CS224W Reaction Paper and Project Proposal

Deepyaman Datta, Jade Huang, Angelica Perez
{deepyamd, jayebird, pereza77}

October 15, 2015

1 Literature Review

1.1 A network theory analysis of football strategies

This article [3] seeks to describe the strategy of football (American soccer) teams by representing passes made between players with a passing network. In this network, nodes are players and edges are weighted by the number of successful passes made between players. Though the football field is a changing landscape throughout the game, the network as a simplification statically represents the players' formation on the field.

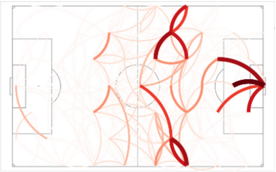
The idea is that by analyzing centrality measures such as closeness, betweenness, Pagerank, and clustering, one can uncover how players perform on a team. For example, if a team's network is very complete with edges between most or all players indicating that all players pass to each other, one could say that the team is very well-connected. Through betweenness, which indicates how ball flow between two other players depend on a particular player, one could uncover which players are more involved and perhaps should be isolated to prevent crucial passes.

By analyzing these measures, the authors were able to uncover interesting information about the teams in the round of 16 during the 2010 World Cup. The top-ranking teams–Spain, Netherlands, Brazil, and Argentina–exhibited a high number of passes, clustering, cliques, edge connectivity and an overall low betweenness score. The low-ranking teams exhibited the opposite: low degree connectivity and high betweenness, which makes the team seem disconnected and relying on a few players instead of all or many.

1.2 A network-based approach to evaluate the performance of football teams

This paper [1] generates football (soccer) match predictions through the use of teams' performance history. Two types of networks are built for each team: a player passing network and a zone passing network. Each network describes the movement of the ball for the duration of one match. In the player passing network, nodes are the players, while in the zone passing network nodes are regions of the field. Edges in both networks are ball movements labeled with weights representing the amount of interaction between two nodes.

The node degree distribution of the zone network illustrates where on the field a team prefers to play, and the degree distribution of the player network



indicates the involvement of each player. For example, the figure below shows the zone network for Argentina's national team during the FIFA World Cup 2014. We can see that most passes occurred in the midfield, and that many shots on goal were taken based on the thickness of edges on their offensive third of the field (right side).

The success of each network is ranked using three measurements:

1. u_i^T: the mean degree of a network's nodes, a proxy for the volume of play expressed by a team in a game
2. o_i^T: the variance of the degree of a network's nodes, a proxy for the diversity of play expressed by a team in a game
3. $H_i^T = 2/(1/u_i^T + 1/o_i^T)$: a combination of mean degree and degree variance represented as the harmonic mean of the two

A match prediction is computed by comparing these three values of two teams. At game i a team has a performance history described as a list $L = P_1, \ldots, P_{i-1}$, where $P_j = (\mu_j, \sigma_j, H_j)$. A prediction is based on the exponentially smoothed means of list L. This prediction algorithm achieved 53

1.3 Degree distribution of position-dependent ball-passing networks in football games

This paper [2] proposes a simpler model to describe the degree distribution of "position-dependent" soccer ball-passing networks. The degree distribution of ball-passing networks is one of many statistical properties that researchers have quantified in efforts to capture soccer match dynamics. Other elements explored in prior works include goal distributions, ball possession timings, and ball and player movements.

Even the interest in ball-passing networks in soccer games is not new. Earlier, we reviewed Lopez Pena and Touchette's analysis of these networks using PageRank, betweenness centrality, and clustering measures. Furthermore, Narizuka, Yamamoto, and Yamazaki have themselves previously modeled position-dependent soccer ball-passing networks. In fact, the result of this paper is not a novel feature for describing the ball-passing process; instead, it yields a simpler representation for the same.

The authors represent position-dependent ball-passing networks as graphs of player-location pairs connected by edges marking passes between the two nodes. Degree in such a network corresponds to the total number of passes made and received by each node. In their past work, Narizuka, Yamamoto, and Yamazaki found that, upon obtaining the aforementioned degree distributions from real data, they are fitted well by a truncated-gamma distribution. Notably, they discovered that imposing an upper bound on degree (hence the truncated distribution) more accurately described the networks than standard gamma distributions.

Another result from the authors' previous publication is the "Markov-chain model" as it applies to the ball-passing process. In this model, the probability of completing a pass between two nodes (as defined for position-dependent ball-passing networks) is directly proportional to (1) the rate of a successful pass of distance equal to that between the two nodes and (2) the likelihood of the receiving player being at the destination node, given a randomized starting position.

This extension changes our degree definitions slightly. Because we now account for both teams, the in-degree of a node increases when it receives a pass from a teammate. Likewise, the out-degree of a node increases when the associated player passes to a teammate. Using real data from nine matches (18 networks), Narizuka, Yamamoto, and Yamazaki calculate the shape and scale parameters to fit the data to a Weibull distribution. (Should we decide to extract a similar feature, we would examine more soccer games.)

2 Critique

2.1 A network theory analysis of football strategies

Unfortunately due to limitation of data from FIFA, the passing networks were computed 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. 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.

Another factor that is unrealistic is the static network since the football field is a constantly changing landscape. Unfortunately, it indeed is a limitation of a network to represent a static state, unless one were to have various networks representing different points in the game.

As noted by the authors, some interesting questions to further investigate would be to examine unsuccessful passes, as, of course, not all passes are successful. This could be taken into account using perhaps negative weights if we were to consider successful passes to have positive weights. In addition, the interception of passes could be another interesting factor to investigate in order to represent how strong is a team's defense.

2.2 A network-based approach to evaluate the performance of football teams

This oversimplified model fails to acknowledge several key features of a football match that could significantly improve a prediction. The team ranking measurements used here may be a good model of a team's general passing dynamics and player involvement, but it does not take into account how those dynamics will perform in relation to the dynamics of an opponent.

Moreover, this model only uses features involving ball movement. Other useful features could be player movements without the ball (although data for this may be challenging to find), halftime adjustments to player configuration, and amount of time spent in the offensive third of the field rather than the number of passes in this region.

2.3 Degree distribution of position-dependent ball-passing networks in football games

The extended Markov-chain model for ball passing proposed in this paper accounts for interactions with the opposing team. A weakness we observe in the authors' approach is the randomized starting position upon which the fitness, or distance to each node, is based. It is also unclear why they further discretized the distances by breaking the the field into identically-dimensioned regions rather than using pointwise distances. We perceive the arbitrary home location assignment as an unrealistic and perhaps unnecessary assumption since these positions should be governed by formations.

The benefits of the extended Markov-chain model over the the truncated-gamma distribution include fewer, more easily interpretable parameters while retaining similar fit. These qualities should prove advantageous should we use degree distributions of ball-passing networks in soccer games as features for predicting other attributes of matches. However, we need to revisit whether some simplifying assumptions (randomized starting positions, dividing the field into equal regions) are worth it when our end goal is not to describe the passing network but to use its characteristics as inputs to a larger problem. An alternative, more realistic approach to dividing the field into regions could involve shapes and sizes proportional to time typically spent in each box over the duration of a match. For our project, a reasonable approach may entail starting by calculating the shape and scale parameters for the ball-passing networks from each match as per the methodology described in this pasdfasdfaper, eventually eliminating some simplifications if initial tests point to degree distributions of these networks being influential predictors of our target.

3 Summary

All three papers model player passing networks based on real games from the World Cup and other sources. In such networks, nodes are players and edges represent ball movements weighted by level of interaction between players or number of successful passes. While Peña and Touchette explore how network centrality measures relate to the success of a team, Cintia et al. attempts to predict match outcomes. Narizuka et al. proposes simpler model to describe the degree distribution of "position-dependent" soccer ball-passing networks.

4 Model Testing

We tested the player passing network model developed by Cintia, Rinzivillo, and Pappalardo, where teams are ranked based on mean out-degree, variance of out-degree, and the harmonic mean of both. Our results were very similar to those achieved in their research. We used the statistics of 18 group round games played by six teams in the 2014/15 UEFA Champions League and tested on ten knock-out games played by the same six teams. The table below shows the prediction accuracy for each of the three metrics.

	Prediction Accuracy
μ	0.5
σ	0.6
$H_{\mu,\sigma}$	0.6

5 Project Proposal

After reviewing prior work related to the soccer network analysis, we identified two potential problems of interest. The first entails improving upon the match prediction conducted by Cintia, Rinzivillo, and Pappalardo. By experimenting with a wider array of features, some from the other literature that we reviewed and others derived from our own understanding, we intend to achieve greater success in identifying better teams. More information about features that we are considering follows below.

The second problem represents a more novel application of soccer network analysis. Training on data from previous seasons and applying the insights using up-to-date statistics from UEFA Champions League press kits (the documents for each matchday are published in advance of the following matchday), we plan to have agents participate in the fantasy UEFA Champions League. We intend to have an AI compete either against other instances of the AI (with different parameters) or against us. A final, cooperative variation on this scenario involves letting one AI acquire whatever players it wants (within fantasy league budget limitations).

For both of the aforementioned tasks, we will build a UEFA Champions League dataset. Unfortunately, while quite comprehensive, the (press kits) aren't readily parseable. We will do the data cleaning ourselves.

The project itself spans both this class and CS 221: Artificial Intelligence: Principles and Techniques. While our proposal is relevant to both classes, certains tasks are more closely tied to the material covered in each class. The portions that focus heavily on concepts from CS 224W relate to extracting network-based features. A few of the features that we are keen on experimenting with are:

- Degree distributions of soccer ball-passing networks. As evidenced in a couple of the papers we discussed earlier, various factors can be taken into account when defining these networks, including player positions from both teams, pass length, and player "fitness" (i.e. the distance covered to receive a pass). Options include utilizing raw data in our predictive model or first fitting to an appropriate distribution and using the model parameters.

- Centrality measures. As with node degree in ball-passing networks, these can either be viewed holistically for each team or individually for each player. While the former is likely more appropriate for the match prediction task, individual statistics will be critical for fantasy team formation.

A couple of the centrality metrics that we will work with include PageRank and betweenness, trying to identify the most effective players for each role.

We also plan to calculate a number of score features based on the press kit data:

- Aggression. Fouls committed, yellow cards, and red cards all contribute to this metric.
- Attack coefficient. Total scoring attempts (on target), time spent in offensive third, and number of passes in offensive third positively influence this measure, while offsides have an opposite effect.
- Defense coefficient. Shots blocked, goalie performance (blocks), occurrences of defensive players making runs to the offensive third, and passes intercepted are indicators in this department.

We are also exploring accounting for a number of other variables, such as performance differences between the first and second half, possession, and player movements not involving the ball (if available). If feasible, one of the factors that we are especially interested in exploring in the context of match prediction are the effects of formations and more nonstandard motifs employed by teams.

Match prediction and fantasy team formation given these features constitute some of the more AI-related work for this project. Stochastic gradient descent combined with least squares regression to learn feature weights is a candidate approach to the first problem. Performing well in the fantasy league, especially in a non-cooperative environment, introduces additional optimization problems. The availability of desired players is not necessarily guaranteed, and funds further need to be managed to ensure a well-balanced fantasy team. Fantasy team formation is also different in the sense that some network features may be less relevant, especially those signaling cooperation between subsets of players in teams.

Evaluation for match prediction will involve cross-validation on historical data. We also have the results from Cintia, Rinzivillo, and Pappalardo to compare against. Fantasy team performance is more interesting to judge. In a cooperative environment, we can compare the team's performance to the theoretical maximum that could have been achieved. On the other hand, in a non-cooperative environment, we will evaluate based on our performance in the league (or compare the performances of multiple AI with different strategies in the same league). Finally, we can also benchmark against results that would have been obtained by actual starting lineups in the league. If competing against other people, our goal is for our algorithm to win the league in which it participates.

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).