A Network-Assisted Approach to Predicting Passing Distributions

Angelica Perez
Stanford University
pereza77@stanford.edu

Jade Huang
Stanford University
jayebird@stanford.edu

Abstract

We introduce an approach of using a linear regression model plus stochastic gradient descent to predict the passing distribution of a soccer team based on data from the UEFA Champions League. Features are derived from player-specific statistics, team-specific history, and network measures while weights are learned to predict the number of passes between each pair of players on a team. Our linear regression model with features and SGD achieved a 25.27% improvement in the average loss from our baseline model of only average past passing networks.

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. 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 B, predicting the passing networks of both teams ahead of time has a number of implications. We can uncover the different patterns of how a team plays another team of a certain difficulty. We can predict the layout of future games, winning more bets, and perhaps giving teams a leg up on their competitors by giving them clues on how their competitor may play based on their playing history.

2 Problem

Given a match between two teams A and B, each of their starting lineups, and each team's passing history comprised of t past passing distributions $PD^{(A)} = \{pd_1^{(A)}, pd_2^{(A)}, , pd_t^{(A)}\}$ and $PD^{(B)} = \{pd_1^{(B)}, pd_2^{(B)}, , pd_t^{(B)}\}$ along with additional team and player statistical data, our model seeks to predict new passing networks $pd_{t+1}^{(A)}$ and $pd_{t+1}^{(B)}$ for each team. We use a linear regression model to learn feature weights using stochastic gradient descent that are then used to predict the edge weight, or number of passes, between each pair of nodes, or teammates, in match $t + 1$.

The features are comprised of a variety of UEFA data and attempt to encapsulate the necessary team and player characteristics to predict edge weights, such as team rankings, average passes between a pair of players, and average betweenness of a player. We carefully choose features that accurately describe a pair of player's historical relationship on the field, the typical involvement of an individual player amongst his team, and, equally as important, a team's strategy and passing history, including how a team's style of play (i.e. heavy midfield or light offense) may change against a specific opposing team.

3 Related Work

(Pena et al., 2012) uncover how players perform on a team by analyzing centrality measures such as closeness, betweenness, PageRank, and clustering. The authors found that through betweenness, one can uncover which players are more involved. If the network is very complete with most edges be-

tween most or all players, one can say the team is very well-connected. (Grund , 2012) expands upon (Pena et al., 2012) with two hypotheses that he confirms in his paper: 1. increased interaction intensity leads to increased team performance and 2. increased centralization of interaction in teams leads to decreased team performance. Using the mean degree (volume of play), variance degree (diversity of play), and harmonic mean of the mean and variance degree of the player passing network, (Cintia et al., 2015) achieved 53% accuracy in predicting match outcomes.

Building upon (Pena et al., 2012), centrality measures, especially betweenness centrality, can be used as features in predicting edges between players. A player with a higher centrality measure perhaps will receive more passes on average. Features used by (Cintia et al., 2015) and (Grund , 2012) such as mean degree, variance degree, interaction intensity, centralization of interaction are indicative of team-specific patterns and can be used in our model when predicting edges.

Due to limitation of data from FIFA, (Pena et al., 2012) computed passing networks by dividing the number of passes by the total number of plays played by each team, lacking a per-game analysis which could be indicative since a team most likely does not play the same way with all opponents. Similarly, (Cintia et al., 2015) and (Grund , 2012) fail to acknowledge how team dynamics will perform in relation to the dynamics of an opponent. (Grund , 2012) defines the most centralized network as the one where most interactions involve the same two individuals. However, such a definition can be further expanded to include cases such as where all members of a team pass to one central player equally.

4 Model

4.1 Baseline

For a baseline model, the average of past passing networks during the group stage is used to predict the passing networks during the round of 16 stage. Thus, during training, for every player to player combination in each team for which there exists a pass, the total number of passes over six games is calculated during the group stage and averaged over the number of games.

The average values calculated during training are used to predict the number of passes for each team in each game in the round of 16 of the 2014-15 season. For example, if in team X, player 1 passed to player 9 an average of five times during the group stage, this baseline model would predict that player 1 would pass to player 9 five times when playing some team in the round of 16.

This is a naive baseline that does not take into account the specific opponent during each of the past games during the group stage, but merely generalizes, ignoring any idiosyncrasies which may arise when facing a higher ranked or lower ranked opponent.

5 Algorithm

To build upon the baseline, a linear regression model was constructed to include features that are indicative of the amount of passes between two players and which would learn weights which would correspond how descriptive a feature is in relation to the amount of passes between two players.

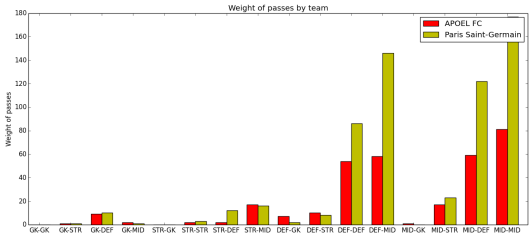
5.1 Features

With guidance from Tim Althoff, we postulated the following are key ingredients leading to how often two certain players pass:

- **Average number of past passes between two players**, a continuous value that is precomputed before the process of training and learning weights and averaged over the group stage. As with the baseline, the average of all passes between two players for each player for each team over all ”past” games, where the notion of past is experimented with later on, is taken.

- **Whether two players are of the same position**, firing 1 if the condition is true and 0 otherwise. The motivation for this feature is that based on analyzing the data, certain positions pass to each other very often, such as from defender to midfielder, while others never pass to each other, such as goalkeeper to goalkeeper. Please see 1 for a bar graph showing the number of passes made between each position in

Figure 1: Number of passes made between each position in a game between APOEL FC and Paris Saint-Germain.



a game between APOEL FC and Paris Saint-Germain. Similarly, we also have a feature for whether two players are of different positions.

- **Whether team A's rank is higher than team B's rank**, firing 1 if true and 0 otherwise. The motivation is that perhaps players may pass more or less depending on whether the opponent team is ranked higher or lower.

- **Whether team A has won against a "similar" team as team B in the past**, firing 1 if true and 0 otherwise. In every subsequent stage, teams play teams that they haven't played in previous stages–thus there is no previously existing data stored for team combinations at testing time. Thus, it may be helpful to find a team C that is similar to a team B in team A's play history and use the outcome of that match in determining if team A has a good chance of winning against team B.

- **Whether team A's average pass completion rate is higher than team B's pass completion rate**, firing 1 if true and 0 otherwise. The average is taken of pass completion rates across past games ignoring specific opponents. Similarly we also have features for whether team A's average passes attempted and average passes failed are higher than team B's pass completion rate. In analyzing the data, a trend was observed where 63.5% of the time the team with a higher pass completion rate and was also the team that lost the match. Also, 62.5% of the time the team with a higher number of passes attempted was also the team that lost the match.

- **Average number of passes between two positions on a team**, a continuous value. The motivation is that perhaps on a team, two given positions generally pass to each other some similar amount across all games.

- **The average betweenness centrality of both players** respectively averaged over games during the group stage. As mentioned in related work, betweenness centrality can indicate the importance of a player, of how often the player is a cog in between passes.

- **Average percentage of passes completed for two specific players** averaged over percentage of passes completed during the group stage, a continuous value. The intuition is perhaps it is more likely for two players to pass if their percentage of passes completed is higher.

5.2 Evaluation

We utilized the squared loss to capture the accuracy of a prediction of number of passes between two players in comparison to the actual number of passes between two players. The score $\phi(x) \cdot \mathbf{w}$ is considered the prediction while y represents the actual number of passes.

$$\begin{aligned} Loss_{squared} &= (\texttt{predicted} - \texttt{actual})^2 \\ &= (\phi(x) \cdot \mathbf{w} - y)^2 \end{aligned}$$

To evaluate our model as a whole, we took the average of the loss over all passes between players. In the below equation, T represents the total number of passes between players over all teams and games during the round of 16 stage.

$$Loss_{model} = \frac{1}{T} \sum_{i=0}^{i=T} (\phi(x) \cdot \mathbf{w} - y)^2 \qquad (1)$$

5.3 Learning Weights

Our linear regression model learns weights using stochastic gradient descent. The objective is to minimize our squared loss, thus with every new training example, i.e. number of passes between two players, we update our weights with the following equation:

$$w \leftarrow w - \eta \nabla_w Loss$$
$$w \leftarrow w - \eta 2(\phi(x) \cdot \mathbf{w} - y)\phi(x)$$

where η is the step size controlling the rate of descent.

5.4 Data

Data used includes team passing distributions, tactical lineups, and squad lists provided by the UEFA Champions League press kits for the Group, Round of 16, Quarter-finals, and Semifinal stages of the 2014-15 season. Team rankings were taken from the UEFA rankings for club competitions for the 2013-14 season. We implemented Python and bash scripts to parse this set information for a total of 124 games for 32 teams, with a total of $26,358$ passes between $3,428$ players over 248 networks with an average of 13.8223 players per team and 106.2823 passes per team.

For each team for each game in each stage, passing distributions included number of passes completed between all players as well as the total number of passes completed and attempted while squad lists included player names and positions.

6 Results

6.1 Baseline

The baseline model has an average loss of 15.00, which was calculated by summing up the individual loss for each player to player pass and dividing by the total number of passes during the round of 16 stage.

For the predicted number of passes to differ from the actual number of passes by 15.00 on average can be explained by the generalization made when averaging over all past passes. In addition, the games played during the Round of 16 have teams playing other teams that they did not play during the Group stage. Thus, the baseline is utilizing data that while generalizes how much players will pass to each other on average, fails to capture any changes in style a team may implement in the face of different opponents.

Feature Set	**ΔLoss**
Average Passing Networks	0.917
Same Position	0.069
Difference in Team Rank	0.506
Mean Degree	3.078
Betweenness of Receiving Player	5.774
Average Pass Completion % of Passing Player	0.096

Table 1: Best-performing feature set in Shared Weights Model. The average loss for the set is 11.21. The losses listed are the increase in average loss when the specified feature is removed from the set.

6.2 Linear Regression Model

6.2.1 Using shared model weights

Experimenting with various subsets of features proved to be very interesting. The features that computed average passes between two players, difference in team rank, whether two players are the same or different position lowered the loss consistently no matter which additional features were added to their sets. Intuitively, these three features have a lot to do with how many times two players will pass to each other. The average passes between two players is arguably the most indicative feature since that is the precise value we are attempting to predict. Difference in team rank is a good indicator of which team will possess the ball for the majority of the game and by how much, so for a team that is far better than their opponent, the feature accurately raises edge weights across the board for the entire team. Examining if two players are in the same position also accurately raises edge weights as it is more frequent for two such players to pass to each other due to sheer proximity on the field.

The average pass completion percentage for the passing player in a pair of teammates slightly reduced the loss consistently as well. It makes sense that a higher completion percentage will increase a player's out-edges.

We observed interesting outcomes during the use of combinations of three features: average betweenness of the passing player, average betweenness of the receiving player, and the mean degree of a team per match. Contrary to our hypothesis, average betweenness of the passing player performed better

Figure 2: Predicted passing distribution for Juventus during the Round of 16 when playing against Borussia Dortmund.

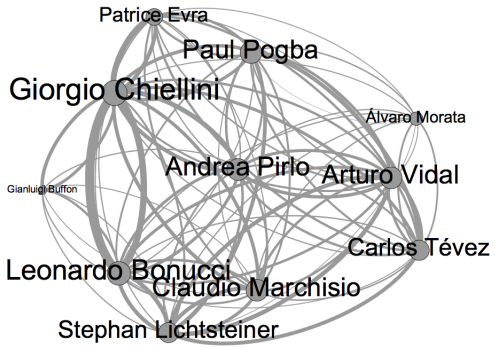
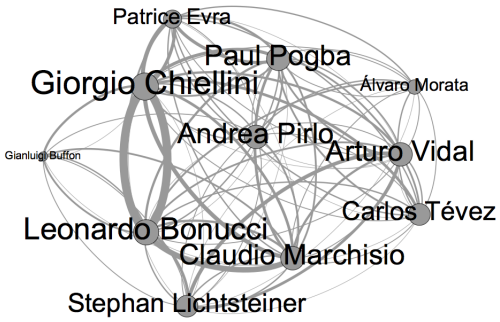


Figure 3: Actual passing distribution for Juventus during the Round of 16 when playing against Borussia Dortmund.



than average betweenness of the receiving player and the use of both features in the same set performed even worse than using only one of them. We expected the combination of both to produce the least amount of loss since both players' centrality in the lineup seems important to the frequency of the two connecting.

Additionally, we found that average betweenness of the passing player consistently reduces loss when used in a feature set that does not include the other two features, but using both average betweenness of the receiving player and mean degree in a feature set without average betweenness of the passing player has an even greater positive impact on the loss. We hypothesized that mean degree would have less of an impact on the magnitude of edge weights since it does not really capture the relationship between two players or the team's strategy against a specific opponent. Moreover, other features in our experiments, such as difference in team rank and whether a team has won against a similar opponent in the past, seem to be better indicators of the number of passes an entire team will make against a certain opponent. Yet, our best-performing set of features includes mean degree, which only performs well with betweenness centrality of the receiving player.

Figures 2 and 3 above depict the predicted and actual passing distributions of the team Juventus against Borussia Dortmund during the Round of 16 stage. Our predicted network generally has more edges than the actual network, as it is more likely to predict a real value that is non-zero than zero due to the nature of our scoring function. What is exciting is that our predicted network was able to capture the concentrated movement along the defensive line of Evra, Chiellini, and Bonucci. This defensive line is not always so concentrated, as can be seen in a game against Malmo FF during the group stage (figure 4 in the Appendix), where the passing concentration is more heavily weighted in the midfield and more spread out, thus for our features and weights to be able to tease this out is promising.

6.3 Using team-specific model weights

While one set of weights performed better overall than using team-specific weights, we noted many differences in the performance of features. For example, in the case of the boolean features of same position and different position between the receiving and passing players, same position worked best without different position in the single-weight model, but the pair worked best together in the team-specific weights model. Using both features most likely works well in the team-specific weights model because it is capturing the position-to-position passing style of each team, which does vary. An aggres-

Feature Set	ΔLoss
Average Passes Per Position	0.054
Same Position	0.532
Different Position	0.248
Difference in Team Rank	0.185
Betweenness of Passing Player	0.265
Average Pass Completion % of Receiving Player	0.129

Table 2: Best-performing feature set in Team-Specific Weights Model. The average loss for the set is 14.195. The losses listed are the increase in average loss when the specified feature is removed from the set.

sive team may have more frequent long passes from the defense or midfield to the offense, while a patient team may pass a lot between midfielders while waiting for an attacking opportunity to present itself.

Interestingly enough, the feature used in the baseline, average passes from one player to another, is not part of the best-performing set for this model depicted in table 2. The teammate relationship reduced the loss by the largest amount and was best indicated by the pair of features same position and different position. Individual passing player features performed decently in this model: betweenness of passing player and average pass completion percentage of receiving player. However, we expected these features' to work best with the opposite player in the teammate pair with betweenness of the receiving player and average pass completion percentage of the passing player.

The features characterizing team dynamics, average passes per position and difference in team rank, significantly lowered the loss. Average passes per position determines where on the field a team prefers to play, so it performed best when trained on team-specific weights as expected. Difference in team rank performed well too, but not as well as in the shared-weights model perhaps because each set of team weights has fewer games to train on than the 96 games of the group stage that were used to trained the shared weights in the previous model and difference in team rank is applicable to all teams.

The continuous feature of mean degree and boolean feature indicating whether a team won against a similar team did not perform as well as

Model	Avg.Loss
Baseline	15.000
Shared Weights	11.210
Team-Specific Weights	14.195

Table 3: Total loss for each model.

expected. Both of these features we believed to have the potential to capture a team's unique style of play and strategy, but perhaps the model needs more training matches for these features have a more positive impact. However, adding additional training matches could possible have a negative effect on the model due to the fact that a team's strategy can change a lot over time. These are experiments we can explore in the future.

6.4 Additional analysis

The superior performance of the shared-weights model is surprising because one would think that using team-specific weights would cater more to the idiosyncrasies of each team, but it seems that certain features are generally of the same importance to each team. Also, by using shared weights, the weights are updated far more often than if we use team-specific weights. So perhaps, it is a bit of both phenomenons coming into play.

Experiments with varying training iterations showed little change in either models' results, while modifying the learning rate produced drastic changes. The optimal values for these parameters resulted in 2 training iterations and a learning rate of 0.008.

We attempted two different training and test sets. First we tried training on the 96 Group Stage matches and testing on the 16 Round of 16 matches, which are the results seen in tables 1, 2, and 3. We also experimented with training on the 112 matches of the Group Stage and Round of 16 and tested on the 8 Quarter Final matches, which significantly increased loss in both the shared-weights and team-specific weights models. We believe this occurred not only because the test set is cut in half, but the nature of the quarter finals matches is more competitive making it more difficult for certain features to pick up on discrepancies between team strategy in the later stages of the tournament.

7 Problems Encountered

7.1 Parsing Data

While we were blessed with a large amount of per-game, per-player, and per-team analysis from the UEFA Champions League, we had to process the data, which was all in pdf form, into parse-able formats such as csv ourselves.

7.2 Feature Engineering

Not all features attempted were successful in lowering average loss. Not all the methods in which we attempted to implement our features were successful, sometimes seemingly intuitive but resulting in massive average loss and skyrocketing weights. We made choices of whether to precompute values for a given set beforehand, sort of simulating a genie knowing values beforehand, or to keep running averages, which is seemingly more realistic and chronological. For example, on matchday 1, a team does not know its average passing network for the entire tournament.

8 Future Direction

As mentioned before, team strategies change over time, so it would be interesting compare our results when we expand our training and dev set to more stages or even train on one season and attempt to predict games in the next season.

Similar to (Pena et al., 2012), we simplified the constantly changing landscape of the soccer field to a static network which was represented by set tactical lineups. It is indeed a limitation of a network to represent a static state, especially since our aim was chiefly to predict the edges between player nodes.

In addition, we also simplified positions to four values: goalkeepers, midfielders, defenders, and forwards. This fails to capture that positions have much more depth: such as defensive forwards, or defenders acting as midfielders. It would be interesting to have various networks representing different points in the game and to take into consideration how different positions evolve throughout a game or per team.

9 Contributions

Angelica: Problem statement, creating and running experiments, results, analysis of results and visualizations

Jade: Crawling data, parsing data, coding up algorithm and features, abstract, introduction, related work, model, evaluation, visualizations

10 Shared Project between CS224W and CS221

We are sharing the data that we parsed for this project for our final project for CS221.

References

[Cintia et al.2015] Cintia, P., Rinzivillo, S., Pappalardo, L. 2011. 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.

[Grund 2012] Grund, T. U. 2012. Network Structure and Team Performance: The Case of English Premier League Soccer Teams. Social Networks 34.4 (2012): 682-90.

[Pena et al.2012] Pena, J. L., Touchette, H. 2012. 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).

Figure 4: Actual passing distribution for Juventus during the Group stage when playing against Malmo FF, a lower-ranked team than Juventus. Note how the strong defensive line seen when playing Borussia Dortmund is not present here, but rather the passes are more spread out.