City University London

MSc in Data Science

Project Report

2019

Understanding player and team performance through football data analysis

Shengqiang Fan

Supervised by: Dr. Cagatay Turkay

1st October 2019

By submitting this work, I declare that this work is entirely my own except those parts duly identified and referenced in my submission. It complies with any specified word limits and the requirements and regulations detailed in the assessment instructions and any other relevant programme and module documentation. In submitting this work I acknowledge that I have read and understood the regulations and code regarding academic misconduct, including that relating to plagiarism, as specified in the Programme Handbook. I also acknowledge that this work will be subject to a variety of checks for academic misconduct.

Signed: Shengqiang Fan

Abstract

This project introduces a novel approach for talented young player identification, team performance evaluation and collaboration performance evaluation. The overall metric includes goal probability creation, possession sequence creation, value assignment for each sequence, sub possession sequence creation, similar sequence check and value assignment for each action, network analysis. Opta's half season data of French Ligue One 2016-2017 were used as the test set and Statsbomb's data of La Liga 2004-2016 seasons and World Cup 2018 were used as the training set in this research. Talent young players in French Ligue One of season 2016-2017 were identified by this metric and it was proved by their market value changes that most of identified player by this metric were worth the investment. In addition, result from action evaluation can be also used to assess team performance by simply aggregating all players' performance value in same team which can provide a new way to analyze current situation of team. Besides, network analysis was applied on Opta's data for collaboration performance evaluation. It is found that betweenness centrality should be minimized ideally, degree centrality, transitivity and density should be maximized in order to improve collaboration between teammate to win the match.

Keywords: Dynamic Time warping, Football Performance Evaluation, Network Analysis, Sports Analytics, Valuing Actions.

Content

1. Introduction and Objectives

1.1. Background and reasons

Sports analysis gains lots of attentions due to the big influence of the book *Moneyball* written by Lewis (2003). It demonstrated how a baseball manager find undervalued player by using sabermetrics. More and more books and papers were published to provide guidance of how to use data for decision makers in sports such as coaches and managers. In recent years, with more and more data can be obtained with rapid development of technology, football player's performance can be judged in a new data driven view instead of subjective view. Shot on target, total goals and assists are well-known performance indicators based on simple statistical information. In addition, expected goal model was put forward by Ensum *et al* (2005) in order to evaluate players' shot ability, which drew a great attention on sports analytics community in the past few years. Regression technique was used to measure the probability of a shot attempt by tracking several shot-related features such as shot angle, shot distance and body part used to perform the shot (left footed shot/right footed shot/headed shot/others). However, these evaluation metrics are focus more on result of a shot event while passing and dribbling events take a large proportion of the match. Similar limitation existed in other research such that Decroos *et al* (2017), Bransen & Van Haaren (2017) and Szczepański & McHale (2016) tried to evaluate player's performance by valuing passes but dribbling events were not included. Another difficulty of performance evaluation is how to address the value of actions in context. For example, a successful long pass may not result in a goal immediately but can create an open space to build up a goal within actions after that long pass. Therefore, one aim of this project is to propose a framework to evaluate players' contribution to fill the gap between existing evaluation metric. More contributions of a player mean more threats/opportunities can be created for his team. Intuitively, aggregate contribution of all players in the same team can indicate how many opportunities were created leading to win compared to opponent. Besides, results of players' contribution can be also used to assess team performance which can provide a new data driven view for football coach. Moreover, football is a team sport since an excellent performance of single player still cannot lead to a win so that collaborative effort by team members should also be considered. Despite the importance of team-level considerations, there are only limited techniques, one possible metric is to evaluate collaboration performance by considering the *whole team as a network*. Hence, passing network analysis will be implemented on the basis of social network theory because passing between teammates can be treated as a small network between each player linked by passing event. Power *et al* (2017) pointed out that the most frequent event during a football match is passing, qualitative and quantitative analysis on passing

data may give some interesting insights. Therefore, this project is also aimed to find differences of network-based indicators between winning teams and losing teams, which can help people get a better understanding of the collaborations between teammates at team level.

1.2. Research questions and objectives

Three main related research questions are addressed in this project:

- Can we develop a data-driven technique to identify talented young players in a league?
- Can we also develop a data-driven technique to quantify team's performance rather than only considering its position on league table?
- Is it possible to evaluate collaboration performance at team level by using network science theory?

In order to answer these research questions, several objectives stated below are needed to be achieved:

1. To understand the source data well provided by Opta Sport company and Statsbomb company in order to make use of these football event-based data.
2. To develop a metric to evaluate players' performance which accounts all types of on-ball actions (such as pass event and dribble event) and considers contextual impact of each action.
3. To develop a data-driven technique to identify talented young players whose value in transfer market has a potential to grow rapidly over years based on the results from the proposed metric of performance evaluation.
4. To develop a data-driven technique to compare team's performance with its position on league table based on the results from the proposed metric of performance evaluation.
5. To develop a metric which can evaluate the collaborations between teammates by considering the whole team as a network. To get some insight based on the characteristics of network-measures for winning team compared to losing team.

1.3. Beneficiaries

- Football coaches and analysts who can use data to support their subjective inference on player's contribution in attacking. Meanwhile, coaches can adjust tactics based on information from individual player's performance and team performance in order to improve team's performance.
- Club manager and football scout who can use data to identify if a talented young player whose value in transfer market has a potential to grow rapidly over the years. Consequently, it can help them to make decisions whether to purchase a young player or not.
- Data scientists / researchers who are working in this area can be inspired by the proposed evaluation metrics for further analysis. Related metrics can be tested on their own dataset and to see if it can resonate with football expert.

1.4. Methods and work plan

Research question 1 and 2 are highly related in performance evaluation at individual level. Research question 3 is relative independent which evaluate performance at team level.

- Method for achieving objective 1: There are documentations provided by Opta Sport and Statsbomb to explain their dataset in detail. Differences should be identified for later transformation before data merge although they are all event-based data.
- Method for achieving objective 2: There are two major challenges discussed in introduction when players' performance evaluation. One is that all types of on-ball actions should be evaluated and the other one is that contextual impact of each action should be considered. Therefore, the proposed evaluation metric is divided into 5 phases:
    i. Goal probability model creation.
    ii. Possession sequence creation and value assignment based on goal probability model.
    iii. Sub possession sequence creation before and after action (including passing event and dribble event)
    iv. Trajectory similarity analysis based on dynamic time warping and assign value to each possession sequence.
    v. Each action is valued by the difference between value of sub sequence after action and before action.
- Method for achieving objective 3: Based on result from action evaluation, individual player's contribution to his team can be assessed by aggregating all actions corresponding to that player. On account of different players have different quantity of actions, total action value cannot be used directly to compare performance among players. Thus, value per action is computed as a normalization technique with specific filter criteria (age, total appearance time) which are used to remove noise. Then, the remaining players with higher value of value per action are considered as top talented young players.
- Method for achieving objective 4: Based on result from action evaluation, team performance value of each match can be simply aggregating all player's performance value in same team. Then, two lines can be tracked match by match of same team to test their correlation, one is cumulative performance value and the other one is cumulative points gained. In addition, the ranking based on team's aggregated performance value compared to the position at league table can reveal some interesting insight.
- Method for achieving objectives 5: It was mainly achieved by comparing average value of six network-based parameters (number of edges, degree centrality, transitivity, density,

eigenvector centrality and betweenness centrality) between winning team and losing team.

Work plan for this approach is divided into 3 phases:

i. Parsing data from original dataset into desired format.
ii. Passing network for each team is generated and network-based parameters are measured.
iii. Difference between winning team and losing team in characteristics of network measures is assessed by correlation analysis.

1.5. Major changes in goals and methods

The main goals of this project did not change but objectives were adjusted to be more specific and result-oriented from the initial proposal. In terms of performance evaluation at individual player level, traditional machine learning techniques (random forest, logistic regression and XGboost) were not used because trajectory analysis was considered as a more suitable algorithm for similarity check, which was more convinced and explainable with the aid of visualization of similar trajectories. In terms of performance evaluation at team level, dynamic network analysis with dashboard in real-time was proposed but did not accomplished by this project due to time constraint. However, comparison on characteristics of network measures for winning team and losing team was accomplished for consistency. Therefore, the current project still maintained the complexity of project with adequate and interesting results.

1.6. Report structures

This project report is divided into 6 chapters:

Chapter 1 introduces the background of this project as well as reasons. Objectives are pointed out in this chapter. In addition, general method is provided to achieve objectives.

Chapter 2 makes a literature review on state of art in sports analysis, especially for Data-intensive team-sport analysis. Some technical foundations are also provided in this chapter.

Chapter 3 gives a detailed method to achieve the aims of this project. Methods for talented young players identification, team performance analysis and collaboration evaluation are provided, which include data manipulation, goal probability evaluation, possession sequence creation, trajectory similarity analysis, passing network creation, and network science analysis.

Chapter 4 presents the outputs from analysis process of this project. talented young players identification, team performance analysis and collaboration evaluation are presented separately.

Chapter 5 examines the output compared to objectives and existing literatures. Implications of result and specific limitations are also discussed in this chapter.

Chapter 6 makes an evaluation of this project including conclusion, general limitation, further work and reflection.

2. Context

2.1. Overview

With rapid development of information technology, lots of high-quality data can be obtained. Opta Sport company and Statbomb company capture all event-based tracking data from every second during the match. More specifically, different football events, such as passing, dribble, shot and tackle, were recorded with corresponding x, y location on football pitch as well as the player number with timestamp. These data make performance evaluation possible rather on stay on basic statistical analysis such as shot on target, passing completion, possession rate and so on. Based on the research aim of this project, literature review is mainly focus on two specific sub-areas: player performance evaluation at individual player level and collaboration performance at team level separately.

2.2. State of the art in team-sport analysis

2.2.1. Player level analysis

Several researches were conducted on performance evaluation of a football player in the past few years. Szczepański and McHale (2016) put forward to use generalized additive mixed model to quantify the passing ability of a football player and point out key players based on that passing model. But only pass difficulty was considered in that model. Brooks and Guttag (2016) designed a player's performance ranking system on basis of the value of passes. A linear classifier was trained to find out the importance of pass start locations and end locations for creating a shot opportunity. Therefore, any passes can be valued. This approach can avoid problems that only passes directly leading to shot events were considered. Moreover, Decroos *et al* (2017) proposed an action rating system based on event-based tracking data which were also called play-by-play data. The rating system consists of three parts. First part is to split event-based match data into different phases. A phase is a consecutive event which starts when possession switches and ends if a dead ball event occurs such as corner/free kick is award as shown in figure 1. Trajectory analysis was used to compute similarity between two phases. Then any phase can be valued by averaging the outcome (whether a goal is scored or not) of k nearest neighbors (k=100 was proposed in their paper). Each event in one phase was assigned a normalized weight on the basis of exponential-decay method, which means the event at the beginning of a phase has relative low weight. Similar research was conducted by Bransen and Van Haaren (2017), and they trained an expected goal model to evaluate the outcome of one phase instead of computing average number of many phases directly leading to a goal from Decroos's research. However, primary concern of that research was only pass event, which means other on-ball events like dribbling were not considered. In term of expected goal method, it was initially put forward by Ensum *et al* (2005)

in order to explore important factors that affect goal probability of a shot on dataset of 37 matches of world cup. Shot distance, shot angle, defender's pressure and whether the shot was followed instantaneously by a cross were pointed out as the most influence factors from his logistic regression model. Expected goal model is widely accepted by sports analysis community nowadays to compute the probability of goal when given a shot opportunity under different situations. It can be used to evaluation player's shot ability by comparing difference of total number of goals expected to score and actual goals in total. Eggels (2016) explored the performance of 4 classifiers (logistic regression, decision tree, random forest and Ada Boost) used to train the expected goal model. Random forest was outperformed in AUC and F-score under 13 features were created for each shot opportunity. Season analysis was put forward to track a club's behavior by aggregating all players' expected goals compared to the team's actual goal during the season. Juventus in 2015-16 season had a bad opening, only won three matches of first 10 games. Fans and news reporters started to question players and coaches. However, expected goal provided a new way to analyze that problem. It is shown that Juventus' s expected goal value was much higher than actual goal scored in first 10 games, which indicated that Juventus were dominating matches and creating lots of opportunities but not seized the goal opportunity. It also indicated that playing style or tactic were not the actual problem.

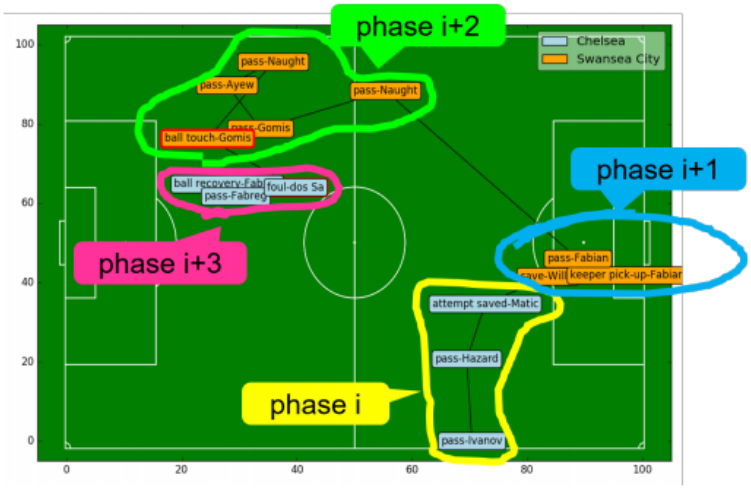


Figure 1: Example of sequence of events are split into several phases (Decroos et al, 2017)

Similar investigations were performed in other sports. A Markov Model was constructed by Schulte *et al* (2015) to find the most valuable action leading to winning in ice hockey. Similarly, Chan and Singal (2016) also proposed a Markov Decision Process system for tennis. Base on that, football game can be also treated as a Markov Model where transient states is different location of football pitch and absorption state could be the outcome of a possession (whether a goal is scored or not). Both Transition probability and absorption probabilities can be calculated from historical match data. Nonetheless, possible limitation of Markov Model approach could be 'memoryless' property of Markov model is not true in football events.

2.2.2. Collaboration performance analysis at team level

A passing network is created from successful passes between teammates, where nodes of network are players with directed connections by passes between players. The concept of passing network in football was firstly introduced by Gould and Gatrell (1979) but did not gain much attention from sports or scientific community. Related researches on passing network just increased gradually in recent years. Pena and Touchette (2012) indicated that passing network can be used as a visualization tool because it can give the insight into statics right away as shown in figure 2. Key player can be easily pointed out through closeness centrality however, it has its bias when using passing network. Midfield players do have higher closeness centrality due to its role in the team. Betweenness centrality is a different concept from closeness centrality, it can calculate how the ball passes among other players depend on one typical player. In other word, a player with high betweenness plays an important role in team's tactic. And if betweenness near to 0, it indicates that the low involvement of that player. An evenly distributed and relatively low betweenness should be preferred as it indicates a balanced passing strategy from tactic view of the whole team in general. PageRank centrality was also introduced by Gould and Gatrell (1979) which can identify a relationship between key players. A good defending tactic is to reduce PageRank centrality of rival's key player in order to decrease the connection between key players. Srinivasan (2017) also applied network analysis on football. Cliques, Average Clustering Coefficient, top eigenvalue and λ distance similarity were used to track the team performance in certain time period. Gonçalves *et al* (2017) suggested that a well-balanced passing network with a lower passing dependency may be a best combination to maximize performance of a team. Recently, social network analysis was applied in Australian Football League 2009-2016 seasons (1516 matches included) by Young *et al* (2019). Seven network measures such as the edge density, betweenness centrality and eigenvector centrality were derived to find out optimal teamwork measures for winning team. Results indicated that edge density, transitivity and betweenness centrality play an important role on team performance while betweenness centrality's correlation is not significant. On the contrast, average path length, eigenvector centrality and degree centrality should be minimized due to its negative effect for team performance. Nevertheless, three new nodes (behind, goal, missed shot) were added during this network analysis which may affect the network measures because the remaining nodes in passing network were all players. In addition, only moderate correlations were identified may due to its complexity of passing network, which can be improved by adding some constrains when passing networks were created. Besides, Kröckel *et al* (2017) evaluated team performance by using dynamic network analysis approach to investigate team performance changes during the 90 mins of the game, which gives a good starting point. There is also agreed by Kröckel *et al* that a well-

distributed passing networks could result in better performance. Based on the dynamic network analysis, decisional helps to the coach in real-time could be one of applications, suggested by Kröckel *et al* (2017).

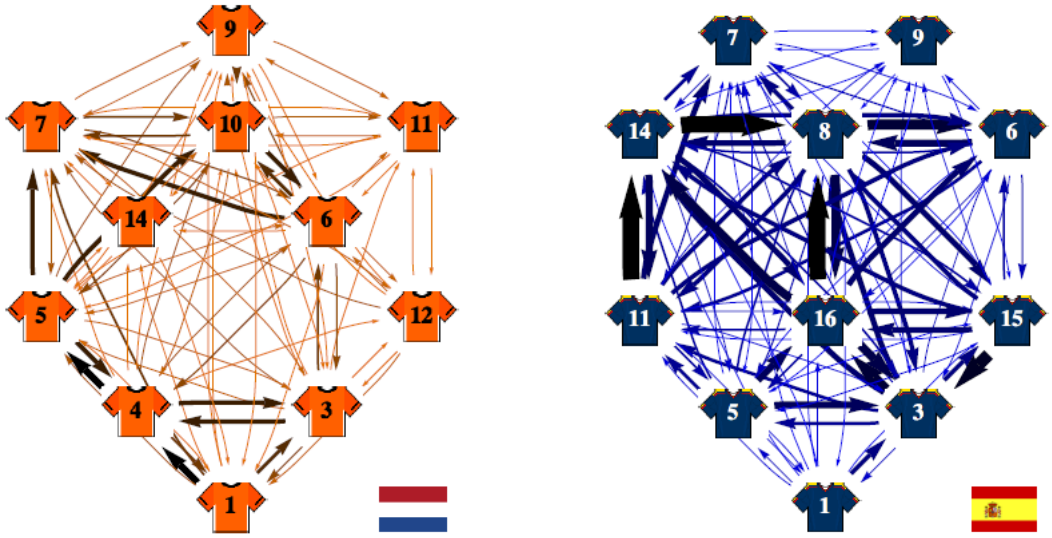


Figure 2: Passing network created by Pena and Touchette (2012)

2.3. Technical foundations

2.3.1. Dynamic time warping

As mentioned in previous chapters, attacking possession sequence can be treated as several trajectories with timestamp. Therefore, dynamic time warping algorithm was used in this project because it is a widely used technique in trajectory analysis (Müller, 2007). It was intensively applied in speech recognition as well as many other areas such as gestures recognition, online signature matching and time series clustering (Senin, 2008). Dynamic time warping can measure the similarity between two time series sequences, which may have different speed or acceleration (shifting and distortion in time). Due to this advantage, dynamic time warping is appropriate for similarity measurement of football possession sequence. Example from Keogh and Pazzani (2001) is used in order to explain this technique more specifically. Suppose there are two time series D and T with length n and m respectively:

$$D = d_1, d_2, d_3, \ldots, d_n$$

$$T = t_1, t_2, t_3, \ldots, t_m$$

A matrix C (n-by-m) is created in order to align sequence D and T, where C (d_i, t_j) is defined as the distance between points d_i and t_j in sequence D and T respectively. Therefore, a typical warping path (W) can be created as shown in figure 3, where $W = w_1, w_2, \ldots, w_k$ and $w_k = (i, j)_k$. This warping path was defined with three overall constrains:

- Boundary condition: The warping path should start at point (1,1) and end at point (m, n)

- Monotonicity: Given $w_{k-1} = (x', y')$ and $w_k = (x, y)$, it must satisfy that $x \geqslant x'$ and $y \geqslant y'$.
- Continuity: Warping path can only move one cell at one time. Given $w_{k-1} = (x', y')$ and $w_k = (x, y)$, it must satisfy that $x \leqslant x'+1$ and $y \leqslant y'+1$.

There will be exponential number of warping paths which satisfy all three conditions, however, only the warping path which has the lowest warping distance will be the optimal path. Thus, dynamic programming method is used to find out this optimal path as defined:

$$\gamma(i, j) = C(d_i, t_j) + \min\{ \gamma(i-1, j), \gamma(i, j-1), \gamma(i-1, j-1)\}$$

where $\gamma(i, j)$ is cumulative distance at (i, j) and $C(d_i, t_j)$ is the distance between points d_i and t_j. Figure 4A is two time series sequence and Figure 4B illustrate an example of dynamic time warping alignment. It is obvious that the alignment between these two sequences can be warped non-linearly in time in order to find similar patterns.

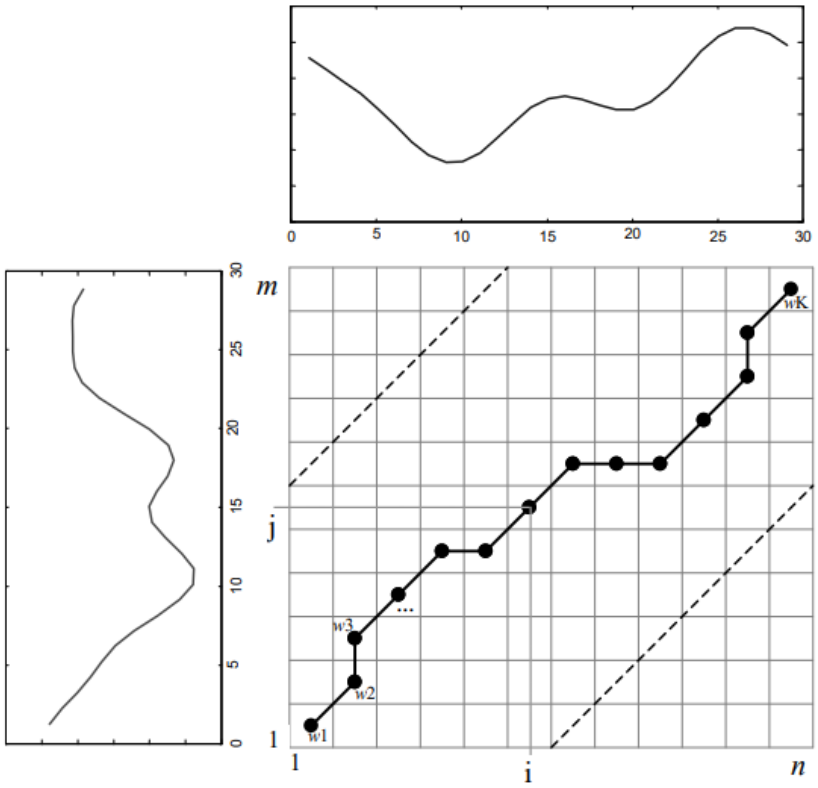


Figure 3: Example of warping path by Keogh and Pazzani (2001)

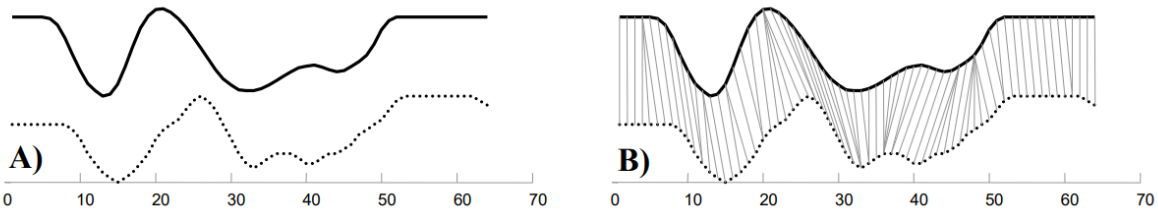


Figure 4: Example of dynamic time warping alignment by Keogh and Pazzani (2001)

2.3.2. Network-based knowledge

As discussed in section 2.2.2, passing network can be created by treating players as nodes, successful passes as directed edges between nodes and number of successful passes as edge weights. Therefore, network-based analysis can be applied to the generated passing network. Centrality measures are often used to determine the importance of a node in a network. Degree centrality is the simplest one in centrality measures, which only counts the number of edges link to target node and be normalized by maximum possible degree numbers. Betweenness centrality is also used in this project, which was proposed by Freeman (1977):

$$C_B(v) = \sum_{s,t \in V} \frac{\sigma(s,t|v)}{\sigma(s,t)}$$

Where σ (s, t) is the number of shortest paths between node s and t, σ (s, t |v) represent the number of the shortest path between node s, t and passing through node v as well. Intuitively, if a player node has higher value in betweenness centrality, this player could be a midfielder with higher probability because midfielder normally plays an important role to link the whole team.

The last centrality measure used in this project is eigenvector centrality. Main difference between eigenvector centrality and degree centrality is that the relative score of adjacent node affects a lot in eigenvector centrality instead of only the number of edges matters in degree centrality (Newman, 2018):

$$Ax=\lambda x$$

Where A is the adjacency matrix of network. And this adjacency matrix. The value in adjacency matrix will be 1 if two nodes are linked otherwise will be 0. Accompany with Perron–Frobenius theorem, unique eigenvector can be identified with the largest eigenvalue.

In addition, transitivity is another valuable network measure. It indicates that the probability of adjacent nodes' trend to connect the selected node, which is inspired from a phenomenon that "friends of friends are often friends". Barrat *et al* (2007) stated that a node (a) with high value of cluster coefficient represents that if a&b, and b&c are linked separately, then the probability of node and c is connected will be relatively high. In terms of football application, transitivity can reveal the collaboration among three players with tacit cooperation to break the defense line.

3. Methods

3.1. Overview

As discussed in chapter 1 and 2, objective 1 is fully understanding of the dataset. Dataset consists of 601 matches in total, of which 64 matches data from World Cup 2018, 347 matches data from La Liga 2004-2016 seasons, 190 matches data from French Ligue One 2016-2017 season. It will be discussed in section 3.2 in detail. Meanwhile, research question 1 and 2 are highly related which are based on action evaluation metric (objective 2). Figure 5 illustrates the metric for action evaluation. It is divided into 5 phases. Overall, 347 matches data from La Liga 2004-2016 seasons are treated as test set and rest of matches as training set in this project. In other word, the evaluation of individual player's performance was only conducted on dataset from French Ligue One 2016-2017 season. In the first phase, goal probability model is created based on historical data from all 601 matches and it will be used in this project to assign value to possession sequences. Possession sequences are created in the following phase, and each sequence is assigned a value based on goal probability model in the meantime. Possession sequence in test set will be split into two sub possession sequences in phase 3, which are sub possession sequence before action and after action. Then, Dynamic time warping technique is proposed to find similar sequence in training set when given a sequence from test set in phase 4. Possession sequences in test set are valued by averaging 50 most similar sequences' value in training set. In the final phase, player's action contribution can be assessed by computing the difference between value of possession sequence after action and before action. Based on result from action evaluation, talented young players can be identified (objective 3) which will be discussed in section 3.4. Value per action will be the key indicator to find player who with the best performance. In section 3.5, team performance evaluation will be illustrated based on result from action evaluation as well. The main idea for team performance evaluation is aggregating all player's performance value in same team. In section 3.6, method for collaboration performance evaluation between teammates will be discussed. Whole team is considered as a network so that social network analysis can be implemented to find collaboration performance indicators.

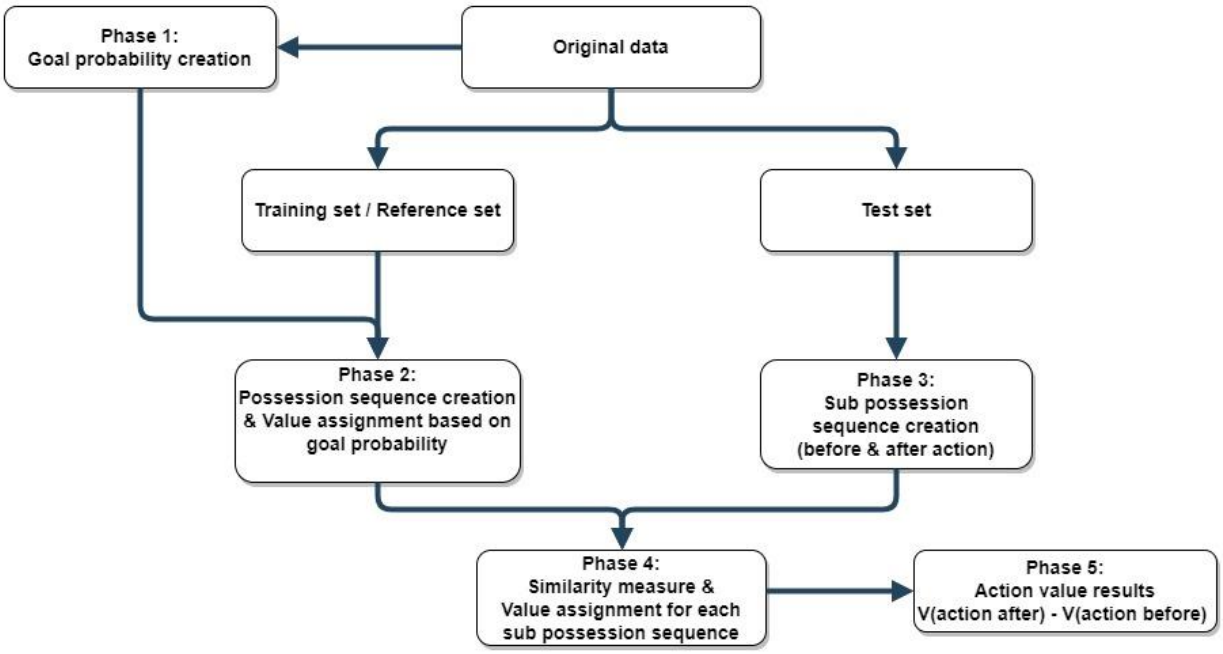


Figure 5: Illustration of metric for action evaluation

3.2. Dataset description (Objective 1)

Opensource data are provided by Opta Sports company (OptaPro, 2019) and Statsbomb company (Statsbomb, 2019) in XML and Json format respectively. Both datasets are event-based sequential data which can be used for this project. The dataset from Opta includes 190 matches from first half season of French Ligue One 2016-2017 season and the dataset from Statsbomb consists of 64 matches data from World Cup 2018 and 347 matches data from La Liga 2004-2016 seasons. Both datasets are recorded in what-who-where-when structure of each match. More specifically, sequential events such as passing, dribble, shot and tackle are recorded with corresponding x, y location on football pitch as well as player number at timestamps. Figure 6 gives an example of XML data of one event information. It is obvious that timestamp, min, sec, x and y location are recorded. However, other information such as 'player_id' and 'type_id', 'qualifier_id' were stored in numerical value, which is not straightforward. They can be translated with the aid of two documents: 'Players and IDs - F40 - L1 20162017' and 'Opta Event Details Feed'. The former was used to look up player's specific information by given 'player_id'. The latter one explained the actual meaning corresponding to 'type_id' or 'qualifier_id'. For example, type_id = 1 in figure 5 which means it is a pass event made by player '59963' (Maxime Gonalons). The outcome, if equals to 1, indicates that this pass is a successful pass otherwise outcome value is 0. 'qualifier_id' gives more detailed information correspond to that event. 'qualifier_id' = 212 represents length of this pass event, 'qualifier_id' = 141 describes pass end y location, 'qualifier_id' = 213 records pass angle in radians and 'qualifier_id' = 140 describes pass end x location. All these values related to 'type_id' or 'qualifier_id' have a detailed

description in the document provided by Opta Sports. Therefore, Python language and Python Standard Library ElementTree package were used to parse XML dataset from Opta Sports into desired format, such as shot data and pass data, which will be discussion in the following sections.

```
- <Event timestamp="2016-08-14T14:00:41.295" id="2094570878" version="1471179643220"
  last_modified="2016-08-14T14:00:44" y="78.8" x="22.1" outcome="1" team_id="143" sec="15" min="0"
  period_id="1" type_id="1" event_id="6" player_id="59963">
      <Q id="1873397150" value="Back" qualifier_id="56"/>
      <Q id="991726621" value="6.0" qualifier_id="212"/>
      <Q id="399659593" value="83.0" qualifier_id="141"/>
      <Q id="259381996" value="2.6" qualifier_id="213"/>
      <Q id="588154786" value="17.1" qualifier_id="140"/>
  </Event>
```

Figure 6: Example of XML data format

As for dataset in json format provided by Statsbomb, it is more straightforward compared to XML data as shown in figure 7. Although it has value named 'id' related to different event types, it includes description in original data. For example, type id = '30' with name = 'Pass', player id = 3244 with name = 'John Stone'. Therefore, it is no need to link any other documents when parse data into desired format. Python statsbomb package (Khan, 2019) was used for automatically parse StatsBomb json data to CSV such as pass and shot data. However, it is not supported to extract all event data in sequence. Hence, Python Standard Library ElementTree package is used to manually parse data to get possession data, which will be discussion in the following sections.


```
"id" : "e0320427-8de2-4dfb-accb
"index" : 131,
"period" : 1,
"timestamp" : "00:01:42.253",
"minute" : 1,
"second" : 42,
"type" : {
  "id" : 30,
  "name" : "Pass"
},
"possession" : 3,
"possession_team" : {
  "id" : 768,
  "name" : "England"
},
"play_pattern" : {
  "id" : 1,
  "name" : "Regular Play"
},
"team" : {
  "id" : 768,
  "name" : "England"
},
"player" : {
  "id" : 3244,
  "name" : "John Stones"
},
"position" : {
  "id" : 4,
  "name" : "Center Back"
},
"location" : [ 39.0, 52.0 ],
"duration" : 1.146,
"under_pressure" : true,
"related_events" : [ "56ab4a7b-
"pass" : {
  "recipient" : {
    "id" : 4597,
    "name" : "Fabian Delph"
  },
  "length" : 17.464249,
  "angle" : -1.1583859,
  "height" : {
    "id" : 1,
    "name" : "Ground Pass"
  },
  "end_location" : [ 46.0, 36.0
  "body_part" : {
    "id" : 40,
    "name" : "Right Foot"
```


Figure 7: Example of json data format

Moreover, pitch coordinate is different between Opta data and Statsbomb data, where $x \in [0,100]$ and $y \in [0,100]$ in Opta data but $x \in [0,120]$ and $y \in [0,80]$ in Statsbomb data. Therefore, a normalization process will be performed in order to make axis of coordinate consistent for both dataset before merging these two datasets.

3.3. Action evaluation metric (Objective 2)

3.3.1. Phase 1 Goal probability evaluation

In the first phase, goal probability model is generated in order to assign a relative threat value when given x, y position of football pitch. As mention in previous section, 601 matches from Opta and Statsbomb datasets are used to construct goal probability model. Firstly, shot data are retrieved from dataset into desired format as shown in table 1. Eight attributes are recorded for shot data, which are team, player, min, sec, shot type, body part, x and y. Shot type is split into 4 categories, which are 'Attempt saved', 'Miss', 'Post' and 'Goal'. Body part is consisting of 'Head', 'Left footed', 'Right footed' and 'Other body part'. In summary, there are 14396 shot events recorded in total and 1688 out of them are goal eventually. The next step is to partition football pitch intro several sub sections. 1*1 sub-area is used because x, y coordinate value from Opta data ranged from 0 to 100. Therefore, football pitch is split into 10000 sub areas with each has one-unit-length in both x and y coordinate. Therefore, goal probability can be simply calculated in each sub-area as:

$$\text{Pr\_Goal(sub\_area\_n)} = \frac{Number\ of\ goal\ in\ sub\_area\_n}{Number\ of\ shoots\ in\ sub\_area\_n}$$

In addition, outliers should be removed. Penalty, own goal and other rare case were all treated as outliers which would influence goal probability model. Normally few shot actions performed in very long distance, therefore if one player had a goal in that distance, that sub-area would have relative high goal probability, which is not expected. Thus, a restrict is proposed that each zone must have more than 5 shot attempts, otherwise, goal probability is assigned to 0. This goal probability should be stored and used in this project to assign value to possession sequence.

Table 1: Example of parsed shot data

team	player	min	sec	shot type	body part	x	y
Atlético Madrid	Diego da Silva Costa	4	55	Attempt saved	Right footed	83.41	55.00
Barcelona	Pedro Eliezer Rodríguez Ledesma	17	10	Attempt saved	Right footed	85.66	66.87
Atlético Madrid	João Miranda de Souza Filho	27	23	Miss	Head	97.75	39.12
Atlético Madrid	Diego da Silva Costa	27	24	Miss	Left footed	97.83	56.62

3.3.2. Phase 2 & 3 Possession sequence and sub possession sequence creation

The following step is to get possession sequence data from each match. Possession of the team is defined as one or more sequential actions belonging to the same team and ended by the opposition gaining control of the ball. Sequence is manually defined: start with making a controlled action on the ball (e.g. successful pass) and ended by successful defensive

actions/stoppages (e.g. foul/corner) or shot event. As state in previous section, 190 matches from French Ligue One 2016-2017 season are treated as test set and rest of matches as training set. Therefore, difference data wrangling processes are conducted to deal with test set and training set. In terms of training sets, which includes 64 matches data from World Cup 2018 and 347 matches data from La Liga 2004-2016 seasons, are parsed and transformed into possession sequences. Once possession sequences are created for training set, a relative value is assigned to each possession sequence based on end event of sequence. Not only sequence ended by a goal is valued but any possession ended with a shot attempt because when a shot attempt is performed, plenty of factors could affect goal or not such as defense pressure, ability of shot player and ability of goalkeeper. Therefore, goal probability model is used to assign a value to any possession sequence ended with a shot attempt. If a sequence is not ended with shot attempt, it is assigned value with 0. When all possession sequences from training set are valued, they are stored in a csv file named 'Reference possession.csv' which includes all x values of sequence, all y values of sequence, length (number of events), all player name of sequence and value of sequence. Subsequently, in terms of test set, possession sequences are also created. Moreover, these possession sequences in test set should be split into sub-possession sequence based on player's action performed before and after. For example, player A gives a pass to B and then B gives a pass to C (A→B→C) is a simple sequence, this sequence should be split into two sub sequence based on B's action: one is A→B (sequence before B's action) and the other one is A →B→C (sequence after B's action). The reason why A and C's actions are not evaluated is that there is no sequence before A or no sequence after C. In other words, the influence of possession sequence cannot be detected in that case. Therefore, there are some boundary conditions when sub-sequence is generated: 1. Length of possession sequence need to be larger than two, otherwise sub-sequence before and after action cannot exist simultaneously. 2. Each sub-sequence belongs to the player who performed that action.

3.3.3. Phase 4 & 5 Similarity check & value assignment

After possession sequences in test set are split into sub-possession sequence, the contribution of action still cannot be assessed due to no value assigned to sub-possession sequence yet. Therefore, dynamic time warping technique is proposed to find similar sequence in training set given a sub possession sequence in test set. This process is accomplished by using FastDTW algorithm proposed by Salvador and Chan (2007), which can provide optimal alignments with an O(N) time. Python package fastdtw by Tanida (2017) is used and google could platform is proposed to

run this code. Since possession sequence in training set has already been assigned value, therefore, the influence of action can be evaluated as:

Value of action = Value of sequence after action – Value of sequence before action

Where Value of sequence after action and Value of sequence before action are calculated from average value of 50 most similar possession sequence in training set.

Therefore, the value of action could be either positive or negative. Higher positive value represents more dangerous situation to the opponent after this action is performed. Negative value indicates that action results in a less dangerous situation to the opponent that is not expected.

3.4. Talented young player identification (Objective 3)

Individual player's contribution to his team can be assessed by aggregating all actions corresponding to that player based on results from action evaluation. On account of different players have different quantity of actions, total action value cannot be used directly to compare performance among players. Thus, value per action is computed as a normalization technique. Two filter criteria are considered to use, which are age and total appearance time. For example, the range of age is set as no older than 21 years old in this project in order to find young players. In addition, total appearance time is set as more than 400 minutes, which means that player gained more than 20 minutes per match in average (there are 19 matches per team and 20 teams in total for test set). Thus, the player who meet these filter criteria with higher value per action is supposed to be talent young player. Because the dataset used is at year 2016, player's market value changes from 2016-2019 can be tracked from data provided by Transfermatkt, which is an authoritative organization with public acceptance. If one player' market value has a big increased from 2016 to 2019, that player can be proved as talented player to some extent.

3.5. Team performance evaluation (Objective 4)

Team performance value of each match can be simply aggregating all player's performance value in same team based on results from action evaluation. Then, two lines can be tracked match by match of same team to test their correlation, one is cumulative performance value and the other one is cumulative points gained. In addition, the other comparison test can be set as the ranking based on team's aggregated performance value compared to the position at league table. Several interesting insights can be found from this comparison group, which will be discussed in result section.

3.6. Method for collaboration performance analysis (Objective 5)

In terms of collaboration performance evaluation at team level, successful passing data are used to construct a network of the team from each match. Each player is the node in the network and passing is the edge link between nodes. Winning team and losing team should be treated separately to generate passing network. Meanwhile, six network measures such as number of edges, degree centrality, transitivity, density, eigenvector centrality and betweenness centrality are computed as indicators of collaboration performance. By summarizing the difference of network measures between winning team and losing, optimal network-based indicators related to winning team can be identified. For example, number of edges should be maximized for better collaboration performance in common sense because it is related to number of passes. 190 matches data from French Ligue One 2016-2017 season are used to test this metric in this project because this dataset covers every match from half season of French Ligue One. Therefore, it is more explainable and reliable to only treat this dataset as an entirety since difference leagues have totally difference play styles which may introduce undesired noise to network measures.

4. Results

4.1. Result from action evaluation (objective 2) and use case study

4.1.1. Result of goal probability at different region (Phase 1)

All historical shot data were collected and stored in csv file named ‘ShotData_Merged.csv’. It can be visualized with the aid of Tableau software as shown in figure 8 (Attacking direction is from left to right), which was stored with name ‘Shot Data Plot_20190810 ShotData_all.twb’. Different shot types are differentiated with different colors, and shot type, body part to shot and team are configurated as filter panels to explore. With pitch partitioning of 1*1 sub-region, goal probability is computed at different region as shown in figure 9 (Attacking direction is from left to right). Darker color means goal probability at that sub region is close to 1 and lighter colors represent low goal probability if a shot performed at that sub-region. However, it is obvious that some outliers are in this original goal probability model, which are outside box-area but still with relative high goal probability. The reason has already been put forward in section 3.3.1, rare case happened in this dataset which is few shot actions performed in very long distance but scored luckily as a result. Therefore, that sub-area will has relative high goal probability, which is not expected. Thus, a restrict was proposed that each zone must has more than 5 shot attempts, otherwise, goal probability is assigned to 0. Finalized goal probability data at different region was saved as ‘Goal_Pr_for_plot.csv’ and ‘Goal_Pr_for_calculation.csv’. Figure 10 (Attacking direction is from left to right) gives a clear demonstration that outliers were removed with boundary conditions. In addition, it is also evident that closer to goal line indicates higher probability to score, which is followed common sense in football. Moreover, there is an interesting observation that goal probability distribution is not symmetric, attacking at left edge (upper side of figure 9) has higher goal probability than right edge (lower side of figure 9). This phenomenon can be explained from two aspect. On the on hand, shot data extracted from original dataset may has bias. On the other hand, number of right footed player is more than left footed player normally, therefore, most of players can shot with their stronger foot (right) on the left edge of the football pitch. If at right edge of the pitch, right footed player can shot with right foot but shot angle will be reduced which leads to reduction in goal probability. Nonetheless, this phenomenon is not a major concern of this project so that the generated goal probability model will still be used for the next stage.

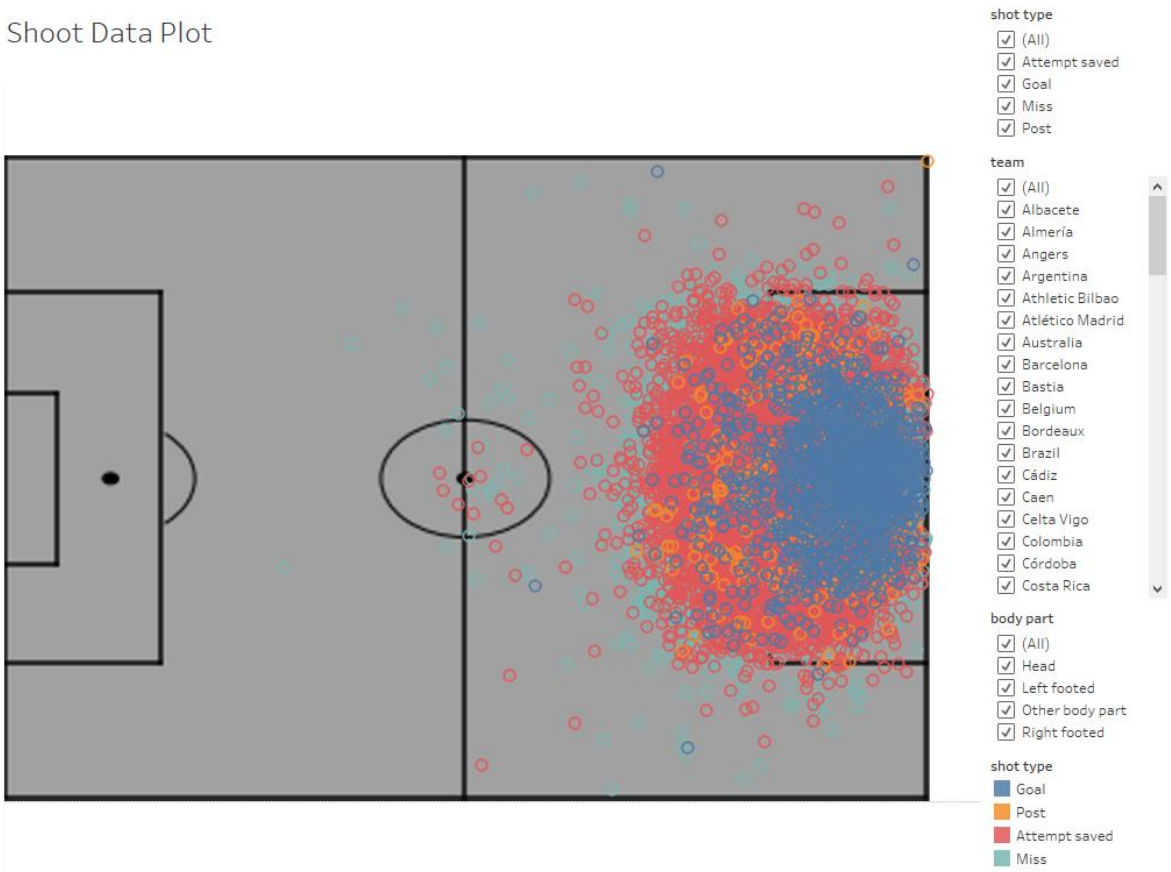


Figure 8: Visualization of shot data (Attacking direction is from left to right).

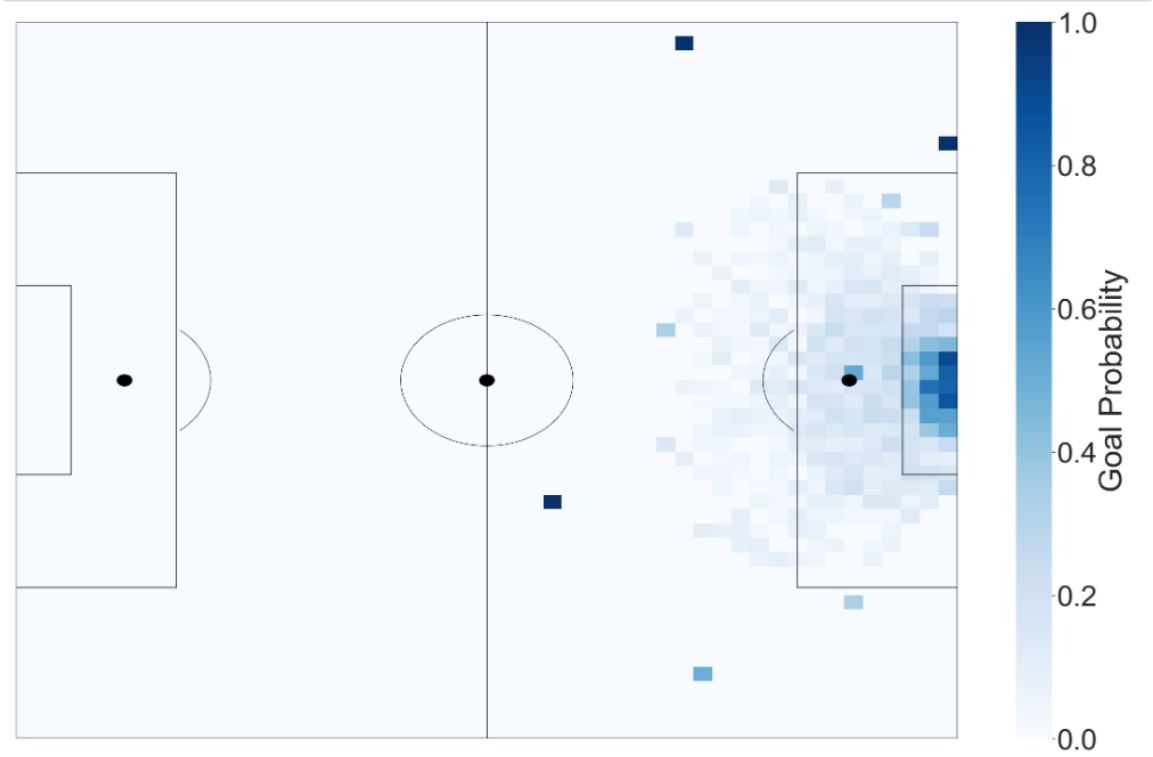


Figure 9: Goal probability with outlier (Attacking direction is from left to right).

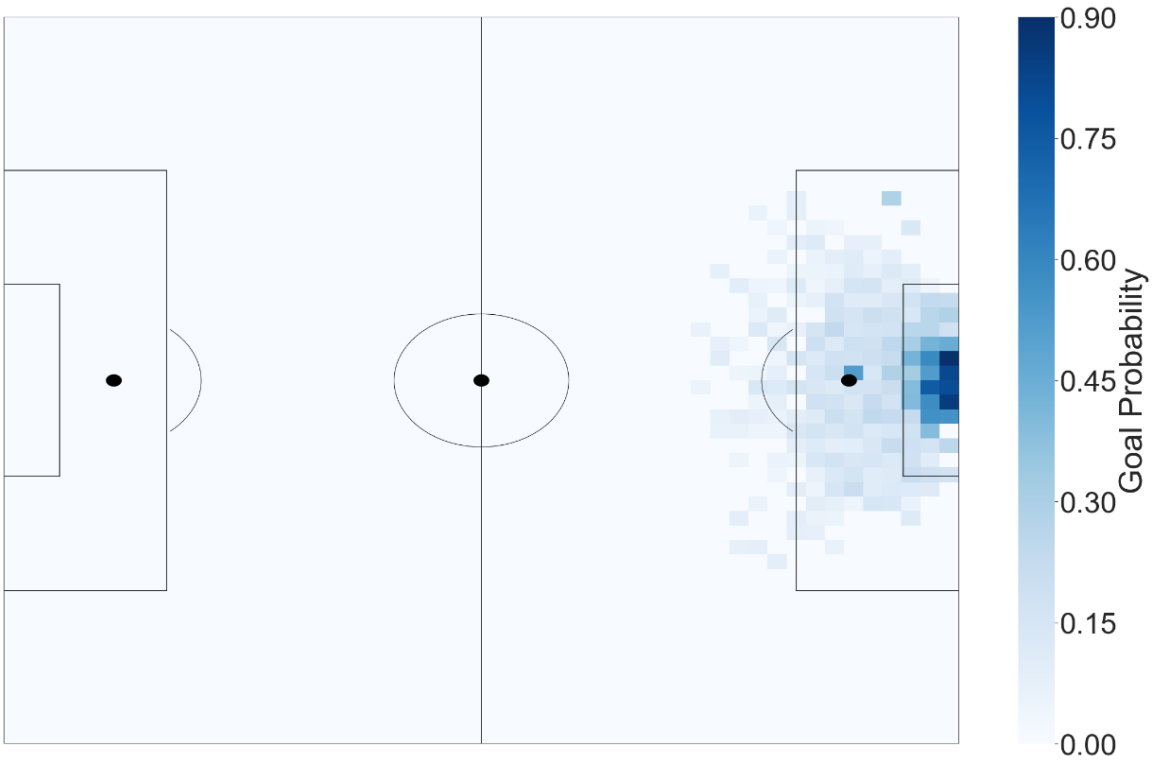


Figure 10: Goal probability without outlier (Attacking direction is from left to right).

4.1.2. Result of possession sequence and sub possession sequence creation (Phase 2 & 3)

With definition of possession sequence as discussed in section 3.3.2 (phase 2), the output data were generated individually match by match from training set (64 matches data from World Cup 2018 and 347 matches data from La Liga 2004-2016 seasons) and stored in a dictionary data structure format: {number of sequence of the match (int): [[X location of each sequence], [Y location of each sequence], [end_event_type] (1 if shot event otherwise is 0), [sequence start time in minute], [sequence start time in second], [sequence end time in minute], [sequence end time in second],[player name by sequence]]. With aid of visualization technique, one example of possession is shown in figure 10. This possession sequence starts at 91'35" of the match and ends at 91'56" of the match by Lyon. Red point is the start location of this possession sequence. Solid line represents pass and dash line indicates dribble event. This possession sequence is ended by shot action performed by Lacazette. Note that the exact dribble route did not recorded in neither dataset provided by Opta or Statsbomb, therefore straight line was used to represent dribble route. After possession sequence of training set was generated, value was assigned to each possession based on rule stated in section 3.3.2, which is if sequence is ended with a shot attempt, goal probability model was used to assign a value otherwise the value would be 0. For example, if end event is shot attempt with coordinate of (x=78.5, y=50.6), this possession sequence will be value as 0.053 based on goal probability. Still taking figure 11 as an example, goal probability at end position (x=93.3, y=51.3) is 0. 287. Thus, this sequence has value of 0.287, which means it is a

sequence with a relative high attacking threat to the opponent. If Lacazette did not attempt to shot, this sequence would still be valued as zero although it was very closer to goal line. All possession sequences in training set were merged, valued and stored with the same file named 'Reference_possession.csv' for later retrieval. In addition, there were 75236 possession sequences stored as reference possession. First 5 rows of reference possession are shown in table 2, X, Y coordinate, length of possession sequence (length here is defined as number of events performed), player name flow in sequence and value are recorded, where values in column of 'X', 'Y' and 'player_name' are folded due to space limit.

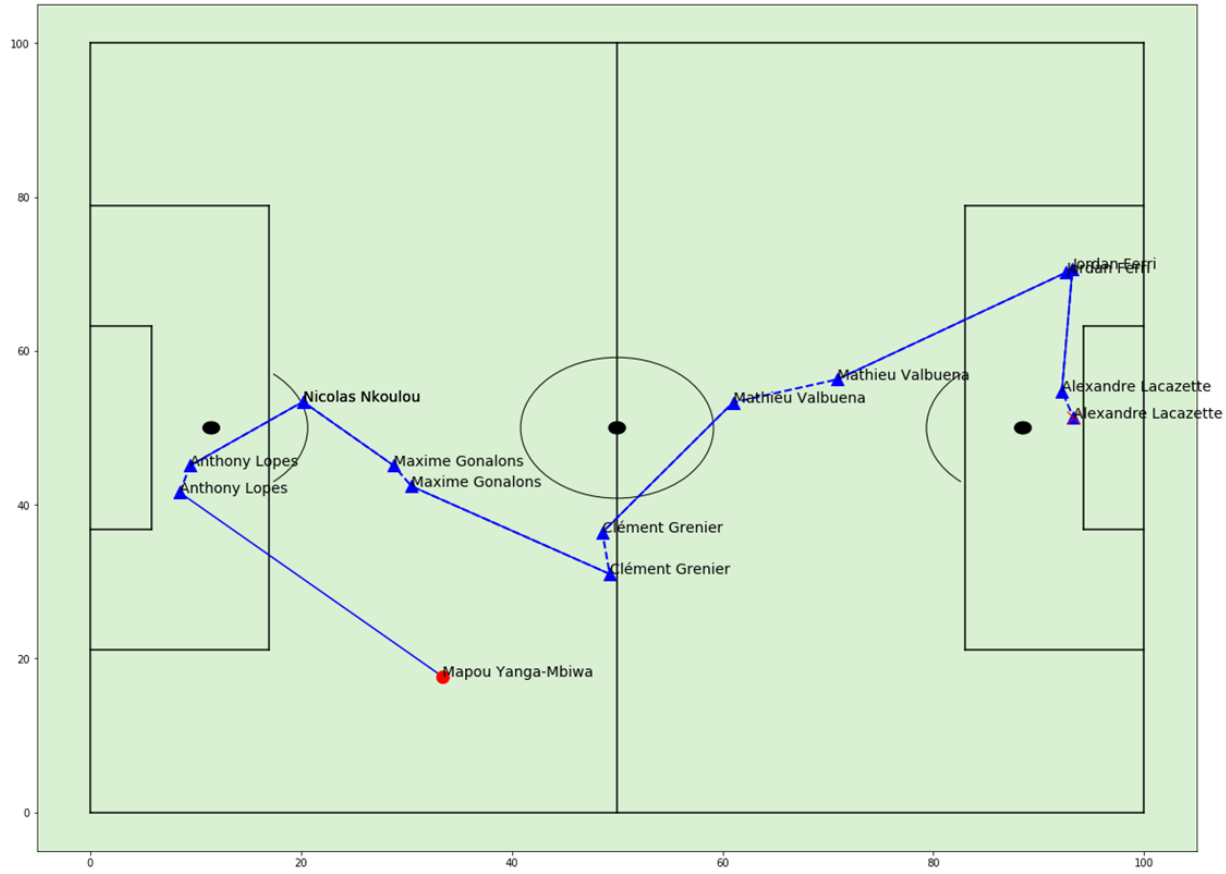


Figure 11: Possession sequence example

Table 2: Head rows of reference possession file

	x	y	length	player_name	value
0	3336, 17.75, 5.160	0, 10.25, 4.75, 41.0, 44.5	5	én', 'Juan Francisco Torres	0
1	33333333336, 74	[0.0, 7.875, 7.875]	3	o Torres Belén', 'Arda Turan	0
2	416666666666678	.75, 87.25, 87.25, 95.1	19	Villa Sánchez', 'David Villa	0
3	4.6666666666666	[26.25, 5.0, 30.25]	3	Silva Costa', 'Arda Turan', 'A	0
4	1.5, 86.916666660	00000000014, 6.25, 60.	5	lio', 'Jorge Resurrección Me	0

In terms of test set (190 matches from French Ligue One 2016-2017 season), possession sequences were extracted and stored with same data format as training set. Then, each possession sequence was split into sub-possession sequence based on player's action and re-assigned to player who performed that action. It was also stored in a dictionary data structure format: {player name (string): {No. of action (int): [x location of sub-sequence before action], [y location of sub-sequence before action], [x location of sub-sequence after action], [y location of sub-sequence after action]}}. Figure 12 and figure 13 give an example of sub-sequence before action and after action separately. The action in this example is highlighted in figure 13, which is a crossing pass made by Corentin Tolisso from right edge of the pitch towards penalty area. Strictly speaking, sub-possession sequence of figure 12 is part of sub-possession sequence of figure 13. Therefore, this crossing action belongs to player Corentin Tolisso and these two sub-possession sequences were attached to this action.

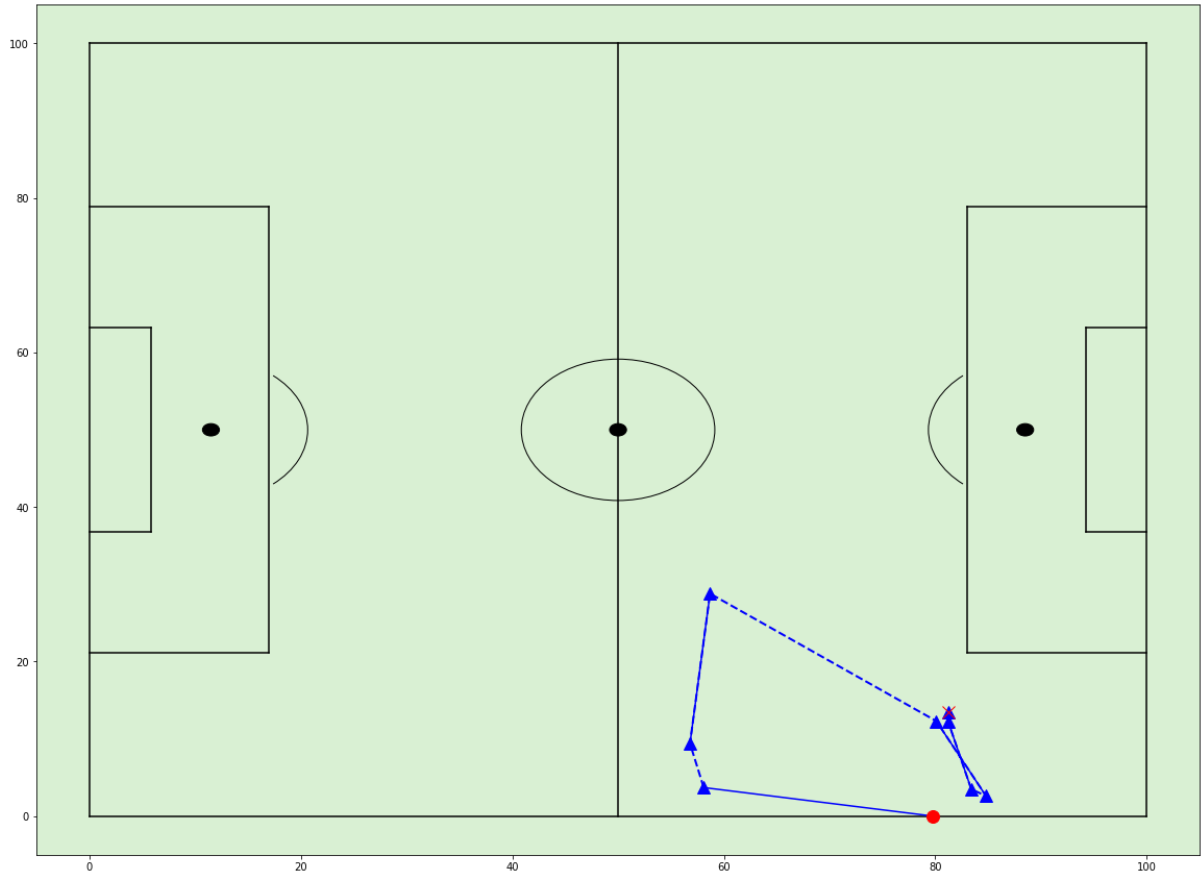


Figure 12: Sub-possession sequence before a crossing pass

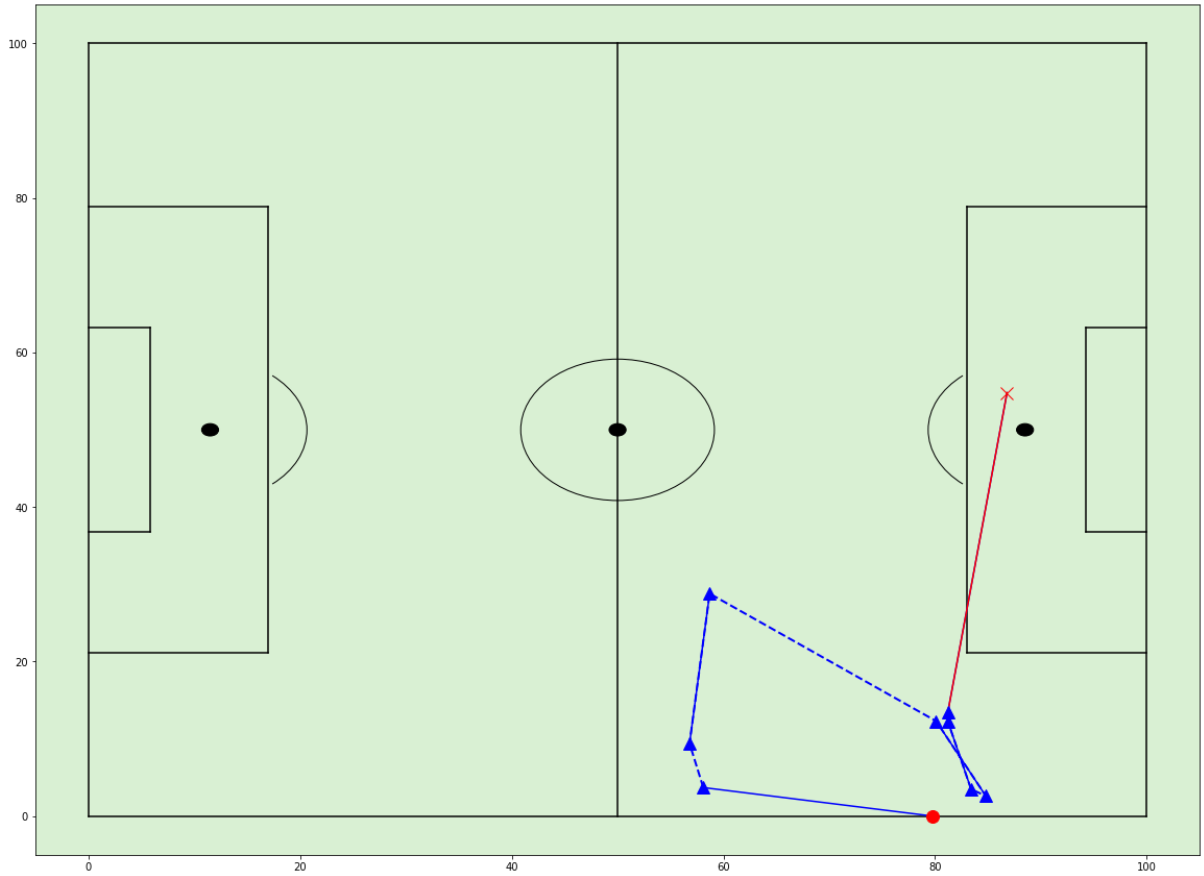


Figure 13: Sub-possession sequence after a crossing pass

4.1.3. Result of similarity check & value assignment (Phase 4 & 5)

Since possession sequences were split and assigned to related player, the action value can be evaluated by value of sub-possession sequence after action subtract sub-possession sequence before action. The value of sub-possession sequence was calculated from average value of 50 most similar possession sequence in training set. When given a sub-possession sequence in test set, similar possession sequences were located by dynamic time warping technique from training set. Figure 14 and figure 15 provide one example for similarity check by dynamic time warping technique. Same target sub-possession sequence before and after action from test set was waited to be valued, which highlighted in blue line in both figure 14 and 15. Target sub-possession sequences are same as example from last section (figure 12 and figure 13) separately. In terms of sub-possession sequence before crossing action, similarity between target sub-possession sequence and sequence in training set was computed one by one. Then top 50 similar sequence in training set were selected. Figure 13 shows top two of the similar sequence in red line and mauve line compared to blue line (target sub-sequence). These three trajectories were all initiated near right sideline of the pitch and ended in similar zone with similar movement curve, which is evidence to say that they are similar trajectories. Therefore, the value of sub-possession sequence before crossing pass in this case is equal to 0.0018, which is the average value of top 50 similar possession sequence's value in training set. In terms of sub-possession sequence after crossing

action, top two of the similar sequence in red line and mauve line compared to blue line (target sub-sequence) are shown in figure 15. It is convinced that these three sequences are highly similar because either start or end location is closed to each other and especially, crossing pass from right wing to penalty area is all performance in all sequences. By averaging value of top 50 similar possession sequence in training set, this sub-possession sequence after crossing pass event in this case equals to 0.0742. Consequently, value of this crossing action is valued as 0.0742 - 0.0018 = 0.0724. It is reasonable that this crossing action was assigned to a positive value because this action made ball moved to a more dangerous area to score. Finally, player's performance was evaluated match by match and stored with game id, team name and player name for later retrieve. Table 3 demonstrates an output from match id 85139 (French Ligue 1 Season 2016/2017, Nancy vs Lyon, 2016-08-14) and Lyon's player Alexandre Lacazette. There are 26 actions in total were evaluated in this match. It is a relative low value of number of actions on account of Alexandre Lacazette is a striker in nature who is not necessary to control the ball all the time during the match compared to midfielders. In addition, total contribution of Alexandre Lacazette for this match is 0.05238 if sum all values of action, which indicates that Alexandre Lacazette played an importance role in Lyon's attacking and had a positive effect for Lyon to win this match.

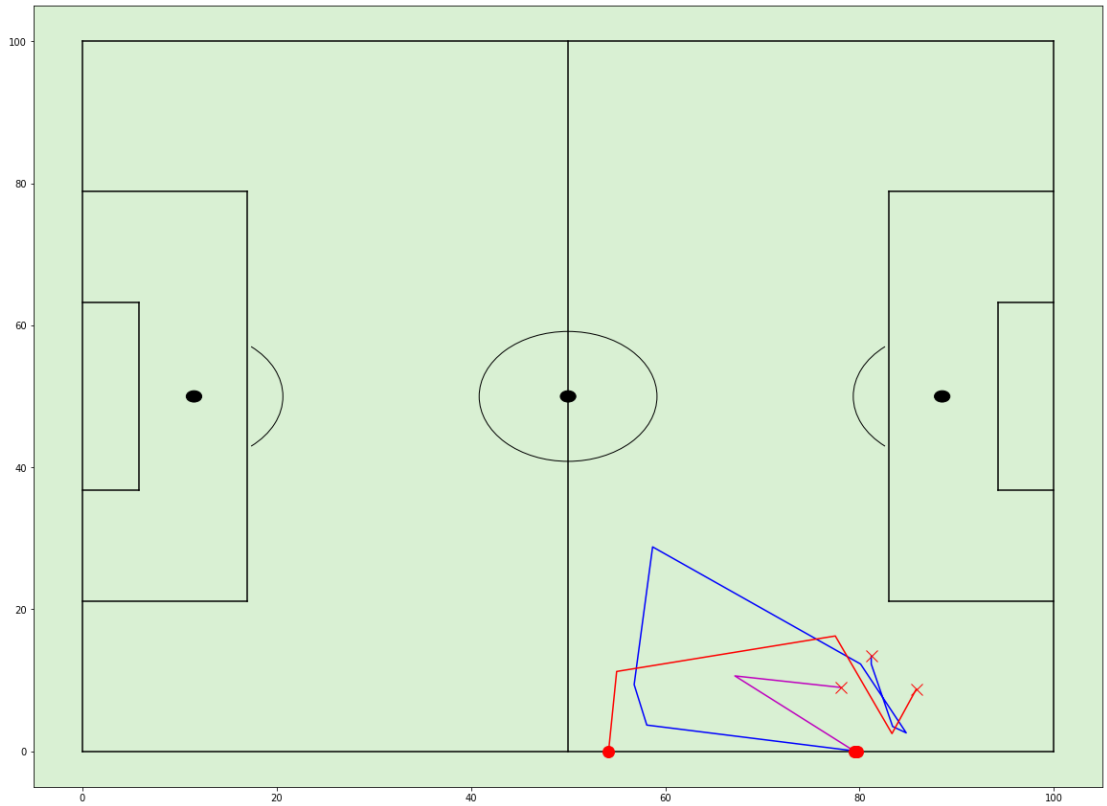


Figure 14: Two similar sequences (red and mauve) identified given sub-possession sequence before crossing action (blue) by dynamic time warping similarity check

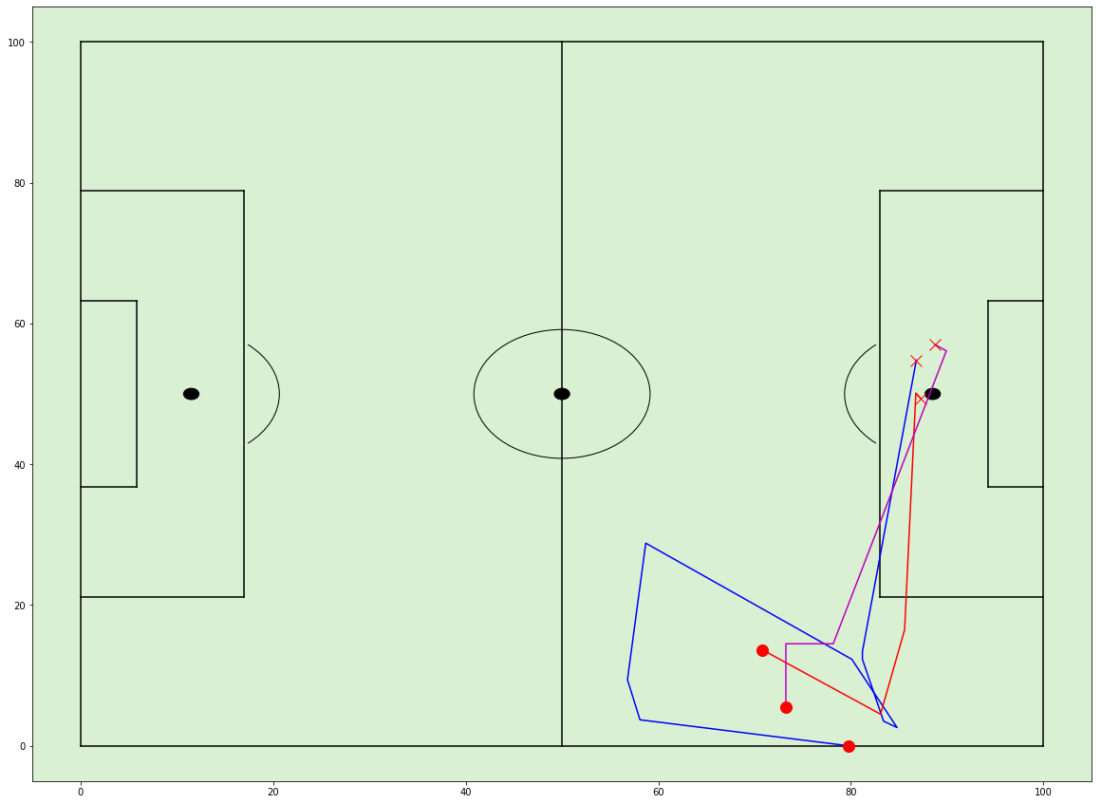


Figure 15: Two similar sequences (red and mauve) identified given sub-possession sequence after crossing action (blue) by dynamic time warping similarity check

Table 3: Alexandre Lacazette's performance in match No. 85139

Action No.	contribution		Action No.	contribution
0	0.00000		13	0.00960
1	0.00000		14	-0.00167
2	0.00000		15	0.00182
3	0.00000		16	-0.00372
4	0.00000		17	0.00681
5	0.00000		18	0.00000
6	-0.02487		19	0.00615
7	0.00000		20	-0.00121
8	0.00000		21	0.00000
9	0.00914		22	0.02747
10	0.00000		23	-0.00280
11	0.00000		24	0.00378
12	0.00000		25	0.02187

4.1.4. Result of identification of young talent player (Objective 3)

When every action was evaluated, individual player's contribution can be simply assessed by aggregating all value of actions. Then, individual player's performance was gathered and merged in same file named 'individual performance all in one.csv'. It not only contains the aggregated value of contribution, but also includes amount (how many actions of one player played during first half season of French Ligue one 2016-2017), minutes played (how many minutes player was attended), age (in order to filter young players), value per action (calculated by value / amount) and value per minute (computed by value / minutes played) as shown in Table 4. When this file was imported in Tableau software, talented young player can be identified with some criteria as no older than 21 years old and played more than 400 minutes as shown in figure 16. Horizontal line represents value per action, higher value is expected for better performance. Vertical line represents actions per 90 minutes, which can indicate player's playing style or position of the team such as striker or midfielder. Higher value in actions per 90 minutes is normally corresponding to a midfielder who plays an important role in connecting teammates. In addition, younger player is highlighted with darker color and player's age close to 21 would be relative light color. Therefore, if a player is in dark color and located in the right position on figure 16, this player is marked as young talent player with relatively good performance in French Ligue One 2016-2017 half season. Although test dataset is from the first half of 2016-2017 season not the latest season, it can be assumed that player's performance was evaluated without knowing player's market value at present. Then identified outperformed young player's market value changes would be tracked from 2016 to 2019 from Transfermarkt data to see whether these young players highlighted by the proposed metric are valuable to invest or not. Top eight players with highest value per action are circled in figure 16. Their market value price changes from 2016-2019 are also tracked and summarized in figure 17. It is obvious that market value of Kylian Mbappe increased 50 times at present (228 million euros) compared to price at 2016 (4.56 million euros) and Kylian Mbappe is identified as the young player with best performance by the proposed performance evaluation metric. Not only for Kylian Mbappe, market value of four players (Mariusz Stepinski / Almamy Touré / Malcom / Allan Saint-Maximin) out of 8 are also at least two times at present than year of 2016. Matkrt value of Enzo Crivelli and Valère Germain also raised but not as much as the other one. Only 1 player (Clinton N'Jie) out of 8, who's market value was reduced at 2019 compared to 2016. However, it is notable that his market value was increased from 2016 to 2018 but dropped from 2018 to 2019 due to his injuries. Therefore, it can be concluded that young talent players found by this proposed metric are truly investable unless uncertain circumstances such as injury. If a club could bring one highlighted young talent club by this metric, the market value of that player would be doubled with high probability. Therefore,

club manager and football scout would be potential beneficiaries because this evaluation metric can help them to make decision whether to purchase a young player or not.

Table 4: Head rows of merged individual performance

	Player	value	amount	Minutes played	Age	Value per action	Value per minute
0	Alexandre Lacazette	1.327	466	1297.9	25	0.00285	0.001022
1	Anthony Lopes	-0.056	90	1620.0	26	-0.00062	-0.000034
2	Clément Grenier	0.159	48	34.5	23	0.00331	0.004611
3	Corentin Tolisso	1.297	860	1499.6	22	0.00151	0.000865
4	Jordan Ferri	0.023	327	661.3	23	0.00007	0.000035

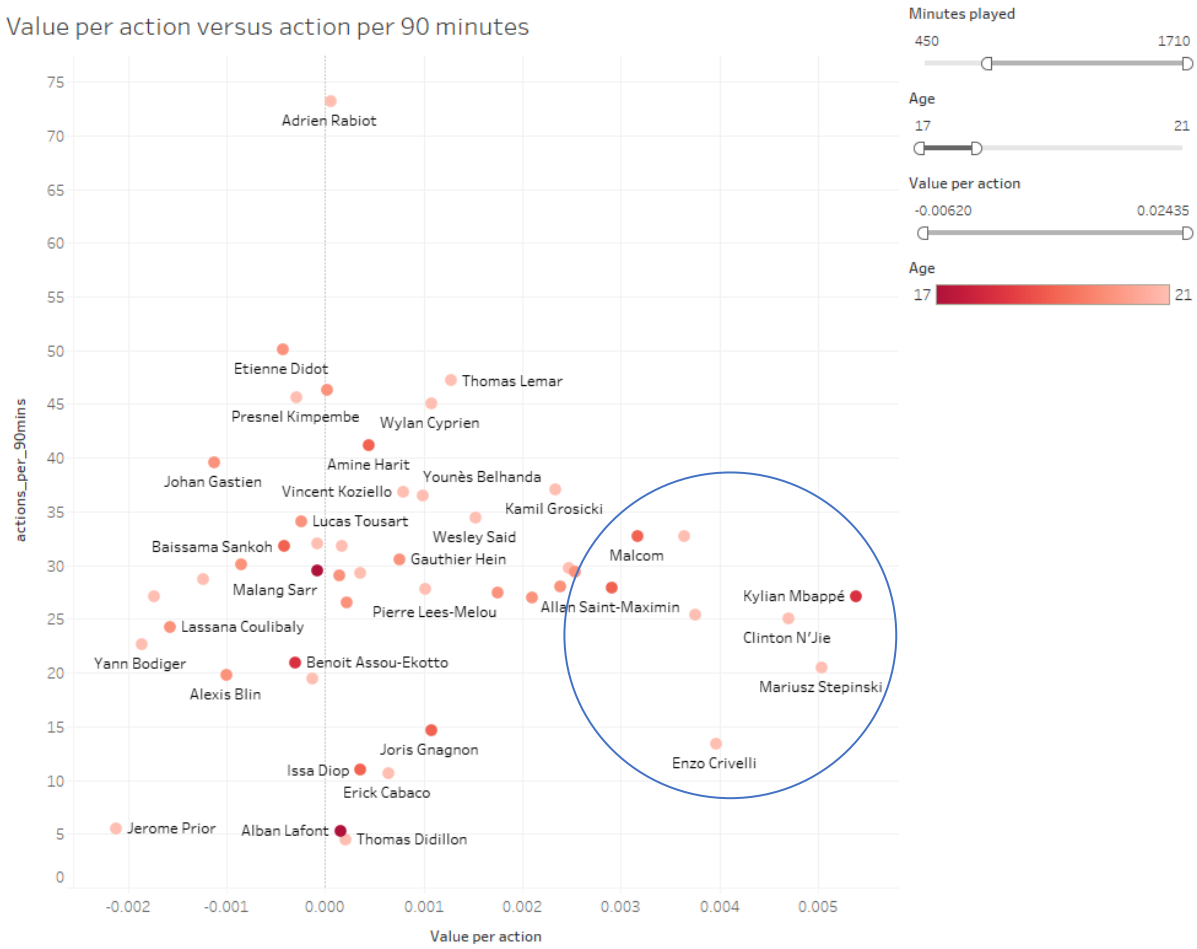


Figure 16: Value per action versus action per 90 minutes with filters

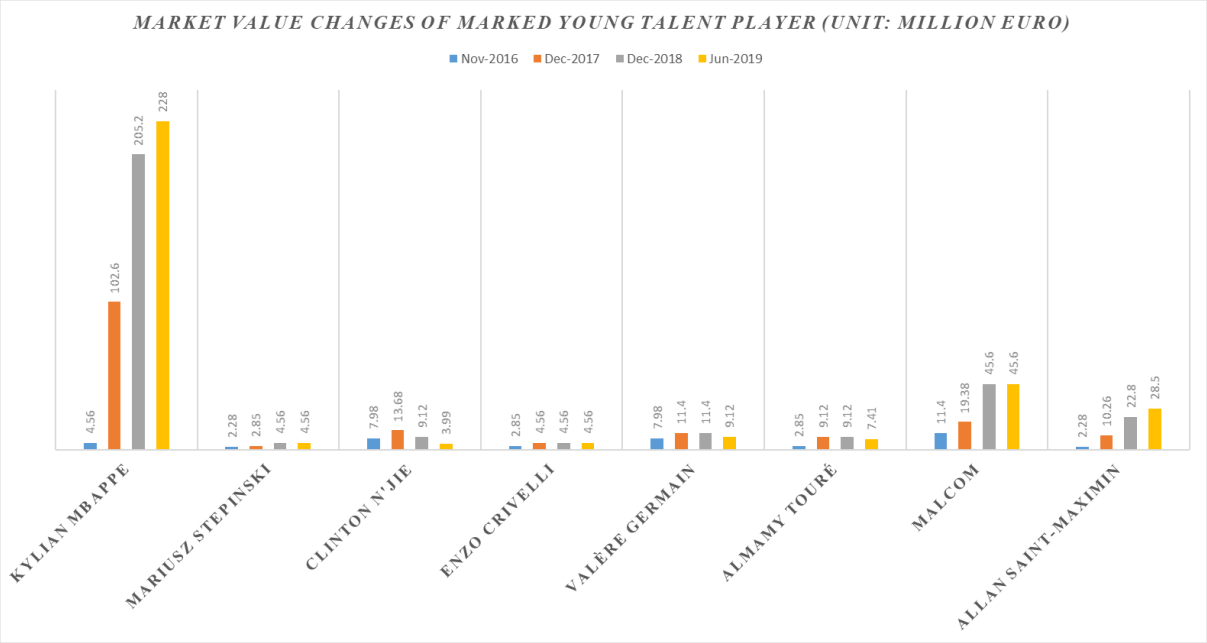


Figure 17: Market value changes from 2016 to 2019 of marked young talent player (unit: million euro)

4.1.5. Result of team performance evaluation (Objective 4)

Other than identification of young talent player, the computed action evaluation metric can be also used to assess team performance by aggregating all players' performance value in same team. On account of dataset is from French Ligue One 2016-2017 half season, league table was found at week 19 of that season as shown in figure 18. It can be concluded that at week 19, OGC Nice was at first place with 44 points and followed by AS Monaco with 42 points. Paris Saint-Germain and Olympique Lyonnais were at third and fourth place separately. These four teams gained similar points at half season while the teams at fifth and sixth place only got 30 points, 7 points less than Olympique Lyonnais. Thus, the top four teams are selected to make comparison. Team performance value and points gained match by match of top 4 teams were plotted to see its trend as shown in figure 19 (left upper: Nice, left bottom: Paris Saint-Germain, right upper: Monaco, right bottom: Lyon), where red line represent cumulative points gained match by match and blue line indicates cumulative performance value of the team match by match. Left vertical axis is value for cumulative performance and right vertical axis is value for points gained. Pearson correlation value of four pairs of growth curves are 0.993, 0.990, 0.991, and 0.967 separately, which imply that the growth curves of cumulative performance value and cumulative points gained are highly correlated. In other words, higher positive team performance value is corresponding to win (3 points) and lower positive or negative performance value are corresponding to lose or draw (0 or 1 points). In addition, blue dotted line is a reference line with same performance value of team Nice because Nice was at first place at week 19. However, it is obvious that cumulative performance value of AS Monaco and Paris Saint-Germain are both

much higher than Nice while Olympique Lyonnais has similar performance value. This phenomenon can be explained by two sides. On the one hand, AS Monaco and Paris Saint-Germain can create more chances in attacking than Nice, however, they were not lucky enough to grasp these opportunities to win the game. On the other hand, Nice might did a greater job on defense than the other two teams so that Nice can win the match even if did not create much score opportunities. This inference is proved by the league table at the end of 2016-2017 season as shown in figure 20. AS Monaco and Paris Saint-Germain got the first and second places in the end of that season while Nice only dropped to third place. Another example could be Lille (LOSC), Nancy, Bastia and Dijon had similar points at middle position (13, 14, 15, 16 separately) on league table at week 19. Their performance value growth curve and points gained growth curve are also plotted in figure 21. It can be concluded from figure 20 that Nancy and Bastia had worse performance value than Lille and Dijon which may indicate that Nancy and Bastia may get into trouble in the second half season due to their bad performance in attacking. Final position of league table at week 38 (figure 20) support this inference. Nancy and Bastia were the worst two team in French Ligue One of 2016-2017 season while Lille (LOSC) and Dijon maintained their position on league table. Therefore, if similar case happens that one team has much higher cumulative performance value but gets less points then the other, some advice for team coach and players could be: 1. Be patient and maintain the performance in attacking. 2. Strikers should focus more on shot training because they got lots of opportunities but did not grasp. 3. Defense ability should be focus more than attacking in this circumstance. On the contrary, if one team gets relative low value in performance evaluation but got more points, it may indicate that this team still need to focus on how to improve their performance to create more opportunities otherwise this team may get into trouble in winning a match in the future.

	Team	Pld	W	D	L	F	A	GD	Pts
1	OGC Nice	19	13	5	1	34	13	+21	44
2	AS Monaco	19	13	3	3	56	20	+36	42
3	Paris Saint-Germain	19	12	3	4	38	15	+23	39
4	Olympique Lyonnais	19	12	1	6	37	19	+18	37
5	EA Guingamp	19	8	6	5	25	19	+6	30
6	Olympique de Marseille	19	8	6	5	22	19	+3	30
7	Stade Rennais FC	19	8	4	7	20	23	-3	28
8	AS Saint-Etienne	19	6	8	5	18	16	+2	26
9	Toulouse FC	19	7	5	7	22	21	+1	26
10	Girondins de Bordeaux	19	6	7	6	20	26	-6	25
11	Montpellier Hérault SC	19	5	7	7	28	31	-3	22
12	FC Nantes	19	6	4	9	13	26	-13	22
13	LOSC	19	6	3	10	18	25	-7	21
14	AS Nancy Lorraine	19	5	6	8	15	23	-8	21
15	SC Bastia	19	5	5	9	17	23	-6	20
16	Dijon FCO	19	4	7	8	26	29	-3	19
17	Angers SCO	19	5	4	10	15	24	-9	19
18	FC Metz	19	5	4	10	18	39	-21	19
19	SM Caen	19	5	3	11	20	33	-13	18
20	FC Lorient	19	4	3	12	20	38	-18	15

Figure 18: League table of French League One 2016-2017 season at week 19 (Ligue 1 Conforama, 2019).

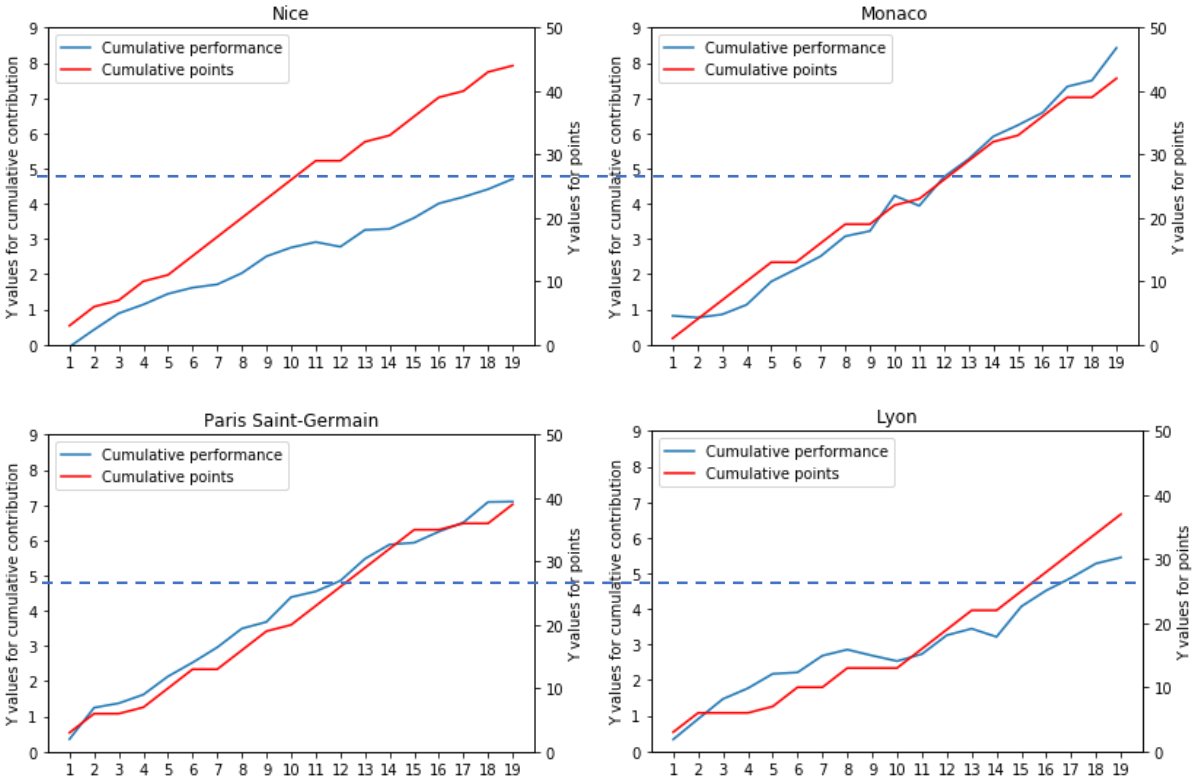


Figure 19: Team performance value and points gained match by match of top 4 teams until week 19 (left upper: Nice, left bottom: Paris Saint-Germain, right upper: Monaco, right bottom: Lyon)

	Team	Pld	W	D	L	F	A	GD	Pts
1	AS Monaco	38	30	5	3	107	31	+76	95
2	Paris Saint-Germain	38	27	6	5	83	27	+56	87
3	OGC Nice	38	22	12	4	63	36	+27	78
4	Olympique Lyonnais	38	21	4	13	77	48	+29	67
5	Olympique de Marseille	38	17	11	10	57	41	+16	62
6	Girondins de Bordeaux	38	15	14	9	53	43	+10	59
7	FC Nantes	38	14	9	15	40	54	-14	51
8	AS Saint-Etienne	38	12	14	12	41	42	-1	50
9	Stade Rennais FC	38	12	14	12	36	42	-6	50
10	EA Guingamp	38	14	8	16	46	53	-7	50
11	LOSC	38	13	7	18	40	47	-7	46
12	Angers SCO	38	13	7	18	40	49	-9	46
13	Toulouse FC	38	10	14	14	37	41	-4	44
14	FC Metz	38	11	10	17	39	72	-33	43
15	Montpellier Hérault SC	38	10	9	19	48	66	-18	39
16	Dijon FCO	38	8	13	17	46	58	-12	37
17	SM Caen	38	10	7	21	36	65	-29	37
18	FC Lorient	38	10	6	22	44	70	-26	36
19	AS Nancy Lorraine	38	9	8	21	29	52	-23	35
20	SC Bastia	38	8	10	20	29	54	-25	34

Figure 20: League table of French League One 2016-2017 season at week 38 (Ligue 1 Conforama, 2019).

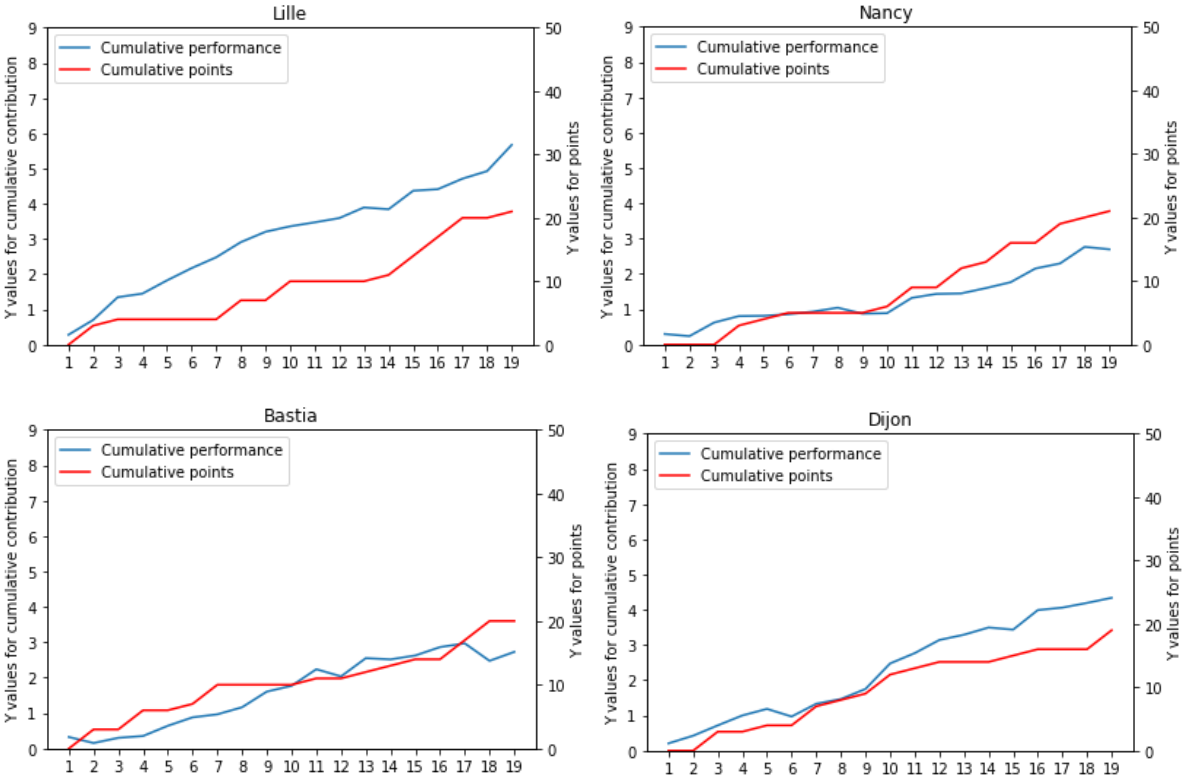


Figure 21: Team performance value and points gained match by match of 4 selected teams until week 19 (left upper: Lille (LOSC), left bottom: Bastia, right upper: Nnacy, right bottom: Dijon)

4.2. Result of collaboration performance by network analysis

Successful passing data were used to construct a network of the team from each match. Each player is the node in the network and passing is the edge links between nodes as shown in figure 22. Node size is differentiated by betweenness centrality, where bigger node indicates a higher betweenness centrality and edge width is ranked by weight. In addition, node position is the average x, y value where the player performed an action. Once the network can be generated, network measures can be computed in order to identify indicator of winning team among number of edges, degree centrality, transitivity, density, eigenvector centrality and betweenness centrality. Only 190 matches data from French Ligue One 2016-2017 season were used because this dataset covers every match from half season of French Ligue One. In addition, matches with 2 or more goal difference (e.g. 2-0, 3-1) can be counted in order to avoid winning with fluke. Moreover, noises are removed by filtering edges in weight less than 3. Therefore, only 67 matches can meet these filter criteria out of 190 matches in total, which were stored in csv file named 'losing_team_indicator_remove_2.csv' and 'wining_team_indicator_remove_2.csv'. A radar plot of each indicator from network analysis of winning (blue line) compared to losing team (orange) during France Ligue One 2016-2017 half season is shown in figure 23, where betweenness centrality was enlarged by 10 times and number of edges was scaled down by 100 times for axis alignment. It is obvious that winning teams have relative larger value in transitivity, degree centrality but lower in betweenness centrality than losing team. An independent samples t-test was conducted to compare the network measures for winning team and losing team statistically, which is summarized in table 5. Two-tailed p values were computed with α value equals 0.05. There is significant difference ($p < 0.05$) for betweenness centrality, degree centrality, transitivity and density while eigenvector centrality and numbers of edges are not significant difference. It reveals that betweenness centrality should be minimized ideally, degree centrality, transitivity and density should be maximized for the sake of winning. Moreover, eigenvector centrality and numbers of edges can be neutral or enlarge to increase team performance. Actual meaning of difference network measures will be discussed in detail in next chapter.

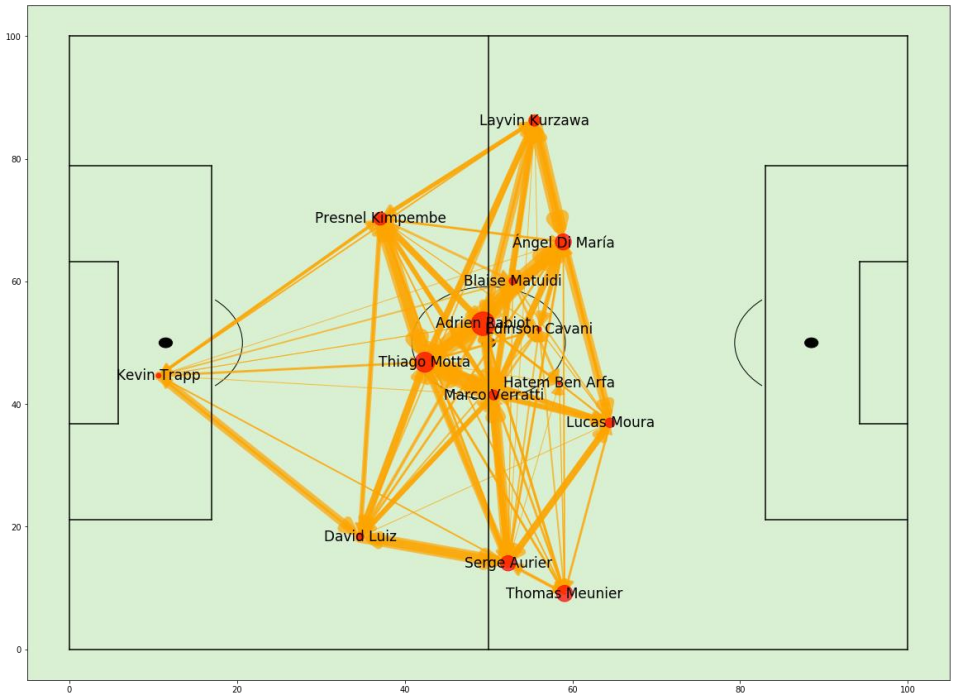


Figure 22: Example of passing network

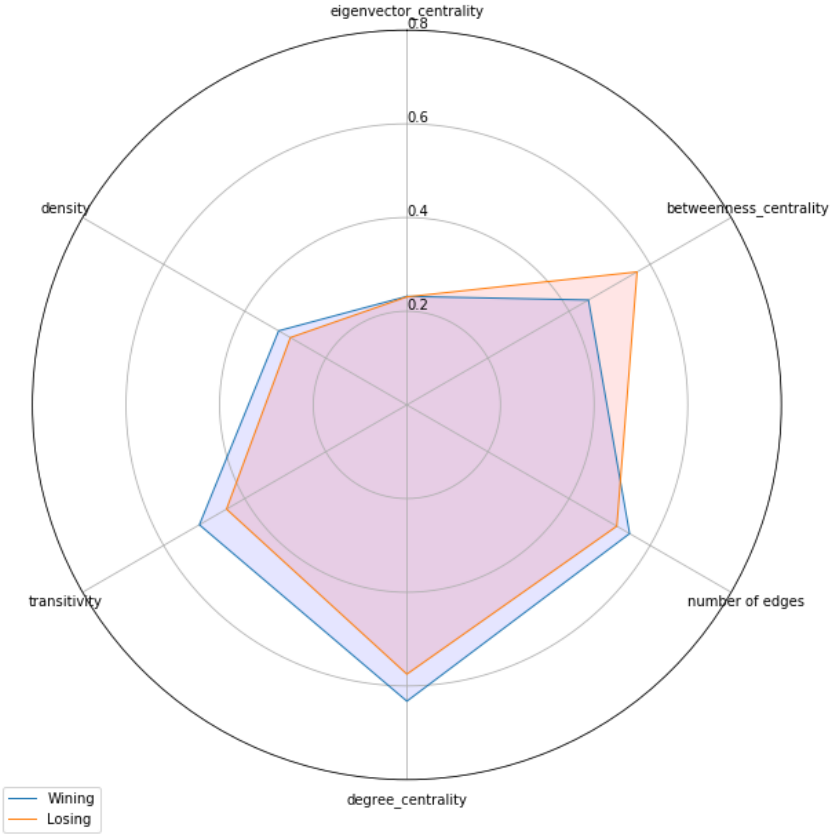


Figure 23: Radar plot of the network measures for winning team and losing team

Table 5: Independent samples t-test result

	Eigenvector centrality	Betweenness centrality	number of edges	degree centrality	transitivity	density
p value	6.7E-01	2.5E-05	1.7E-01	2.7E-02	4.3E-04	2.7E-02

5. Discussion

5.1. Result compared with objectives

Objective 1: *'To understand the source data well provided by Opta Sport company and Statsbomb company in order to make use of these football event-based data'*: Both datasets are used efficiently in this project. Documents provided by Opta and Statsbomb demonstrate their data format detailly. Nearly all event cases were used including pass, shot, dribble and foul event. Both datasets record ball location in x,y coordinate however, different coordinate systems are used by two company. Therefore, transformation is needed before merge two datasets.

Objective 2: *'To develop a metric to evaluate players' performance which accounts all types of on-ball actions (such as pass event and dribble event) and considers contextual impact of each action'*: In literature review, several methods for action evaluation were reviewed such as traditional machine learning technique, Markov model approach and trajectory analysis. As a result, a new metric to evaluate player's performance approach was proposed as followed by goal probability model creation, possession sequence creation and be value based on goal probability model, sub possession sequence creation (including passing and dribble event), dynamic time warping technique used to find similar sequence in training set and compute value of each action (passing and dribble) based on the value of sub sequence after and before action.

Objective 3: *'To develop a data-driven technique to identify talented young players whose value in transfer market has a potential to grow rapidly over years based on the results from the proposed metric of performance evaluation'*: Section 4.1.4 presents the results of identified young player by proposed metric. It was accomplished by aggregating all value of actions for the same player and value per action (calculated by value / amount) was tracked to find player with best performance during first half season of French Ligue One 2016-2017. Other filter criteria include age (no older than 21) and attendance time (played more than 400 minutes). Kylian Mbappe was pointed out as the young player (18 years old at that season) with best performance. His market price was just 4.56 million euro at that season, however, Kylian Mbappe is one of the most expensive players in the world at present (228 million euro). Other top 8 young talent players filtered by this proposed metric such as Almamy Touré, Malcom and Allan Saint-Maximin, most of their market price was double from 2016 to 2019, which makes this evaluation metric more convincing.

Objective 4: *'To develop a data-driven technique to compare team's performance with its position on league table based on the results from the proposed metric of performance evaluation'*: This objective was completed by aggregating all players' performance value in same team based on the computed performance contribution of actions. Team with largest value in team contribution is Monaco by the end of week 19 2016-2017 season but Monaco was behind

Nice at the league table. Nonetheless, Monaco got final league championship at that season, which can support this evaluation metric to some extent. Other cases represented in section 4.1.5 also indicate that this evaluation metric cannot identify the exact team's league position, but it can provide a new way to judge team's performance which can give some insights of team's performance in the future.

Objective 5: *'To develop a metric which can evaluate the collaborations between teammates by considering the whole team as a network. To get some insight based on the characteristics of network-measures for winning team compared to losing team'*: Successful passing data is used to construct a network of the team from each match. Each player is the node in the network and passing is the edge links between nodes. Two more boundary condition were set in order to reduce noise of network. One is that edges less than 3 in weight were removed. The other one is that only matches with 2 or more goal difference (e.g. 2-0, 3-1) were counted. Therefore, only 67 matches can meet these filter criteria out of 190 matches from dataset. Once passing network was generated, network measures of winning team and losing team can be computed by using NetworkX Python library separately. Finally, significant difference ($p < 0.05$) was found for betweenness centrality, degree centrality, transitivity and density by statistical t test on network measures of winning team and losing team. Therefore, it is found that some network measures should be optimized such as minimize betweenness centrality, maximize degree centrality, maximize transitivity and maximize density in order to improve collaboration between teammate to win the match.

5.2. Result compared with existing literatures

The results not only need to be compared with objectives, but also need to be compared with existing literatures. Discussion in this part will followed the same structure of result chapter.

- **Goal probability at different region**: It is the simplest way to compute goal probability by only considering shot position other than lots of features (such as context, body part to shot and shot pressure) used for machine learning modelling (logistic regression, decision tree, random forest and Ada Boost) because the main objective of this project is not for predicting expected goal given context condition. Goal probability model generated in this project is theoretical correct, with position closer to goal line with higher goal probability, which is similar to the result by Decroos *et al* (2017).
- **Possession sequence and sub-possession sequence creation**: Possession sequence was also created by Decroos *et al* (2017) and Bransen & Van Haaren (2017) but dribbling events were not included in their result. Both of their researches only evaluated passing event so that the identified high value players by their metric were normally midfielder players who make

plenty of passing just in one match. Differ from their metrics, sub-possession sequence creation in this project includes both passing and dribbling event so that identified high value players by this proposed metric contains both strikers and midfielders.

- **Similarity check & value assignment**: Similar sequence identification was achieved by using dynamic time warping technique in this project. Section 4.1.3 illustrates that result of similar sequences are very close to test sub-sequence with visualization technique, they were all initiated and ended at similar areas. In addition, they all maintain unique characteristic of sequence such as long pass and crossing pass. Decroos *et al* (2017) examined three method for similarity measures in football which are dynamic time warping, Frechet distance measure and the longest common subsequence distance measure. The conclusion put forward by that research was that dynamic time warping and Frechet distance measure are both efficient to find similar possession sequence in football, but the longest common subsequence distance measure did not work well for this job. Results from this project also agree that conclusion.
- **Identification of young talent player**: As mention in previous part, the identified high value players by this proposed metric contains both strikers and midfielders while other metrics only focus on one of them. Meanwhile, the identified young talent player by this metric is compared to its market value changes from 2016 to 2019 because dataset is the French Ligue One first half season. Player's market value estimated by Transfermarkt was proposed to use because it is an authoritative organization with public acceptance. Data from Transfermarkt proves that investable young talent player can be identified by this proposed metric.
- **Team performance evaluation**: There is no previous work found which using action evaluation value from individual player to get some insight for the whole team performance. One similar work was conducted by Eggels (2016). Expected goal model was created to compare team's actual goal during the season in his research. Juventus in 2015-16 season had a bad opening, that they only won three matches of first 10 games, however, expected goal showed that Juventus' s expected goal value was much higher than the actual goal scored in first 10 games. It indicated that Juventus was dominating matches and creating lots of opportunities but did not seize the goal opportunity. Same inference logic was used in this project, if a team has relative high value in team performance value while its actual points gained is less than other teams. It may indicate that playing style or tactic in attacking were not the actual problem. Team should have patience and focus more on how to improve shot quality or defense ability. Opposing explanation works vice versa.
- **Collaboration performance by network analysis**: Result shows that there is significant difference ($p < 0.05$) for betweenness centrality, degree centrality, transitivity and density while eigenvector centrality and numbers of edges do not have significant difference. It is

very similar to conclusion by Young et al (2019). Social network analysis was applied in Australian Football League 2009-2016 seasons (1516 matches included) in their research with conclusion that edge density, transitivity and betweenness centrality play an important role on team performance while betweenness centrality's correlation is not significant. More specifically, betweenness centrality should be minimized ideally, degree centrality, transitivity and density should be maximized for the sake of winning. Therefore, this conclusion could be widely used because both of different projects with different datasets in different leagues leading to similar result.

5.3. Implications of result

This project is successful with effective results to give several insights:

- In terms of result from individual player performance analysis, it is proved that player's performance in attacking can be evaluated in more advanced way instead of using basic statistical analysis such as shot on target, passing completion, and possession rate.
- It is also can be concluded that young talent players found by this proposed metric are truly investable unless uncertain circumstances such as injury from French Ligue One 2016-2017 half season. Therefore, same procedure can be conducted in any other leagues only if you got trustable dataset. If a club could bring one highlighted young talent club by this metric, the market value of that player would be doubled at least with high probability. Therefore, club manager and football scout would be potential beneficiaries because this evaluation metric can help them to make decision whether to purchase a young player or not.
- Other than identification of young talent player, the computed action evaluation metric can be also used to assess team performance by aggregating all players' performance value in same team. The team performance was not orientated only by match result but considering all actions performed during the match. Thus, it can provide a new way to analyze current situation of team, which higher team performance value can indicate that playing style or tactic in attacking were not the actual problem. Specifically, some insights or advices can be put forward if one team has much higher cumulative performance value but got less points then the other: 1. Be patient and maintain the performance in attacking. 2. Strikers should focus more on shot quality training because they got lots of opportunities but did not grasp. 3. Defense ability should be focus more than attacking in this circumstance. Coach and players become the potential beneficiaries if proposed evaluation metric applied in team performance assessment.
- Besides, dynamic time warping technique plays an important role in whole evaluation process. Results from this project proved that trajectory analysis can be used for similarity check in

football analysis. Therefore, other advanced trajectory technique can be considered such as Frechet distance measure or more advanced technique like spatial-temporal convolution kernels instead of using dynamic time warping technique.

- In terms of result from collaboration performance analysis by network analysis, it shows that betweenness centrality should be minimized ideally, degree centrality, transitivity and density should be maximized for the sake of winning. Each network measure has its own implications of how to improve the team performance. As for betweenness centrality, it calculates how the ball passes among other players depends on one typical player. Therefore, optimal betweenness centrality should be minimized indicates the ball should be shared among all teammate other than only depends on star player to deal with the ball. In terms of degree centrality and transitivity, they should be maximized because they can reveal the collaboration among players with tacit cooperation to break the defense line, higher value corresponding to higher collaboration. Therefore, coaches can adjust the tactics based on feedback of network indicators real time during the match or post-match analysis in order to construct a more balanced network which can has a better performance.

6. Evaluation, Reflections, and Conclusions

6.1. Overall conclusion

In a conclusion, action contribution was evaluated successfully which can be used to identify talented young players and team performance evaluation. The action evaluation metric is based on dynamic time warping technique, which includes goal probability model creation, possession sequence creation and value assignment based on goal probability model, sub possession sequence creation (including passing and dribble event), dynamic time warping technique used to find similar sequence in training set and compute value of each action (passing and dribble) based on the value of sub sequence after and before action. Talent young players in French Ligue One of season 2016-2017 were identified by this metric and proved by their market value changes that most of identified player by this metric are worth the investment. Besides, action evaluation results can be also used to assess team performance by simply aggregating all players' performance value in same team which can provide a new way to analyze current situation of team other than only focus on current position at league table. It can give some insights for coach and players on how to improve performance in the future. In addition, network science analysis was applied to collaboration performance evaluation. It is found that some network measures should be optimized such as minimize betweenness centrality, maximize degree centrality, maximize transitivity and maximize density in order to improve collaboration between teammate to win the match.

6.2. Limitation

Although the proposed metric in this project achieves objectives successfully, it still has its own limitations. Firstly, only on-ball events were evaluated due to lack of off-ball data. However, positioning of teammate and pressure from the opponent all affect player's choices when making dribble or pass decisions. Secondly, only complete possession sequence more than 3 actions can be evaluated in this metric so that short sequences were not counted which may reduce the accuracy of results. Thirdly, player's decisions during match can be influenced by playing style of team. For example, some coach may encourage player to shot at long range whereas some may demand player to control the ball by short pass other than long pass. Different decision corresponds to various contribution value such as long pass would be assigned higher value than short pass in the proposed metric. In addition, playing with good teammates is normally easier to perform risky but valuable actions than playing in a small club with weaker teammates. Thus, performance evaluated value may biased to players from big club with stronger teammate. Moreover, fast dynamic time warping technique can provide optimal alignments with an O(N) time but if reference database is very large, it will be time-consuming to implement this proposed

evaluation metric. In terms of team performance evaluation by network analysis, the major limitation is lack of data meets the boundary conditions, only 67 from French Ligue One half season of 2016-2017 were used to generate passing network. In addition, results of significant network measures calculated based on French League may cannot applied to other leagues such as Premier League because different leagues have different playing style leading to diverse network measures. Moreover, results from proposed metrics still not verified with football experts.

6.3. Further work

Based on current results of project, ideas of how to improve and extend, this project has been discovered. More data should be gathered in order to improve the accuracy of result. As explained in limitation, off-ball event data should be tracked which will provide new view to conduct performance evaluation analysis. Moreover, proposed metrics can be tested to see if it can resonate with football expert. In addition, different techniques for trajectory analysis such as Frechet distance measure and spatial-temporal convolution kernels can be merged to find trajectory similarity more accurate and efficient. However, these approaches cannot reduce computational time when conduct similarity check. Therefore, image processing analysis may be a possible way to train a model to find similar sequence with same pattern. Each trajectory can be treated as a white line plotted on a fixed size black background (football pitch) to create an image. All images generated by reference possession sequence can be the training data of a convolutional neural network model. Once a CNN model is trained, similarity check can be implemented extreme fast by input test image. Besides, in terms of team performance evaluation by network analysis, more data can also improve the accuracy of result. One possible extension is to construct dynamic network in real time (e.g. per 10 minuets of the match) which can be used to monitor network measures during the match and to help coach make tactic changes if necessary.

6.4. Reflection

From my perspective of view, there were plenty works done during last three month. The word plan set in the initial project was not appropriate because there were lots of preliminary data preparation but necessary for this project. It was very time consuming to do that preparation especially for me who had no experience before with related topics. In addition, the method to achieve objectives was changed from initial proposal due to new inspiration got from other literatures. Therefore, if this project starts again, I would plan more time on literature review before preliminary data preparation to prevent method changes when project was already started for a while. Dynamic network in real time was proposed initially but not be implement due to

lack of time. However, the results of current project are also interesting and rewarding. This project helps me to learn plenty of trajectory analysis and network analysis knowledge in both theoretical and practical ways. My interest in time series analysis and network analysis was motivated after this project. Overall, it is a rewarding project with plenty of benefits for my future study.

References

Barrat, A., Barthelemy, M. and Vespignani, A., 2007. The architecture of complex weighted networks: Measurements and models. In Large scale structure and dynamics of complex networks: from information technology to finance and natural science (pp. 67-92).

Bransen, L., van de Velden, M. and Van Haaren, J., 2017. Valuing passes in football using ball event data. Master of Science. Erasmus University Rotterdam. URl: http://hdl. handle. net/2105/41346.

Brooks, J., Kerr, M. and Guttag, J., 2016, August. Developing a data-driven player ranking in soccer using predictive model weights. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 49-55). ACM.

Chan, T.C. and Singal, R., 2016. A Markov Decision Process-based handicap system for tennis. Journal of Quantitative Analysis in Sports, 12(4), pp.179-188.

Decroos, T., Van Haaren, J., Dzyuba, V. and Davis, J., 2017, September. STARSS: A Spatio-Temporal Action Rating System for Soccer. In MLSA@ PKDD/ECML (pp. 11-20).

Eggels, H.P., 2016. Expected goals in soccer: Explaining match results using predictive analytics. In the Machine Learning and Data Mining for Sports Analytics workshop (p. 16).

Ensum, J., Pollard, R. and Taylor, S., 2005, May. Applications of logistic regression to shots at goal at association football. In Science and football V: the proceedings of the Fifth World Congress on Science and Football (p. 214).

Freeman, L.C., 1978. Centrality in social networks conceptual clarification. Social networks, 1(3), pp.215-239.

Gould, P. and Gatrell, A., 1979. A structural analysis of a game: the Liverpool v Manchester United Cup Final of 1977. Social Networks, 2(3), pp.253-273.

Gonçalves, B., Coutinho, D., Santos, S., Lago-Penas, C., Jiménez, S. and Sampaio, J., 2017. Exploring team passing networks and player movement dynamics in youth association football. PloS one, 12(1), p.e0171156.

Jayal, A., McRobert, A., Oatley, G. and O'Donoghue, P., 2018. Sports Analytics: Analysis, Visualisation and Decision Making in Sports Performance. Routledge.
Lewis, M. 2003. Moneyball: the art of winning an unfair game. New York, W.W. Norton.

Kempe, M., Goes, F.R. and Lemmink, K.A., 2018, October. Smart Data Scouting in Professional Soccer: Evaluating Passing Performance Based on Position Tracking Data. In 2018 IEEE 14th International Conference on e-Science (e-Science) (pp. 409-410). IEEE.

Keogh, E.J. and Pazzani, M.J., 2001, April. Derivative dynamic time warping. In Proceedings of the 2001 SIAM international conference on data mining (pp. 1-11). Society for Industrial and Applied Mathematics.

Kröckel, P., Piazza, A. and Neuhofer, K., 2017, August. Dynamic Network Analysis of the Euro2016 Final: Preliminary Results. In 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW) (pp. 114-119). IEEE.

Ligue 1 Conforama. 2019. League Table 2016-2017 season. [ONLINE] Available at: https://www.ligue1.com/ligue1/classement?saison=61. (Accessed 21 August 2019).

Müller, M., 2007. Information retrieval for music and motion (Vol. 2). Heidelberg: Springer.

Newman, M., 2018. Networks. Oxford university press.

OptaPro (2019) Opta Pro Open data. [ONLINE] Available at: http://www.optasportspro.com/ (Accessed: 01 August 2019).

Pena, J.L. and Touchette, H., 2012. A network theory analysis of football strategies. arXiv preprint arXiv:1206.6904.

Power, P., Ruiz, H., Wei, X. and Lucey, P., 2017, August. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1605-1613). ACM.

Salvador, S. and Chan, P., 2007. Toward accurate dynamic time warping in linear time and space. Intelligent Data Analysis, 11(5), pp.561-580.

Schulte, O., Zhao, Z. and Routley, K., 2015. What is the Value of an Action in Ice Hockey? Learning a Q-function for the NHL. In Proceedings of the 2nd Workshop on Machine Learning and Data Mining for Sports Analytics.

Senin, P., 2008. Dynamic time warping algorithm review. Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA, 855(1-23), p.40.

Srinivasan, B., 2017, December. A Social Network Analysis of Football-Evaluating Player and Team Performance. In 2017 Ninth International Conference on Advanced Computing (ICoAC) (pp. 242-246). IEEE.

StatsBomb (2019) StatsBomb Open Data. Available at: https://github.com/statsbomb/open-data (Accessed: 01 August 2019).

Szczepański, Ł. and McHale, I., 2016. Beyond completion rate: evaluating the passing ability of footballers. Journal of the Royal Statistical Society: Series A (Statistics in Society), 179(2), pp.513-533.

Tanida, K. (2017) Fastdtw 0.3.2. Available at: https://pypi.org/project/fastdtw/ (Accessed: 15 August 2019).

Young, C.M., Luo, W., Gastin, P., Lai, J. and Dwyer, D.B., 2019. Understanding effective tactics in Australian football using network analysis. International Journal of Performance Analysis in Sport, 19(3), pp.331-341.

Appendix A: Project Proposal for MSc in Data Science

Name: Shengqiang Fan

E-mail address: Shengqiang.Fan@city.ac.uk

Contact Phone number: (+86) 13586377263

Project Title: Understanding player and team performance through visual analysis of football data

Supervisor: Dr. Cagatay Turkay

1. Introduction

Football is the most attraction sport in the world, which has more than 4 billion fans all over the world (Giulianotti, 2012). It is therefore no surprise that large amounts of researches on football analysis were conducted from different professions such as journalist, football coach, club manager and so on. Nowadays, with more and more data can be achieved with rapid development of technology, football player's performance is judged from not only subjective view but also data driven view. Some widely accepted data by coach, players and fans are ball possession percentage, pass accuracy, shot on target *etc.*, which will also be presented at halftime/fulltime tactic analysis. Moreover, expected goal or expected pass model drew a great attention on sports analytics community in the past few years which was initially put forward by Ensum *et al* (2005). Regression technique was used to evaluate evaluation player's shot ability. Nevertheless, existing approaches like expected goals are focus more on result of an action (e.g. only consider goal probability given a shot or pass completion rate when perform a pass) rather than player's performance contribution. Football is a team sport not an individual game, each individual contribution should be emphasized however personal ability of a super star can dominate a win in some cases. For example, player A slices the opponent's defense by finding player B with a penetrate pass and player B makes a delta-pass to penalty area results in player C scores a goal. In this attacking example, player C is marked with a goal and player B is rewarded with an assist, however, player A did not get any big reward in statistical information, just a successful pass will be recorded. But technically, the penetrate pass from player A could be the key to goal. In addition, there is also a common phenomenon that player's contribution is not recorded when a wonderful chance was created but striker missed to goal. Therefore, one aim of this project will evaluate player's performance based on quantifying their pass contribution to fill the gap between existing evaluation metrics and use visualization techniques to help people have a better understanding on overall contribution of a player in attacking. More contribution of a player means more threats/opportunities can be created. Moreover, football is a team sport since an excellent performance of single player still cannot lead to a win so that performance evaluation should also be considered at the team level. It can be evaluated by considering the whole team as a network. Hence, passing network analysis will be implemented on basis of social network theory because passing between teammates can be treated as a small network between each player linked by passing event. Power *et al* (2017) pointed out that the most frequent event during football match is passing, qualitative and quantitative analysis on passing data may give some interesting insights. After passing network is created, topological analysis will be conducted in order to evaluate the performance at team level.

1.1 Title

Understanding player and team performance through visual analysis of football data.

1.2 Research questions

- How can we use machine learning techniques to find out who has best performance in a league?
- Is it possible to evaluate performance at team level by using network science theory?

- Are there any relationships can be found between individual player's performance and team performance?
- How to use visualization techniques make above analysis result easier to use and explore by coaches and journalist?

1.3 Research objectives

- To identify and understand existing approaches used by other researchers in player's performance evaluation. Find limitations of current metrics and then to propose a new metric to evaluate player's performance in attacking play.
- To understand the source data well provided by Opta Sport company and Statsbomb company in order to make use of these football event-based data sets.
- To evaluate team performance by using network science theory based on passing network.
- To find relationships between individual player's performance and team performance.
- To build a dashboard with visualization techniques includes above analysis results for easier to use and explore by coaches and journalist.

1.4 Outcome

- A new metric to evaluate football player's performance in attacking and then top players will be ranked who can create more opportunities / threats per game (90 minutes).
- A systematic way to use network science theory to evaluate team performance based on network theory.
- A dashboard including above analysis by visualization techniques for people to explore.

1.5 Beneficiaries

- Football analyst and journalist who can stand on a new data driven side to evaluate football player and team performance to tell difference stories.
- Coaches and club manager who can use data to support their subjective inference on player's contribution in attacking. Meanwhile, coaches can adjust tactics based on relationship between individual player performance and team performance in order to improve team's performance. Coaches can also get tactic suggestion in real-time by using dynamic network analysis.

2. Critical Context

Sports analysis gains lots of attentions due to big influence of book Moneyball written by Michael Lewis (2003). It demonstrated how a baseball manager to find undervalued player by using sabermetrics. More and more books and papers were published to provide guidance of how to use data in sports for decision makers in sports such as coaches and managers. With rapid development of information technology, lots of high-quality data can be obtained. Opta Sport company and Statbomb company capture all event-based tracking data from every second during the match. More specifically, different football events such as passing, dribble, shot and tackle were recorded with corresponding x,y location on football pitch as well as player number with timestamp. These data make performance evaluation possible rather on stay on basic statistical analysis such as shot on target, passing completion, possession rate and so on. Based on research question of this project, literature review is mainly focus on two specific topics: player performance evaluation of football player and network science in football analysis for team performance.

Player performance evaluation

Several researches were conducted on performance evaluation of a football player in past few years. Szczepa´nski and McHale (2016) put forward to use generalized additive mixed model to quantify passing ability of a football player and point out key players based on that passing model. But only pass difficulty was considered in that model. Brooks and Guttag (2016) designed a player's performance ranking system on basis of the value of passes. A linear classifier was

trained to find out importance of pass start locations and end locations for creating a shot opportunity therefore, any passes can be valued. This approach can avoid problems that only passes directly leading to a shot were considered. Moreover, Decroos *et al* (2017) proposed an action rating system based on event-based tracking data or called play-by-play data. The rating system consists of three parts. First part is to split event-based match data into different phase. A phase is a consecutive event which is started when possession switches and ended if dead ball event occurs such as corner/free kick is award as shown in figure 1. Dynamic time warping method was used to compute similarity between two phases. Then any phase can be valued by average the outcome (whether a goal is scored or not) of k nearest neighbors (k=100 was proposed in their paper). Each event in one phase was assigned a normalized weight on the basis of exponential-decay method, which means event at the beginning of a phase has relative low weight. Similar research was conducted by Bransen and Van Haaren (2017), they trained an expected goal model to evaluate the outcome of one phase instead of computing average number of many phases directly leading to a goal from Decroos's research. Expected goal method was initially put forward by Ensum *et al* (2005) in order to explore important factors that affect goal probability of a shot on dataset of 37 matches of world cup. Shot distance, shot angle, defender's pressure and whether the shot was followed instantaneously by a cross were pointed out as most influence factors from his logistic regression model. Expected goal model is widely accepted by sports analysis community nowadays to compute probability of goal when given a shot opportunity under different situations. It can be used to evaluation player's shot ability by comparing difference of total number of goals expected to score and actual goals in total. Eggels (2016) explored performance of 4 classifiers (logistic regression, decision tree, random forest and Ada Boost) used to train the expected goal model. Random forest was outperformed in AUC and F-score under 13 features were created for each shot opportunity. Season analysis was put forward to track a club's behavior by aggregating all player's expected goal compared to team's actual goal during the season. Juventus in 2015-16 season had a bad opening, only won three matches of first 10 games. Fans and news reporters started to question players and coaches. However, expected goal provided a new way to analyze that problem. It is shown that Juventus' s expected goal value was much higher than actual goal scored in first 10 games, which indicated that Juventus were dominating matches and creating lots of opportunities but not seized the goal opportunity. It also indicated that playing style or tactic were not the actual problem.

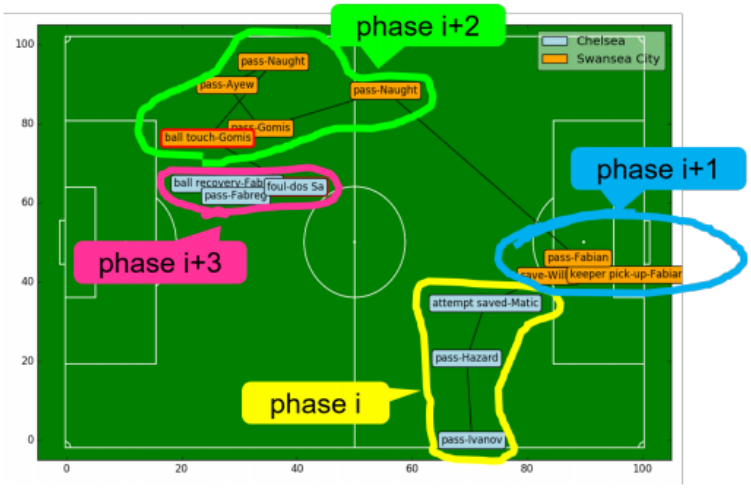


Figure 1: Example of sequence of events are split into several phases (Decroos et al, 2017)

Similar investigations were performed in other sports. A Markov Model was constructed by Schulte *et al* (2015) to find the most valuable action leading to winning in ice hockey. Similarly, Chan and Singal (2016) also proposed a Markov Decision Process system for tennis. Base on that, football game can be also treated as a Markov Model where transient states is different location of football pitch and absorption state could be the outcome of a possession (whether a goal is scored or not). Both Transition probability and absorption probabilities can be calculated

from historical match data. Nonetheless, possible limitation of Markov Model approach could be 'memoryless' property of Markov model is not true in football events.

Network science in football analysis for team performance

A passing network is created from successful passes between teammates, where nodes of network are players with directed connections by passes between players. The concept of passing network in football was firstly introduced by Gould and Gatrell (1979) but did not gain much attention from sports or scientific community. Related researches on passing network just increased gradually in recent years. Pena and Touchette (2012) indicated that passing network can be used as a visualization tool because it can give the insight into statics right away as shown in figure 2. Key player can be easily pointed out through closeness centrality however, it has its bias when using passing network. Midfield players do have higher closeness centrality due to its role in the team. Betweenness centrality is a different concept from closeness centrality, it can calculate how the ball passes among other players depends on one typical player. In other word, a player with high betweenness plays an important role in team's tactic. And if betweenness near to 0 indicates that the low involvement of that player. Generally speaking, an evenly distributed and relatively low betweenness should be preferred as it indicates a balanced passing strategy from tactic view of the whole team. PageRank centrality was also introduced by Gould and Gatrell (1979) which can identify a relationship between key players. A good defending tactic is to reduce PageRank centrality of rival's key player in order to decrease the connection between key players. Srinivasan (2017) also applied network analysis on football. Cliques, Average Clustering Coefficient, top eigenvalue and λ distance similarity were used to track the team performance in certain time period. Gonçalves *et al* (2017) suggested that a well-balanced passing network with a lower passing dependency may be a best combination to maximize performance of a team. Besides, Kröckel *et al* (2017) evaluated team performance by using dynamic network analysis approach to investigate team performance changes during the 90 mins of the game, which gives a good starting point. There is also agreed by Kröckel *et al* that a well-distributed passing networks could result in better performance. Based on dynamic network analysis, decision help to coach in real-time could be one of applications suggested by Kröckel *et al* (2017).

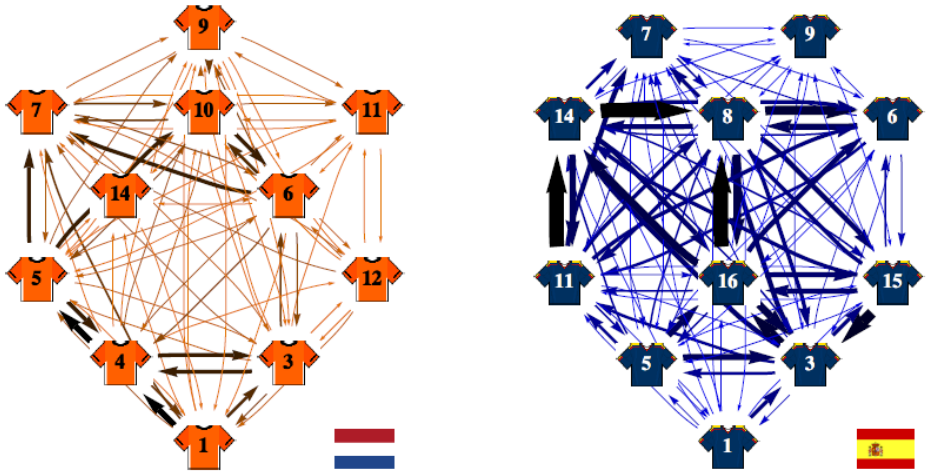


Figure 2: Passing network created by Pena and Touchette (2012)

3. Approaches: Methods & Tools for Design, Analysis & Evaluation

Literature review

Initial literature review was conducted as discussed in critical context section. However, more scholarly or peer reviewed papers should be reviewed in order to get ideas and knowledges from previous research in performance/passing evaluation and network science topics. The overall goal of literature review is to have a better understanding of current work and get ideas to answer research questions in this project.

Gathering, parsing and merging data

Opensource data is provided by Opta Sports company and Statsbomb company in XML and json format respectively. Both datasets are event-based sequential data which can be used for this project. The datasets include 190 matches from first half season of French Ligue One 2016-2017 season, 43 matches from 2018 FIFA World Cup, 37 matches from FA Women's Super League 2018-2019 season and 49 matches from National Women's Soccer League (US) 2018 season. Data were recorded in what-who-where-when structure of each match. More specifically, sequential event such as passing, dribble, shot and tackle was recorded with corresponding x,y location on football pitch as well as player number at timestamps. Therefore, first task is to parse these event-based data into csv file and store them in a database for convenience of further query. In addition, related documents and specifications provided by Opta and Statbomb should be read and fully understood because there are hundreds of event types and qualifier types. It is necessary to have a better understanding of dataset before using it. Moreover, pitch coordinate is different between Opta data and Statsbomb data, where x ∈ [0,100] and y ∈ [0,100] in Opta data but x ∈ [0,120] and y ∈ [0,80] in Statsbomb data. Therefore, a normalization process will be performed in order to make meaning of x,y coordinate consistent for both dataset before merging these two datasets.

Create a player's performance metric based on machine learning techniques

Approach for player's performance evaluation is divided into 3 phases.

- In the first phase, an expected goal model will be trained with three types of classifiers as logistic regression, random forest and XGboost. Features will be extracted and created from prepared dataset such as shot location (distance to goal and shot angle), last event type before shot, current score state (lead/tie/behind). Then, an expected threat model will be created based on expected goal model. The expected threat model can avoid problems when only passes directly leading to a shot were considered. It will be computed from two parts, goal probability at current location and goal probability if ball is passed to next location. Therefore, the expected threat model can be expressed below:

$$\text{Expected\ Threat}(\{xT\}\_\{x,y\}) = \left(\{Ps\}\_\{x,y\} \times xG_{x,y}\right) + \{Pp\}\_\{x,y\} \times \sum_a \sum_b Ts_{(x,y)\rightarrow(a,b)} \times xT_{a,b}$$

  Where $Ps_{x,y}$ represents probability of taking a shot at location x,y and $Pp_{x,y}$ stands for the probability of taking a pass at location x,y. $xG_{x,y}$ is expected goal probability from trained model. $Ts_{(x,y)\rightarrow(a,b)}$ is the transition probability at current location x,y to pass the ball to location a,b.

- In the second phase, sequence of events is split into several phases in similar way as shown in figure 1 in section 2. When given a new phase, similar phase will be clustered by using spatial-temporal convolution kernels for trajectories. Each phase's outcome will be calculated based on expected threat.
- Finally, pass contribution can be calculated as average of expected threat with similar trajectories after the pass minus expected threat before a pass is performed. A player's performance can be calculated by aggregating all the pass contribution in one match.

Create a team performance metric based on network analysis

Team performance can be evaluated by considering the whole team as a network. Successful passes will be extracted from prepared dataset of each match. Passing network can be created by using these successful pass data where nodes are players linked by passes. Topological analysis can be used to explore which topology characteristic leads to better performance. For example, clustering coefficient can be used to measure how frequent interactions between players. If a player has very high betweenness centrality indicates that player plays an important role in attacking. It is proved that an evenly distributed passing network with relatively low betweenness is preferred as it represents a balanced passing strategy from tactic view of the whole team. However, the links between team performance and network theory is still not clear. Important features in topology which affect team's performance will be explored in this project. Moreover, timestamp will also be recorded for each pass in order to make dynamic network analysis. With

dynamic network analysis, it may can provide useful suggestion for coaches in real-time. NetworkX Python library will be used to conduct this process.

Evaluation method

Dataset will be split into training set and evaluation set for training the expected goal model. F-score/ROC-AUC curve will be used to judge the performance of trained expected goal model with different classifier. Top players who can create more opportunities / threats per game (90 minutes) in will be ranked based on dataset of French Ligue One 2016-2017 season. The result will be compared to their goal and assists stats as well as market value at that time (2016-2017). The TransferMarkt will be used to obtain the market value of each player, which has high reputation in assessing a player's market value. A player with high market value indicates he is a key player with outstanding performance or high potential. Therefore, there is no doubt that that player can make more threat in attacking play to help team win the game especially for attackers and midfield. In terms of team performance, passing network evaluation will be tested on 2018 FIFA World Cup to figure out key topological characteristics help France to win the World Cup.

Develop a dashboard

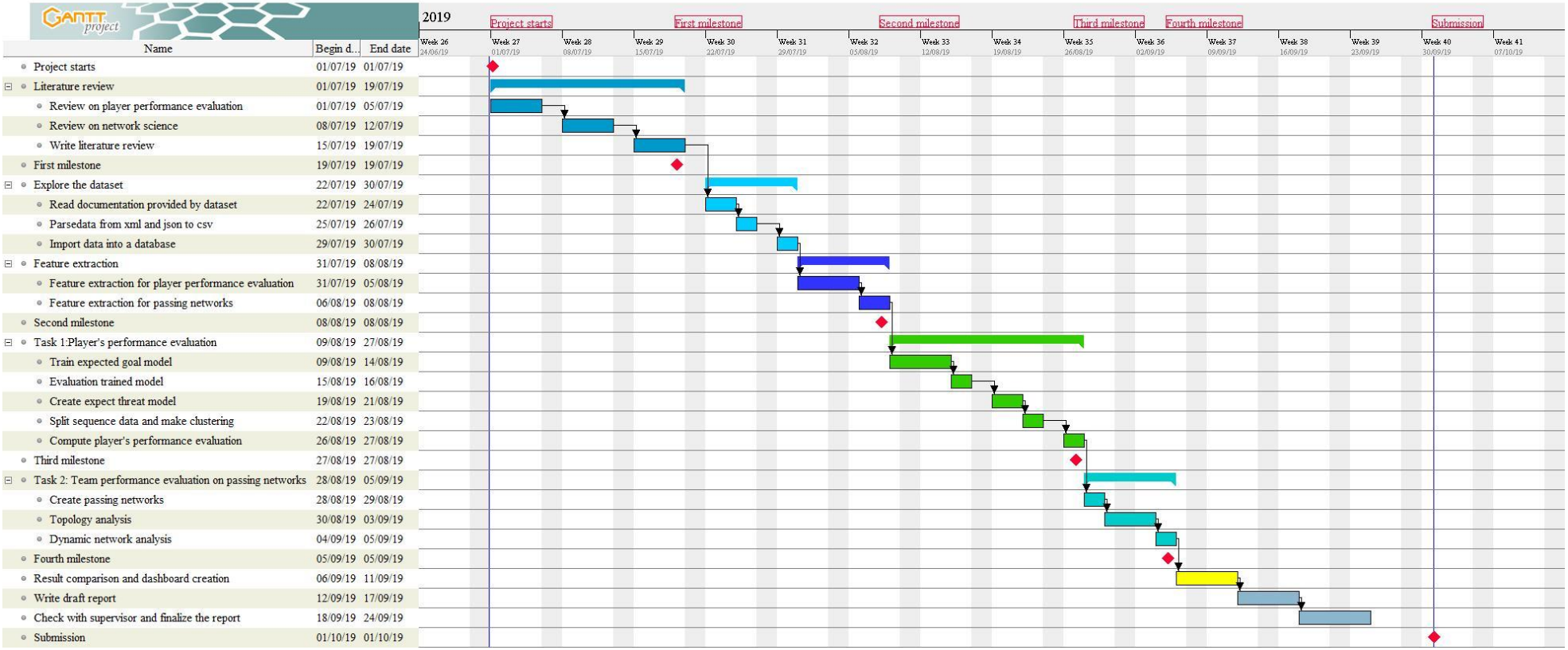
An interactive dashboard will be created by Tableau software in order to present the evaluation results of player and team performance. Coaches can use player's performance index versus match-match of the season to track if player is making progress or in trouble from match by match. Journalist can use top ranking players or even last ranking players based on performance evaluation to tell different stories for football fans. Moreover, coaches can also get some real-time tactic suggestion from dynamic network analysis.

Work plan, risk register and reporting

Work plan and Gantt chart is shown in next section. This project will start at 1st Jul 2019 and end in 1st Oct 2019. The Gantt chart should be up to date with project going on. The draft report is expected to be finished 2 weeks before the deadline, which will be used for seeking supervisor's suggestion and reserving enough time to finalize the report. A risk register is also created in section 5. Mitigation or prevention plan can be followed if any risk happens. The risk register is also needed to be up to date since new risk may appear with project going on.

4. Work plan

A Gantt chart is created for this project, where left side demonstrates all tasks to be done and right side is the Gantt chart.



Name	Begin d...	End date
Project starts	01/07/19	01/07/19
Literature review	01/07/19	19/07/19
Review on player performance evaluation	01/07/19	05/07/19
Review on network science	08/07/19	12/07/19
Write literature review	15/07/19	19/07/19
First milestone	19/07/19	19/07/19
Explore the dataset	22/07/19	30/07/19
Read documentation provided by dataset	22/07/19	24/07/19
Parsedata from xml and json to csv	25/07/19	26/07/19
Import data into a database	29/07/19	30/07/19
Feature extraction	31/07/19	08/08/19
Feature extraction for player performance evaluation	31/07/19	05/08/19
Feature extraction for passing networks	06/08/19	08/08/19
Second milestone	08/08/19	08/08/19
Task 1:Player's performance evaluation	09/08/19	27/08/19
Train expected goal model	09/08/19	14/08/19
Evaluation trained model	15/08/19	16/08/19
Create expect threat model	19/08/19	21/08/19
Split sequence data and make clustering	22/08/19	23/08/19
Compute player's performance evaluation	26/08/19	27/08/19
Third milestone	27/08/19	27/08/19
Task 2: Team performance evaluation on passing networks	28/08/19	05/09/19
Create passing networks	28/08/19	29/08/19
Topology analysis	30/08/19	03/09/19
Dynamic network analysis	04/09/19	05/09/19
Fourth milestone	05/09/19	05/09/19
Result comparison and dashboard creation	06/09/19	11/09/19
Write draft report	12/09/19	17/09/19
Check with supervisor and finalize the report	18/09/19	24/09/19
Submission	01/10/19	01/10/19

5. Risk register

No.	Risk description	Likelihood (1-3)	Consequence (1-5)	Impact (l*c)	Mitigation/prevention Plan
1	Source data cannot be access from Opta and Statsbomb	1	5	5	Always check open source availability, download and make backup on cloud drive or hard drive with original open source data.
2	Broken/Crashed of laptop or even loss of laptop	2	5	10	Make backup on cloud drive or hard drive with all files includes scripts/data set/results/word file/images etc. on weekly or daily basis. Make sure every progress is recorded.
3	Loss of Motivation	1	4	4	Choose the specific topic which I am interested in and keep in touch with supervisors and classmate to remain the motivation.
4	Procrastination/Time overrun	2	3	6	Make work plan by using Gantt Chart, reserve enough time for contingencies. Follow and keep work plan up to date when project begins.
5	Underestimate workload or underestimate work time	3	2	6	Keep work plan updated and adjust work load based on remaining time as well as keep in touch with supervisor.
6	New tasks arise or change of tasks from proposal	3	3	9	Keep informed with supervisor and get advice whether it need to be changed of tasks or not. If changed, make work plan up to date accordingly.
7	Difficulty in implementing theory into practice e.g. coding issues	2	2	4	Seek for other expertise's help, could be supervisor, classmate or online forum.
8	Data wrangling process may need to be repeated after some findings which takes extra-long time	2	2	4	Data wrangling process always cost majority of time when play with data for data scientist, always reserve enough time for conducting all machine learning process more than once.
9	Illness or accident occurs	1	5	5	Inform supervisor and school as soon as possible for applying extenuating circumstances

6. References

Bransen, L., van de Velden, M. and Van Haaren, J., 2017. Valuing passes in football using ball event data. Master of Science. Erasmus University Rotterdam. URl: http://hdl. handle. net/2105/41346.

Brooks, J., Kerr, M. and Guttag, J., 2016, August. Developing a data-driven player ranking in soccer using predictive model weights. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 49-55). ACM.

Chan, T.C. and Singal, R., 2016. A Markov Decision Process-based handicap system for tennis. Journal of Quantitative Analysis in Sports, 12(4), pp.179-188.

Decroos, T., Van Haaren, J., Dzyuba, V. and Davis, J., 2017, September. STARSS: A Spatio-Temporal Action Rating System for Soccer. In MLSA@ PKDD/ECML (pp. 11-20).

Eggels, H.P., 2016. Expected goals in soccer: Explaining match results using predictive analytics. In the Machine Learning and Data Mining for Sports Analytics workshop (p. 16).

Ensum, J., Pollard, R. and Taylor, S., 2005, May. Applications of logistic regression to shots at goal at association football. In Science and football V: the proceedings of the Fifth World Congress on Science and Football (p. 214).

Giulianotti, R., 2012. Football. The Wiley‐Blackwell Encyclopedia of Globalization.

Gould, P. and Gatrell, A., 1979. A structural analysis of a game: the Liverpool v Manchester United Cup Final of 1977. Social Networks, 2(3), pp.253-273.

Gonçalves, B., Coutinho, D., Santos, S., Lago-Penas, C., Jiménez, S. and Sampaio, J., 2017. Exploring team passing networks and player movement dynamics in youth association football. PloS one, 12(1), p.e0171156.

Jayal, A., McRobert, A., Oatley, G. and O'Donoghue, P., 2018. Sports Analytics: Analysis, Visualisation and Decision Making in Sports Performance. Routledge.

Lewis, M. (2003). Moneyball: the art of winning an unfair game. New York, W.W. Norton.

Kempe, M., Goes, F.R. and Lemmink, K.A., 2018, October. Smart Data Scouting in Professional Soccer: Evaluating Passing Performance Based on Position Tracking Data. In 2018 IEEE 14th International Conference on e-Science (e-Science) (pp. 409-410). IEEE.

Kröckel, P., Piazza, A. and Neuhofer, K., 2017, August. Dynamic Network Analysis of the Euro2016 Final: Preliminary Results. In 2017 5th International Conference on Future Internet of Things and Cloud Workshops (FiCloudW) (pp. 114-119). IEEE.

Pena, J.L. and Touchette, H., 2012. A network theory analysis of football strategies. arXiv preprint arXiv:1206.6904.

Power, P., Ruiz, H., Wei, X. and Lucey, P., 2017, August. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1605-1613). ACM.

Schulte, O., Zhao, Z. and Routley, K., 2015. What is the Value of an Action in Ice Hockey? Learning a Q-function for the NHL. In Proceedings of the 2nd Workshop on Machine Learning and Data Mining for Sports Analytics.

Szczepański, Ł. and McHale, I., 2016. Beyond completion rate: evaluating the passing ability of footballers. Journal of the Royal Statistical Society: Series A (Statistics in Society), 179(2), pp.513-533.

Srinivasan, B., 2017, December. A Social Network Analysis of Football-Evaluating Player and Team Performance. In 2017 Ninth International Conference on Advanced Computing (ICoAC) (pp. 242-246). IEEE.

7. Ethical Issues

Research Ethics Review Form: BSc, MSc and MA Projects
Computer Science Research Ethics Committee (CSREC)

http://www.city.ac.uk/department-computer-science/research-ethics

Undergraduate and postgraduate students undertaking their final project in the Department of Computer Science are required to consider the ethics of their project work and to ensure that it complies with research ethics guidelines. In some cases, a project will need approval from an ethics committee before it can proceed. Usually, but not always, this will be because the student is involving other people ("participants") in the project.

In order to ensure that appropriate consideration is given to ethical issues, all students must complete this form and attach it to their project proposal document. There are two parts:

PART A: Ethics Checklist. All students must complete this part. The checklist identifies whether the project requires ethical approval and, if so, where to apply for approval.

PART B: Ethics Proportionate Review Form. Students who have answered "no" to questions 1 – 18 and "yes" to question 19 in the ethics checklist must complete this part. The project supervisor has delegated authority to provide approval in such cases that are considered to involve MINIMAL risk.

The approval may be provisional: the student may need to seek additional approval from the supervisor as the project progresses and details are established.

A.1 If you answer YES to any of the questions in this block, you must apply to an appropriate external ethics committee for approval and log this approval as an External Application through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
1.1	Does your research require approval from the National Research Ethics Service (NRES)? *e.g. because you are recruiting current NHS patients or staff?* *If you are unsure try - https://www.hra.nhs.uk/approvals-amendments/what-approvals-do-i-need/*	**NO**
1.2	Will you recruit participants who fall under the auspices of the Mental Capacity Act? *Such research needs to be approved by an external ethics committee such as NRES or the Social Care Research Ethics Committee - http://www.scie.org.uk/research/ethics-committee/*	**NO**
1.3	Will you recruit any participants who are currently under the auspices of the Criminal Justice System, for example, but not limited to, people on remand, prisoners and those on probation? *Such research needs to be authorised by the ethics approval system of the National Offender Management Service.*	**NO**
A.2 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee, you must apply for approval from the Senate Research Ethics Committee (SREC) through Research Ethics Online - https://ethics.city.ac.uk/		*Delete as appropriate*
2.1	Does your research involve participants who are unable to give informed consent? *For example, but not limited to, people who may have a degree of learning disability or mental health problem, that means they are unable to make an informed decision on their own behalf.*	**NO**
2.2	Is there a risk that your research might lead to disclosures from participants concerning their involvement in illegal activities?	**NO**
2.3	Is there a risk that obscene and or illegal material may need to be accessed for your research study (including online content and other material)?	**NO**
2.4	Does your project involve participants disclosing information about special category or sensitive subjects? *For example, but not limited to: racial or ethnic origin; political opinions; religious beliefs; trade union membership; physical or mental health; sexual life; criminal offences and proceedings*	**NO**

2.5	Does your research involve you travelling to another country outside of the UK, where the Foreign & Commonwealth Office has issued a travel warning that affects the area in which you will study? *Please check the latest guidance from the FCO - http://www.fco.gov.uk/en/*	**NO**
2.6	Does your research involve invasive or intrusive procedures? *These may include, but are not limited to, electrical stimulation, heat, cold or bruising.*	**NO**
2.7	Does your research involve animals?	**NO**
2.8	Does your research involve the administration of drugs, placebos or other substances to study participants?	**NO**
A.3 If you answer YES to any of the questions in this block, then unless you are applying to an external ethics committee or the SREC, you must apply for approval from the Computer Science Research Ethics Committee (CSREC) through Research Ethics Online - https://ethics.city.ac.uk/ **Depending on the level of risk associated with your application, it may be referred to the Senate Research Ethics Committee.**		*Delete as appropriate*
3.1	Does your research involve participants who are under the age of 18?	**NO**
3.2	Does your research involve adults who are vulnerable because of their social, psychological or medical circumstances (vulnerable adults)? *This includes adults with cognitive and / or learning disabilities, adults with physical disabilities and older people.*	**NO**
3.3	Are participants recruited because they are staff or students of City, University of London? *For example, students studying on a particular course or module.* *If yes, then approval is also required from the Head of Department or Programme Director.*	**NO**
3.4	Does your research involve intentional deception of participants?	**NO**
3.5	Does your research involve participants taking part without their informed consent?	**NO**
3.5	Is the risk posed to participants greater than that in normal working life?	**NO**
3.7	Is the risk posed to you, the researcher(s), greater than that in normal working life?	**NO**
A.4 If you answer YES to the following question and your answers to all other questions in sections A1, A2 and A3 are NO, then your project is deemed to be of MINIMAL RISK. **If this is the case, then you can apply for approval through your supervisor under PROPORTIONATE REVIEW. You do so by completing PART B of this form.** **If you have answered NO to all questions on this form, then your project does not require ethical approval. You should submit and retain this form as evidence of this.**		*Delete as appropriate*
4	Does your project involve human participants or their identifiable personal data? *For example, as interviewees, respondents to a survey or participants in testing.*	**NO**

Appendix B: Example code used for this project

There are plenty codes used in this project in order to achieve different data analysis objectives. All codes are submitted as additional zip file. Appendix B represents an example code used to create possession sequence from Opta data:

```
# coding: utf-8
# The following code is used to create possession sequence from opta data
import os
from pathlib import Path
import xml.etree.ElementTree as et
import csv
import numpy as np
import pandas as pd
initial_file_path = os.path.dirname(os.getcwd())
player_path = os.path.join(initial_file_path, "Origional data\Data from Opta in XML")
os.chdir(player_path)

tree_player = et.ElementTree(file = "Players and IDs - F40 - L1 20162017.xml")
playerinfo = tree_player.getroot()
player_ids = []
player_names = []
for child in playerinfo:
    for grchild in child:
        if grchild.tag == "Team":
            for grgrchild in grchild:
                if grgrchild.tag == "Player":
                    player_ids.append(grgrchild.attrib["uID"].lstrip('p'))
                    for grgrgrchild in grgrchild:
                        if grgrgrchild.tag == 'Name':
                            player_names.append(grgrgrchild.text)
        if grchild.tag == "PlayerChanges":
            for grgrchild in grchild:
                for grgrgrchild in grgrchild:
                    if grgrgrchild.tag == "Player":
                        player_ids.append(grgrgrchild.attrib["uID"].lstrip('p'))
                        for grgrgrgrchild in grgrgrchild:
                            if grgrgrgrchild.tag == 'Name':
                                player_names.append(grgrgrgrchild.text)


player_dict = dict(zip(player_ids, player_names))
# path = 'XPSG - available resources\Learning database - French Ligue One 20162017 season - Match Day 1- 19\French Ligue One 20162017 season - Match Day 1- 19'
# os.chdir(path)
opta_path = os.path.join(initial_file_path, "Origional data\Data from Opta in XML\French Ligue Data\Data_Opta")
os.chdir(opta_path)


def parse_each_team_sequence(df,teamname,gameid):
    df_lyon = df[df.team==teamname]
    count = 1
    label = []
    for i in range(len(df_lyon)):
        if i == 0:
            label.append(count)
        if i > 0:
            if df_lyon.iloc[i,0] - 1 == df_lyon.iloc[i-1,0]:
                if df_lyon.iloc[i,1] == 'Shot' and df_lyon.iloc[i-1,1] != 'Pass':
                    count += 1
```

```
                    label.append(count)
                else:
                    label.append(count)
            else:
                count += 1
                label.append(count)

    df_lyon.loc[:,'Label'] = label

    label_count = {}
    for i in range(len(df_lyon)):
        key = df_lyon.iloc[i,-1]
        if key not in label_count.keys():
            label_count[key] = 1
        else:
            label_count[key] += 1

    for i in range(len(label_count)):
        if label_count[i+1] <= 1:
            label_count.pop(i+1, None)

    keys = list(label_count.keys())
    df_lyon_filtered = df_lyon[df_lyon.Label.isin(keys)]

    count = 1
    x = []
    y = []
    possession_2 = {}
    outcome = [0]
    start_min = []
    start_sec = []
    end_min = []
    end_sec = []
    player_flow = []
    for i in range(len(df_lyon_filtered)):
        if i == 0:
            start_min = [df_lyon_filtered.iloc[i,5]]
            start_sec = [df_lyon_filtered.iloc[i,6]]
            x.append(df_lyon_filtered.iloc[i,7])
            y.append(df_lyon_filtered.iloc[i,8])
            x.append(df_lyon_filtered.iloc[i,9])
            y.append(df_lyon_filtered.iloc[i,10])
            player_flow.append(df_lyon_filtered.iloc[i,3])
            possession_2[count] = [x,y,outcome,start_min,start_sec,player_flow]
        if i > 0 and i < len(df_lyon_filtered) - 1:
            if df_lyon_filtered.iloc[i,-1] == df_lyon_filtered.iloc[i-1,-1]:
                        if df_lyon_filtered.iloc[i,1] == 'Pass':
                            if df_lyon_filtered.iloc[i+1,-
    1] != df_lyon_filtered.iloc[i,-1]:
                                end_min = [df_lyon_filtered.iloc[i,5]]
                                end_sec = [df_lyon_filtered.iloc[i,6]]

                                x.append(df_lyon_filtered.iloc[i,7])
                                y.append(df_lyon_filtered.iloc[i,8])
                                player_flow.append(df_lyon_filtered.iloc[i,3])
                                player_flow.append(df_lyon_filtered.iloc[i,3])
                                possession_2[count] = [x,y,outcome,start_min,start_se
    c,end_min,end_sec,player_flow]
                            else:
                                end_min = [df_lyon_filtered.iloc[i,5]]
                                end_sec = [df_lyon_filtered.iloc[i,6]]

                                x.append(df_lyon_filtered.iloc[i,7])
                                y.append(df_lyon_filtered.iloc[i,8])
                                x.append(df_lyon_filtered.iloc[i,9])
                                y.append(df_lyon_filtered.iloc[i,10])
```

```
                                player_flow.append(df_lyon_filtered.iloc[i,3])
                                player_flow.append(df_lyon_filtered.iloc[i,3])
                                possession_2[count] = [x,y,outcome,start_min,start_sec,end_min,end_sec,player_flow]
                        if df_lyon_filtered.iloc[i,1] == 'Shot':
                            outcome = [1]
                            end_min = [df_lyon_filtered.iloc[i,5]]
                            end_sec = [df_lyon_filtered.iloc[i,6]]
                            x.append(df_lyon_filtered.iloc[i,7])
                            y.append(df_lyon_filtered.iloc[i,8])
                            player_flow.append(df_lyon_filtered.iloc[i,3])
                            player_flow.append(df_lyon_filtered.iloc[i,3])
                            possession_2[count] = [x,y,outcome,start_min,start_sec,end_min,end_sec,player_flow]
                    else:
                        if df_lyon_filtered.iloc[i,1] == 'Pass':
                            count += 1
                            outcome = [0]
                            x = []
                            y = []
                            player_flow = []
                            start_min = [df_lyon_filtered.iloc[i,5]]
                            start_sec = [df_lyon_filtered.iloc[i,6]]
                            x.append(df_lyon_filtered.iloc[i,7])
                            y.append(df_lyon_filtered.iloc[i,8])
                            x.append(df_lyon_filtered.iloc[i,9])
                            y.append(df_lyon_filtered.iloc[i,10])
                            player_flow.append(df_lyon_filtered.iloc[i,3])
                            possession_2[count] = [x,y,outcome,start_min,start_sec,end_min,end_sec]
                if i == len(df_lyon_filtered) - 1:
                    if df_lyon_filtered.iloc[i,1] == 'Pass':
                        end_min = [df_lyon_filtered.iloc[i,5]]
                        end_sec = [df_lyon_filtered.iloc[i,6]]

                        x.append(df_lyon_filtered.iloc[i,7])
                        y.append(df_lyon_filtered.iloc[i,8])
                        player_flow.append(df_lyon_filtered.iloc[i,3])
                        player_flow.append(df_lyon_filtered.iloc[i,3])
                        possession_2[count] = [x,y,outcome,start_min,start_sec,end_min,end_sec,player_flow]
                    if df_lyon_filtered.iloc[i,1] == 'Shot':
                        outcome = [1]
                        end_min = [df_lyon_filtered.iloc[i,5]]
                        end_sec = [df_lyon_filtered.iloc[i,6]]
                        x.append(df_lyon_filtered.iloc[i,7])
                        y.append(df_lyon_filtered.iloc[i,8])
                        player_flow.append(df_lyon_filtered.iloc[i,3])
                        player_flow.append(df_lyon_filtered.iloc[i,3])
                        possession_2[count] = [x,y,outcome,start_min,start_sec,end_min,end_sec,player_flow]
            # store dict to txt for later retrieve
            team_data = "Possession_data_%s_%s.txt" % (gameid,teamname)
            Path(team_data).touch()
            f = open(team_data,'w',encoding="utf-8")
            f.write(str(possession_2))
            f.close()
            print(team_data + ' is saved')


        def parse_xml_PossessionData_to_csv(x):
            tree = et.ElementTree(file = os.listdir()[x])
            games = tree.getroot()
            match_details = games[0].attrib
            team_dict = {match_details["home_team_id"]: match_details["home_team_name"],
```

```
                        match_details["away_team_id"]: match_details["away_team_name"]}
            shot_dict = {'13': 'Miss',
                         '14': 'Post',
                         '15': 'Attempt saved',
                         '16': 'Goal'}
            body_dict = {"15": "Head",
                         "72": "Left footed",
                         "20": "Right footed",
                         "21": "Other body part"}
            event_type = [] # pass-1. shot-13/14/15/16. Corner-6. Turnover-9. Foul-4. throw-in/goal kick-5. Interception/blocked-8/74. End-30.
            x = [] # pass/shot location
            y = []
            minutes = []
            seconds = []
            period = []
            team = []
            player = []
            # Only for pass
            Pass_outcome = []
            Pass_length = []
            Pass_angle = []
            Pass_end_x = []
            Pass_end_y = []
            # Only for shot
            Shot_outcome = []
            body_part = []
            for game in games:
                for event in game:
                    # if is pass event
                    if event.attrib.get("type_id") == '1':
                        event_type.append('Pass')
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        Pass_outcome.append(event.attrib.get("outcome"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
                        period.append(event.attrib.get("period_id"))
                        team.append(team_dict[event.attrib.get("team_id")])
                        if event.attrib.get("player_id") is None:
                            player.append('Not Recorded')
                        else:
                            player.append(player_dict[event.attrib.get("player_id")])

                        # Append empty value to shot events
                        Shot_outcome.append('')
                        body_part.append('')
                        for q in event:
                            qualifier = q.attrib.get("qualifier_id")
                            if qualifier == "140": # Pass End X
                                Pass_end_x.append(q.attrib.get("value"))
                            if qualifier == "141": # Pass End Y
                                Pass_end_y.append(q.attrib.get("value"))
                            if qualifier == "212": # Pass Length
                                Pass_length.append(q.attrib.get("value"))
                            if qualifier == "213": # Angle/Radians
                                Pass_angle.append(q.attrib.get("value"))
                    # if is shot event
                    if event.attrib.get("type_id") in shot_dict:
                        event_type.append('Shot')
                        Shot_outcome.append(shot_dict[event.attrib.get("type_id")])
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
```

```
                        period.append(event.attrib.get("period_id"))
                        team.append(team_dict[event.attrib.get("team_id")])
                        player.append(player_dict[event.attrib.get("player_id")])
                        # Append empty value to pass events
                        Pass_outcome.append('')
                        Pass_length.append('')
                        Pass_angle.append('')
                        Pass_end_x.append('')
                        Pass_end_y.append('')
                        for q in event:
                            qualifier = q.attrib.get("qualifier_id")
                            # print(qualifier)
                            if qualifier in body_dict.keys():
                                body_part.append(body_dict[qualifier])
                    # if corner awarded
                    if event.attrib.get("type_id") == '6' and event.attrib.get("outco
    me") == '1':
                        event_type.append('Corner awarded')
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
                        period.append(event.attrib.get("period_id"))

                        team.append('')
                        player.append('')
                        # Append empty value to shot events
                        Shot_outcome.append('')
                        body_part.append('')
                        # Append empty value to pass events
                        Pass_outcome.append('')
                        Pass_length.append('')
                        Pass_angle.append('')
                        Pass_end_x.append('')
                        Pass_end_y.append('')

                    # if Foul
                    if event.attrib.get("type_id") == '4' and event.attrib.get("outco
    me") == '1':
                        event_type.append('Foul')
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
                        period.append(event.attrib.get("period_id"))

                        team.append('')
                        player.append('')
                        # Append empty value to shot events
                        Shot_outcome.append('')
                        body_part.append('')
                        # Append empty value to pass events
                        Pass_outcome.append('')
                        Pass_length.append('')
                        Pass_angle.append('')
                        Pass_end_x.append('')
                        Pass_end_y.append('')

                    # if throw-in / goal kick
                    if event.attrib.get("type_id") == '5' and event.attrib.get("outco
    me") == '1':
                        event_type.append('Out')
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
```

```
                        period.append(event.attrib.get("period_id"))

                        team.append('')
                        player.append('')
                        # Append empty value to shot events
                        Shot_outcome.append('')
                        body_part.append('')
                        # Append empty value to pass events
                        Pass_outcome.append('')
                        Pass_length.append('')
                        Pass_angle.append('')
                        Pass_end_x.append('')
                        Pass_end_y.append('')
                    # if half time/full time
                    if event.attrib.get("type_id") == '30':
                        event_type.append('End')
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
                        period.append(event.attrib.get("period_id"))

                        team.append('')
                        player.append('')
                        # Append empty value to shot events
                        Shot_outcome.append('')
                        body_part.append('')
                        # Append empty value to pass events
                        Pass_outcome.append('')
                        Pass_length.append('')
                        Pass_angle.append('')
                        Pass_end_x.append('')
                        Pass_end_y.append('')

                    # if Interception/blocked
                    if event.attrib.get("type_id") in ['8', '74']:
                        event_type.append('Interception')
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
                        period.append(event.attrib.get("period_id"))

                        team.append('')
                        player.append('')
                        # Append empty value to shot events
                        Shot_outcome.append('')
                        body_part.append('')
                        # Append empty value to pass events
                        Pass_outcome.append('')
                        Pass_length.append('')
                        Pass_angle.append('')
                        Pass_end_x.append('')
                        Pass_end_y.append('')

                    # if Turnover
                    if event.attrib.get("type_id") == '9':
                        event_type.append('Turnover')
                        x.append(event.attrib.get("x"))
                        y.append(event.attrib.get("y"))
                        minutes.append(event.attrib.get("min"))
                        seconds.append(event.attrib.get("sec"))
                        period.append(event.attrib.get("period_id"))

                        team.append('')
                        player.append('')
```

```
                        # Append empty value to shot events
                        Shot_outcome.append('')
                        body_part.append('')
                        # Append empty value to pass events
                        Pass_outcome.append('')
                        Pass_length.append('')
                        Pass_angle.append('')
                        Pass_end_x.append('')
                        Pass_end_y.append('')

            possession_sequence = np.array(list(zip(event_type, team, player, period, minutes, seconds, x,
                                    y, Pass_end_x, Pass_end_y, Pass_length, Pass_angle,Pass_outcome,Shot_outcome, body_part)))

            fieldnames = ["Event","team", "player", "period", "min", "sec", "x", "y", "Pass_end_x", "Pass_end_y",
                          "pass length", "pass angle", "Pass outcome", "Shot outcome", "Shot body part"]

            file_name = "Possession_data_%s_%s_%s.csv" % (match_details["id"],match_details["home_team_name"], match_details["away_team_name"])
            pd.DataFrame(possession_sequence).to_csv(file_name, header=fieldnames, encoding='utf-8-sig') #already saved in output file
            print(file_name + ' is saved')

            df = pd.read_csv(file_name)
            df = df.rename(columns = {'Unnamed: 0':'index'})
            parse_each_team_sequence(df,match_details["away_team_name"],match_details["id"])
            parse_each_team_sequence(df,match_details["home_team_name"],match_details["id"])
        #  There will be three files generated for each match data, one is a intermediate csv file, two txt files are possession sequence files for home team and away team. Generated txt file are already stored in output data with file path
        # '..\Shengqiang Fan\Code and output\Output\Action evaluation\Phase 2 and 3\Possession sequence\possesion sequence of each team\test set
```