Graph Analysis of Major League Soccer Networks: CS224W Milestone

Evan Huang
Stanford University
ehuang@stanford.edu

Sandeep P. Chinchali
Stanford University
csandeep@stanford.edu

1 INTRODUCTION

Given the natural existence of networks in team-based sports, we hope to use network analysis and predictive modeling to better analyze soccer. Currently, research in soccer analytics has focused on individual player statistics, and predictive modeling has been limited to simple logistic regression, decision trees, and LSTMs [3]. We believe that leveraging network structure will create better results given the importance of teamwork within the sport. In particular, a better predictive model can benefit both the sports betting industry (last year, the total betting on sport was $4.9 billion in Nevada alone [1]) and the soccer team's management.

2 RELATED WORK

One of our main novel contributions in this project would be assessing whether graph-based learning algorithms like Graph Convolutional Networks (GCNs) [2] perform better than other traditional learning algorithms in the context of prediction and analysis of sports games.

There has been some related work in sports analytics conferences like the MIT Sloan Sports Analytics conference; for example, one paper [3] describes data-driven ghosting in soccer games which enables coaches and managers to âĂIJscalably quantify, analyze and compare fine grained defensive behaviorâĂİ. Their learning task was different than our proposed one here because they were more interested in analyzing and modeling player movements and defense styles, not predicting game statistics. Other related work includes an MIT Master's Thesis [4] and a journal paper [5] which also use similar data from soccer games to infer passing patterns and styles [4], assess playersâĂŹ passing effectiveness, and predict shots [5]. However, to the best of our knowledge, none of these papers represent the data as a network graph and utilize the graph structure in their inferences, which is where our current work is situated.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM. . . $15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

3 DATASET AND GRAPH STRUCTURE

Our dataset consists of 6 seasons of âĂIJplay-by-playâĂİ soccer games from major professional leagues such as Li Liga (Spain), English Premier League (EPL), and Major League Soccer (MLS). For a given game between two teams, âĂIJplay-by-playâĂİ data identifies each player, their x and y coordinates in the field, a timestamp of hour, minute, and second, and the major action taken by that player, such as a pass, interception, aerial shot, goal etc. in a standardized vocabulary.

We collected the annotated data from OPTA, a sports analytics company, which ensured data fidelity and annotated player actions (passing, aerial shots) using uniform video annotation techniques. Overall, the dataset consists of 2,280 games from players across 60 diverse teams, leading to 3,893,304 player actions (rows). Since only one player is considered in each timestep, we can identify player A passed to player B by considering the action and players in successive timestamps.

3.1 EPL Dataset

To prototype our methods, we have only tested our analysis on the EPL dataset. For the final project, we will repeat the analysis for all leagues, cluster playing styles across leagues, and compare node feature vectors between leagues.

The EPL dataset consists of a single season in 2012, consisting of 380 matches between 20 teams, such as Liverpool, Manchester United, Southampton, etc. The dataset consists of 648,883 unique plays made by 524 unique players, who are annotated to have 5 positions of Strikers, Defenders, Midfielders, Goalkeepers, and Substitutes. Each play is annotated using a standardized vocabulary of 46 actions, including 'Goals' and 'Interceptions'.

3.2 Graph Structure

The team structure within a given match is represented as a weighted, directed graph, where nodes are players and edges are actions (pass, kick, etc.). Our initial networks also include additional event nodes to represent non-player states such as the gaining of the ball, the loss of a possession, and a shot taken. Respectively these are named âĂIJGainâĂİ, âĂIJLossâĂİ, âĂIJShotâĂİ, and âĂIJGoalâĂİ.

(1) *Nodes:* 14 to 17 nodes consist of 11 players from each team plus events and substitute players. As of now we encode no additional information on the nodes except

for ID, but will eventually add in position, location, and time played.

(2) *Edges:* Actions between states where the ball is located. If player (node) A passed to player (node B) some amount of times within a game, a passing edge is created. Concurrently, shot, gain, and loss rates are also used to connect player nodes to event nodes. The weights on the edges can be formulated as follows:

(1) Passes (A, B) = (num. successful passes between A and B) / (time shared between A and B)
(2) Shots(A) = (num. saved attempts, post hits, misses, or goals) / (time A on field)
(3) Gain(A) = (num. of ball recoveries, corners awarded, out of bounds rewarded) / (time A on field)
(4) Loss(A) = (num. unsuccessful passes, out of bounds, dispossessions) / (time A on field)

Networks currently in production represent a single team in a single game, such as:

(1) Single game networks with 2 teams. Each team is connected through respective losses and gains of possessions, transforming original âĂIJGainâĂİ and âĂIJLossâĂİ nodes into âĂIJSwitchâĂİ nodes.
(2) Aggregated team network consisting of data within a whole season. Each teamâĂŹs individual networks is aggregated across all games, giving one unique network per team. Rates are now based on total time rather time in a game.

4 GRAPH ANALYSIS

5 FUTURE WORK

6 FUTURE RESEARCH DIRECTIONS

REFERENCES