A Network-Based Approach to Predicting Passing Distributions
CS224W Project Milestone

Angelica Perez (pereza77), Jade Huang (jayebird)

November 17, 2015

1 Introduction

The passing network of a team can reveal much about game strategy–who the key player or key players are, whether a team had a hard time against its opponent and focused mainly on defense or if it was completely dominating its opponent and was rocking the offense, and so on. Furthermore, the passing network changes based on the opposing team–team A will play differently when faced with inferior team B than when faced with superior team C. Given the history of passing networks of team A and another team D, what if we could predict the passing networks of both teams ahead of time? And based on our predicted passing networks, predict the outcome of the match? Such is our envisioned task for our CS224W Final Project.

2 Review of Relevant Prior Work

Cintia et al, Narizuka et al, and Pena and Touchette model player passing networks where nodes are players and edges represent ball movements weighted by level of interaction between players or number of successful passes. While Pena and Touchette explore how network centrality measures relate to the success of a team, Cintia et al. attempt to predict match outcomes. Narizuka et al. propose a Markov-chain model to describe the degree distribution of "position-dependent" soccer ball-passing networks. For our goal of predicting a passing distribution, we can take into account network centrality measures such as PageRank and betweenness of each individual team member to predict individual links between players based on past performance. Once we predict passing distributions for two teams, then we can attempt to predict the match outcome based on our predicted passing distributions, using the mean and variance of the degrees of the nodes of the opposing teams.

Due to limitation of data from FIFA, Pena and Touchette computed passing networks by dividing the number of passes by the total number of plays played by each team. Thus, we lacked a per-game analysis which could be indicative since a team most likely does not play the same way with all opponents. Just as the authors noted how Spain's "total football" or "tiki-taka" style was evident in their centrality measures, passing strategies are particular to teams. Similarily, the model of Cintia et al. fails to acknowledge how team dynamics will perform in relation to the dynamics of an opponent. It would be interesting to analyze how different teams behave when facing an opponent, or how a single team changes its strategy depending on the opponent. Luckily, the data we have from the UEFA Champions League contains per-game analysis for players and teams including passing distributions for each team for each game, thus the goal is that we will be able to distinguish how a team plays against say, a highly-ranked team vs. a lesser-ranked team.

While it is a trade-off to be able to represent a mobile, non-static game with a static network, which the first two papers mentioned did, a weakness of Narizuka et. al's approach is the randomized starting position upon which the fitness, or distance to each node, is based. In order to present clear analysis of the passing distributions of a team and how they change upon facing other teams, we will use the static tactical lineups of each team from the UEFA Champions League to represent each node's position in the passing distribution as represented in network form.

However, what Narizuka et al. are able to do is theoretically approximate the degree distributions of these passing networks using a Markov chain model, which may assist our efforts when predicting the weight of degrees of our players in our predicted passing networks.

3 Description of Data Collection Process

We were blessed with the abundance of data provided by the UEFA Champions League, including tactical lineups, team passing distributions, and individual player statistics for 4 matchdays in the 2015-16 season, where there are 8 games per matchday. This data, unfortunately, was not pre-processed, and thus, we had to process and parse the data ourselves.

For the team passing distributions and individual player statistics, we first converted the pdf documents into xlsx and csv files using PDF Tables, which was able to mostly preserve the table format that pervades the documents. We then used Python scripts to parse the documents into formats that we could feed into our program as well as use to visualize and analyze the passing distributions as networks.

For the individual player statistics, we mapped features such as goals scored to numbers as well as players to their player number, resulting in the format of

```
player_number, feature_number: feature_value
```

For the passing distributions, we generated a csv file for nodes, i.e. players, and a csv file for weighted edges, i.e. passes between players. For nodes, we mapped player number to player name. For edges, the file was of the following format:

```
player_1 player_2 weight
```

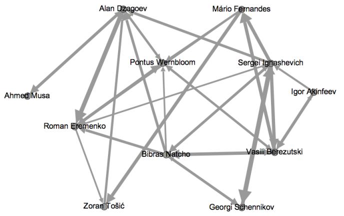
We utilized the tactical lineups as models for how to position nodes in our passing networks.

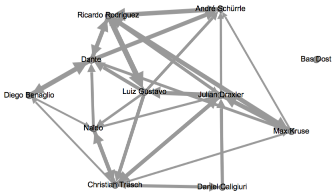
4 Initial Findings and Summary Statistics

During our analysis, we classified passing distributions as one of the following:

1. heavy defense
2. heavy midfield
3. heavy offense
4. balanced
5. light defense
6. light midfield
7. light offense

The most prevalent PD classifications in our data are heavy midfield, heavy defense, balanced and light offense. Higher ranked teams tend to have heavy midfield, heavy defense, or balanced, as they usually prefer to rebuild from the defense or midfield and and patiently develop offensive opportunities. A light offense PD occurs most often for a low-ranked team that faces a much higher-ranked team, since offensive opportunities for such a team are rare. Such a phenomenon was seen with a game between PFC CSKA Moskva (left) and VfL Wolfsburg (right), where VfL Wolfsburg won 1-0. While the distribution was fairly balanced in Moskva's passing network, there was heavy passing from defender to forward in Wolfsburg's network, showing a command of the midfield and dominance of the midfield that Moskva lacked.





Heavy offense, light defense, and light midfield are extremely rare. A heavy offense implies that one team is so much better than their opponent that they are always playing in their offensive third of the field, which is hardly ever seen in professional soccer. Midfielders always see a lot of action since they are the nodes that hold the network together, so a light midfield is also unheard of. This is definitely seen in Manchester United's game against PFC CSKA Moskva, whose game is mostly based on their central midfielders and playing it out to their wingers. An opposing team may have a weak set of offensive players, but defenders are also used to develop plays from the back field and often make runs up the field or even along the sides of the field to the offensive third, which is what PSV Eindhoven did in the 2nd matchday against Moskva.

5 Important Algorithms

5.1 Predicting Passing Distribution

In order to produce the passing distributions of two teams in a single match, we approach the problem as a simulated soccer match. We will represent a match as a network of players, the nodes, and their passes, the edges. There are two sets of nodes, $\{x_1, x_2, \ldots, x_{11}\}$ and $\{y_1, y_2, \ldots, y_{11}\}$, which represent the two teams. Edges can only be created between nodes in the same set, as teammates pass to each other. At a given time step t_i, a single player x_j possesses the ball and we determine if x_j passes to one of his teammates x_k (an edge is created) or if he loses the ball to an opponent y_l. This is determined by computing scores or edge strength for every pair of nodes (x_j, x_k) for $k \in range(1, 11)$ and $k \neq j$ as well as a teleporting probability that represents the likelihood of player x_j losing the ball to an opposing player. If teleportation to an opponent node does not occur, the teammate pair scores are converted to a normalized probability distribution and an edge is chosen. This process continues until we have reached the sum of each team's average total passes in a given match.

The success of this approach relies on the score functions for teammate pairs and each player's ball loss likelihood. We are experimenting with the use of a large array of past match statistics to develop the best score functions.

5.2 Baseline Model

The score functions for teammate pairs in our baseline model are computed using position passing ratios, while the opponent teleportation probability uses team rankings. A position passing ratio is the proportion of total passes in a match to each player position (i.e. defenders, midfielders, strikers, and goalkeepers). Both functions are listed below:

$$score(x_j, x_k) = \frac{\text{average number of passes to } Position(x_k)}{\text{average total number of passes}}$$

$$teleportProb(x_j) = \frac{rank(Team(x))}{rank(Team(x)) + rank(Team(y))}$$

A major characteristic of the PD's that the baseline fails to capture is how a team modifies its style of play based on its opponent's PD. This can be improved by using passing distribution data for individual players. How often do they pass to certain players? Against which opponent lineups do they perform the best? These are statistics that we will incorporate into our final model. The teleport probability also needs to use match statistics that are more indicative of a player's capability and situation on the field, such as which opposing player is he most likely facing. We will include individual player statistics to compute a more accurate measurement.

5.3 Supervised Link Prediction

In our final supervised link prediction model, we will use match data that gives a more accurate prediction of where the ball will travel in a given state. The score/edge strength function for a pair of nodes and the teleporting probability are defined as:

$$score(x_j, x_k) = \alpha assortativityScore + \beta maximumLikelihood$$

$$teleportProb(x_j) = \gamma clusteringCoefficient + \delta betweenness + \rho capacity$$

Each sub-score is described in the table below.

Sub-score	**Application**	**Data Used**
assortativityScore	tendency of players in the same position to pass to each other given the opponent's PD	team passing distributions
maximumLikelihood	tendency of a specific player to pass to another player	individual player passing distribution
clusteringCoefficient	do the players belong to a cluster	individual player passing distributions
betweenness	how vital is a player to the network	team passing distributions
capacity	how often a player fails to make a pass	player pass completion rates and team ranks

5.4 Evaluation

In evaluating our predicted passing networks, we can use past games in the 2014-15 season as well as existing games in the 2015-16 season to compare the accuracy of our predictions. We can aggregate a certain amount of matchdays to use as training data, while holding out matches with the two teams involved in our prediction, and then evaluate accuracy on our held-out set.

6 General Difficulties

So far, the greatest hurdle has been indeed, as our mentor predicted, processing the large amount of un-processed, raw data into readable and workable formats as well as losing a team member. Surely, we will run into more difficulties as we explore different methods and features in predicting passing networks.

References

[1] Paolo Cintia, Salvatore Rinzivillo, and Luca Pappalardo *A network-based approach to evaluate the performance of football teams.* Machine Learning and Data Mining for Sports Analytics workshop (MLSA'15). ECML/PKDD conference 2015.

[2] Takuma Narizuka, Ken Yamamoto, Yoshihiro Yamazaki. *Degree distribution of position-dependent ball-passing networks in football games.* Journal of the Physical Society of Japan. Vol.84, No.8. Article ID: 084003 (2015).

[3] Javier López Peña and Hugo Touchette. *A network theory analysis of football strategies.* C. Clanet (ed.), Sports Physics: Proc. 2012 Euromech Physics of Sports Conference, p. 517-528, Éditions de l'École Polytechnique, Palaiseau, 2013. (ISBN 978-2-7302-1615-9).