprisingly, the regression analysis of the study reveals that in the absence of gender information, the perceived quality of women's and men's soccer did not differ. Same conclusions when they explored the willingness to pay for the matches.

Hopefully, these insights contribute to understand the importance of deconstructing centuries of gender role stereotypes, not only in sports but in all aspects of life. While biological differences undoubtedly exist, the most significant one arises from how our brains have been shaped throughout history in distinct social contexts.

5. Limitations

Despite the valuable insights assessed in the discussion of the present study, several limitations should be acknowledged to understand potential areas for improvement. These shortcomings can be categorized into two groups: data and methodology.

Data Limitations

Some of the weaknesses related to the data were already mentioned during the study, such as the absence of quarter-final matches from the 2019 Women's World Cup. This fact may lead to results being less reliable and not fully comparable with other seasons.

Conversely, for a real understanding of the relationship between the network metrics and the team performance, it would have been beneficial to access more historical World Cups data or to evaluate regular league performance. In the former case, more data from the knockout phases is available, which allows more consistency in metrics. Similarly for the regular league, in which data tend to be more stable across the season, as teams have more time to develop a consistent playing style, which might be reflected in the metrics. However, StatsBombs repository has insufficient data for women's league, which limits the analysis and comparison between genders.

Methodological Limitations

For simplicity, the graphs analyze were undirected. In undirected passing networks, all connections are assumed to be bidirectional, which may not fully reflect the reality of the game (where a pass generally has an origin and a destination). This decision may influence network metrics, for example assortativity that provides a better coefficient when there is a distinction between in-degree and out-degree. What is more, PageRank is more effective with directed graphs because it identifies important nodes based on the number of incoming links from other significant nodes.

Graphs and microscale metrics only considered the starting eleven. This significantly limit the insights derived from them, specially of how the strategies might change after certain player substitutions at halftime.

Although network graphs represent the weight in both nodes and edges, the metrics used for the statistical procedures do not incorporate these weights, only the global clustering coefficient. The reason of this is based on the idea that weights are relevant for metrics that measures local interactions, like the local clustering coefficient – component of the global clustering coefficient studied – where weights ponder certain passes of the subgraphs. Likewise, weights are less crucial for measurements that study global characteristics, such as total number of links, density, diameter or assortativity, in which weights do not change substantially the general analysis of the network. This fact might impact slightly the results, although the main conclusions are still valid.

6. Conclusions

While results from men's Worl Cups do not show evidence of significant differences between the maximum stage reached and the network metrics, results for women's Worl Cup tournaments are more revealing. Ball possession and team cohesion influence decisively passing or not the group stage of the championship. These findings could be explained by psychological factors derived from the socio-cultural context that women face. Yet, there is still insufficient evidence to explains why men's soccer is perceived as more thrilling than women's.

Bibliography

Antunes, M. (s. f.). *Análise da variância e regressão* [Diapositivas]. [Apuntes de clase]. DEIO - Faculdade de Ciências da Universidade de Lisboa.

average_clustering — NetworkX 3.3 documentation. (s. f.). https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.average_clustering.html

Beheshtian-Ardakani, A., Salehi, M., & Sharma, R. (2023). CMPN: Modeling and analysis of soccer teams using Complex Multiplex Passing Network. *Chaos, Solitons & Fractals/Chaos, Solitons And Fractals*, *174*, 113778. https://doi.org/10.1016/j.chaos.2023.113778

Caicedo-Parada, S., Lago-Peñas, C., & Ortega-Toro, E. (2020). Passing Networks and Tactical Action in Football: A Systematic Review. *International Journal Of Environmental Research And Public Health/International Journal Of Environmental Research And Public Health*, *17*(18), 6649. https://doi.org/10.3390/ijerph17186649

Clemente, F. M. (2012). Study of Successful Teams on FIFA World Cup 2010 through Notational Analysis. *Pamukkale Journal Of Sport Sciences*, *3*(3), 90-103. https://dergipark.org.tr/en/download/article-file/191782

Clemente, F. M., Martins, F. M. L., Kalamaras, D., Del Wong, P., & Mendes, R. S. (2015). General network analysis of national soccer teams in FIFA World Cup 2014. *International Journal Of Performance Analysis In Sport*, *15*(1), 80-96. https://doi.org/10.1080/24748668.2015.11868778

clustering — NetworkX 3.3 documentation. (s. f.). https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.cluster.clustering.html

Contributor. (2024, 3 mayo). *The Most Popular Sports in 2024: A Comprehensive Guide*. BNO News. https://bnonews.com/index.php/2024/05/the-most-popular-sports-in-2024-a-comprehensive-guide/

Davis, S. N., & Greenstein, T. N. (2009). Gender Ideology: Components, Predictors, and Consequences. *Annual Review Of Sociology*, *35*(1), 87-105. https://doi.org/10.1146/annurev-soc-070308-115920

degree_assortativity_coefficient — NetworkX 3.3 documentation. (s. f.). https://networkx.org/documentation/stable/reference/algorithms/generated/networkx.algorithms.assortativity.degree_assortativity_coefficient.html#id1

Easley, D., & Kleinberg, J. (2010). *Networks, Crowds, and Markets: Reasoning about a Highly Connected World*. Cambridge University Press.

Everitt, B. S., Landau, S., Leese, M., & Stahl, D. (2011). *Cluster analysis*. Wiley.

Gomez-Gonzalez, C., Dietl, H., Berri, D., & Nesseler, C. (2023). Gender information and perceived quality: An experiment with professional soccer performance. *Sport Management Review, 27*(1), 45-66. https://doi.org/10.1080/14413523.2023.2233341

Herrero Candela, R. (2022). *Análisis de las estrategias utilizadas en el fútbol femenino mediante el modelado de pases* [Trabajo de fin de grado]. Universitat Politècnica de València.

Ievoli, R., Palazzo, L., & Ragozini, G. (2021). On the use of passing network indicators to predict football outcomes. *Knowledge-based Systems, 222*, 106997. https://doi.org/10.1016/j.knosys.2021.106997

Introduction to Generalized Linear Mixed Models. (s. f.). https://stats.oarc.ucla.edu/other/mult-pkg/introduction-to-generalized-linear-mixed-models/

Kacanski, S., & Lusher, D. (2017). The Application of Social Network Analysis to Accounting and Auditing. *International Journal Of Academic Research In Accounting Finance And Management Sciences, 7*(3). https://doi.org/10.6007/ijarafms/v7-i3/3286

Leela, J., Rahaman, K., & Comissiong, D. M. G. (2024). Analyzing the Efficiency of Passing Networks in Soccer. *Athens Journal Of Sports, 11*(1), 37-58. https://doi.org/10.30958/ajspo.11-1-3

Mclean, S., Salmon, P. M., Gorman, A. D., Stevens, N. J., & Solomon, C. (2018). A social network analysis of the goal scoring passing networks of the 2016 European Football Championships. *Human Movement Science, 57*, 400-408. https://doi.org/10.1016/j.humov.2017.10.001

Newman, M. E. J. (2002). Assortative Mixing in Networks. *Physical Review Letters, 89*(20). https://doi.org/10.1103/physrevlett.89.208701

Norberg, J. (2020). *Open: The story of human progress*. Atlantic Books.

Nunes, M., & Abreu, A. (2020). Applying Social Network Analysis to Identify Project Critical Success Factors. *Sustainability*, *12*(4), 1503. https://doi.org/10.3390/su12041503

Page Rank in Network Analysis - Andrea Perlato. (s. f.). https://www.andreaperlato.com/graphpost/page-rank-in-network-analysis/

Prabhakaran, S. (s. f.). *Mahalanobis Distance – Understanding the math with examples (python)*. Machinelearningplus. Recuperado 2 de septiembre de 2024, de https://www.machinelearningplus.com/statistics/mahalanobis-distance/

Sabot, K., Wickremasinghe, D., Blanchet, K., Avan, B., & Schellenberg, J. (2017). Use of social network analysis methods to study professional advice and performance among healthcare providers: a systematic review. *Systematic Reviews*, *6*(1). https://doi.org/10.1186/s13643-017-0597-1

Salehi, M., Sharma, R., Marzolla, M., Magnani, M., Siyari, P., & Montesi, D. (2015). Spreading Processes in Multilayer Networks. *IEEE Transactions On Network Science And Engineering*, *2*(2), 65-83. https://doi.org/10.1109/tnse.2015.2425961

Scoulding, A., James, N., & Taylor, J. (2004). Passing in the Soccer World Cup 2002. *International Journal Of Performance Analysis In Sport*, *4*(2), 36-41. https://doi.org/10.1080/24748668.2004.11868302

Shin, H., & Park, C. (2023). Gender differences in social networks and physical and mental health: are social relationships more health protective in women than in men? *Frontiers In Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1216032

Sigrist, F. (2024, 26 marzo). Generalized Linear Mixed Effects Models in R and Python with GPBoost. *Medium*. https://towardsdatascience.com/generalized-linear-mixed-effects-models-in-r-and-python-with-gpboost-89297622820c

Thakur, D. (2024, 7 marzo). *10 Most Profitable Sports Leagues in The World*. SportsUnfold. https://www.sportsunfold.com/10-most-profitable-sports-leagues-in-the-world/

Ward, M. D., Stovel, K., & Sacks, A. (2011). Network Analysis and Political Science. *Annual Review Of Political Science*, *14*(1), 245-264. https://doi.org/10.1146/annurev.polisci.12.040907.115949

Wikipedia contributors. (2024, 23 junio). *Mahalanobis distance*. Wikipedia. https://en.wikipedia.org/wiki/Mahalanobis_distance

Workday. (2018, 31 julio). *In good company: organizational networks* [Vídeo]. YouTube. https://www.youtube.com/watch?v=6faV0v0yVFU

Xiang, S., Nie, F., & Zhang, C. (2008). Learning a Mahalanobis distance metric for data clustering and classification. *Pattern Recognition*, *41*(12), 3600-3612. https://doi.org/10.1016/j.patcog.2008.05.018

Xing, E. P., Jordan, M. I., Russell, S. J., & Ng, A. Y. (2002). Distance Metric Learning with Application to Clustering with Side-Information. *Neural Information Processing Systems, 15*, 521-528. http://papers.nips.cc/paper/2164-distance-metric-learning-with-application-to-clustering-with-side-information.pdf

Annexes

Dataset

MWC 2018 2022
Macroscale metrics.xl

WWC 2019 2023
Macroscale metrics.xl

Network graphs and metric tables

Network graphs and metrics.zip

R Scripts

ANOVA R script - MWC 2018 & 2022

ANOVA R script - WWC 2019 & 2023

Cluster R script - MWC 2018 & 2022

Cluster R script - MWC 2019 & 2023

GLMM R script - MWC 2018 & 2022

GLMM R script - MWC 2019 & 2023