Network Analysis:
Network Structure and Team Performance: Euro 2020

Marco Benito Tomasone 1038815
Luca Genova 1038843
Master Degree in Computer Science

2022-2023

1 Abstract

Brief summary of the whole study (around 60-120 words), summarising the salient parts of the sections below.

2 Context

Human interaction are very important in our society. In a lot of situations humans tend to group themselves in subset to operate in a more efficient way. Also in sports, humans organized themselves in teams to achive common goal and demostrate their skills and their ability to work in a group. In this case of interaction, the ways in which the members of the team interact with each other are more important, because the way people inside the team interact influence directly the results the team will obtain. In this study we will analyze the network structure of football teams and how it affects the performance of the team. In this work we focused on the last UEFA Euro 2020 tournament.

2.1 Previous Work

This work is heavely based on the work done by Thomas U. Grund in which he analyzed two seasons of the english Premier League.

3 Problem and Motivation

In last year data science is a very growing field and this it's impacting a lot of different fields, sport included. In the football case, data are in most case an unknown technology

and a lot of teams (even at highest level) don't use this type of knowledge to understand the performance of the team itself, but event of the opponents. Given that the way player interact each other can result in a better or worse performance of the team, it's important to understand how the network structure of the team can affect the performance of the team.

4 Datasets

The data have been provided by StatsBomb, one the biggest provider for football data. The dataset is composed by every single event during the match: every pass, every shoot, every press and so on. The gathering process is based on a python library called StatsBombPy, which is a wrapper for the StatsBomb API. StatsBomb's Data are not for free, except for some free data, Euro 2020 is one of them. From StatsBombPy you get data in form of Pandas Dataframe, which is a very useful tool for data analysis. We used python and Pandas (a python library) to manage data and then we gathered data and store Data for all matches in .csv o .xlsx format. All other manipulation on data have been done using python and metrics have been computed using NetworkX (a python library for network analysis). The data we gathered contains all the matches of the tournament, so we have data for 51 matches. In a football match there are two team facing, so we have a total of 102 networks. We computed metrics for each network and then we analyzed the results.

5 Validity and Reliability

TODO: add validity and reliability

6 Hypothesis

Out work is based on two fundamental hypothesis:

1. Team performance is affected by interaction opportunities: increased interaction intensity leads to increased team performance.

2. Increased centralization of interaction in teams leads to decreased team performance.

So the first hypothesis is based on the idea that the more a team interacts with each other, the more the team will be able to perform better. The second hypothesis is based on the idea that the more a team is centralized on a single player, the less the team will be able to perform better, because if a team makes to much reliance on a single player, the team will be more vulnerable to the opponent.

7 Measure

Based on the hypothesis we want to prove we focused on two metrics in particular:

- Network Intensity
- Network Centralization

7.1 Network Intensity

Usually one of the most computed metrics of networks is density, which is traditionally calculated as the number of existing ties in a network divided by the number of potential ties. In our case, we can't use this metric, because in a football match we can easily expect a pass between each pair of players in a team. So we need to define a new metric, which is weighted on the number of passes. To compute this new metric we need the information of the number of passes received and made by each player. So we used:

- **Out-strenght:** the number of passes made by the player i

$$C_{OS}(i) = \sum_{j=1}^{N} w_{ij}$$

- **In-strenght:** the number of passes received by the player i

$$C_{IS}(i) = \sum_{j=1}^{N} w_{ji}$$

Where N is the number of nodes in the network, in our case is 11, because we consider only the starting XI of each team. We focused our attention on the starting XI instead of the best eight player for number of passes (as in Grund CITAAAAA) because we are analyzing a tournament. In a tournament in the final phase if two teams draw go to extra time so some player from the bench could impact the match in a more important way than others. To limit this problem we only focused on the starting XI. Another reason is that these tournament are very short (7 match at maximum) and the match are knockout game, so all the coach try to always play with the best players he has for that match. Another reason for which we focused only on the starting XI is that from 2020 the number of subs a coach can do is 5, while in the past it was 3 (that's because Grund used only the most eight players, because he ideally consider only the players who played the entire match).
So the network intensity for a team is defined as:

$$I = \sum_{i=1}^{N} \frac{C_{OS}(i) + C_{IS}(i)}{2}$$

This is just the number of total passes a team has made. Differently from Grund we can't normalize this metric for the time of possession of the team, because we don't have this information.

7.2 Network Centralization

The network centralization is computed by computing two different metrics: the weighted centralization and the strenght centralization.

7.2.1 Weighted centralization

The weighted centralization is one of the simplest way to examine the distribution of the weights in a network. It is defined as:

$$C_w = \frac{\sum_{i=1}^{N} \sum_{j=1}^{N} (w^* - w_{ij})}{(N^2 - N - 1)IT}$$

Where w^* is the biggest tie value in the network (so the biggest number of passes between two players), w_{ij} is the weight of the link between the node i and the node j, N is the number of nodes in the network (11) and IT is the total number of passes of the team for that match.
This metric is zero in the the most decentralized interaction pattern so when everybody interacts with everybody with the same intensity. In contrast, this metric is maximum (1) in the case the most interactions involve the same two individuals.

7.2.2 Strenght centralization

The streght centralization is used instead of the degree centralization, beacuse as said for the network density in (almost all) football matches we can expect a pass between each pair of players in a team. So we used a centralization for incoming and outcoming node strenght:

$$C_I = \frac{\sum_{i=1}^{N} (C_{IS}^* - C_{IS}(i))}{(N-1)IT}$$

$$C_O = \frac{\sum_{i=1}^{N} (C_{OS}^* - C_{OS}(i))}{(N-1)IT}$$

Where C_{IS}^* is the biggest incoming node strenght in the network (so the biggest number of passes received by a player), $C_{IS}(i)$ is the incoming node strenght of the node i, C_{OS}^* is the biggest outcoming node strenght in the network (so the biggest number of passes made by a player), $C_{OS}(i)$ is the outcoming node strenght of the node i, N is the number of

nodes in the network (11) and IT is the total number of passes of the team for that match.

8 Results

We start the result section presenting a table summarising the results of the metrics we obtained for each of the 102 observation we made.

	Mean	Std_dev	Min	Max	Obs
I	376.372549	132.701028	89	722	102
C_w	0.030206	0.011425	0.001864	0.065984	102
C_I	0.073909	0.022539	0.032302	0.134383	102
C_O	0.077249	0.022527	0.032	0.137046	102

Tabella 1: Statistical summary of the metrics

8.1 Conclusion (not needed for the project proposal)

Qualitative analysis of the quantitative findings of the study.

8.1.1 Critique (not needed for the project proposal)

Do you think your work solves the problem presented above? To which extent (completely, what parts)? Why? What could you have done differently to answer your research problems (e.g., gather data with additional information, build your model differently, apply alternative measures)?