BACHELOR

Empirical Study of Temporal Networks

Ashran, Izz Faris

Award date:
2024

Awarding institution:
Tilburg University

Eindhoven University of Technology &
Tilburg University

Department of Mathematics and Computer Science
Tilburg Law School (TLS)

Bachelor Thesis Draft

Empirical Study of Temporal Networks

Izz Faris Ashran
Student Number: 1770608

Time frame: February 2024 - June 2024

Supervisors
Dr. George Fletcher
Dr. Nikolay Yakovets

Contents

1 Introduction **1**
- 1.1 Background . . . 1
  - 1.1.1 Networks and Temporal Networks . . . 1
  - 1.1.2 State of the Art Network Analysis . . . 2
- 1.2 Problem Statement . . . 2

2 Related Work **4**
- 2.1 Data Quality . . . 4
- 2.2 Quality Metrics . . . 4
- 2.3 Temporal Network Modeling . . . 5

3 Methodology **6**
- 3.1 Data Collection . . . 6
  - 3.1.1 Data Quality . . . 6
  - 3.1.2 Assessing Quality . . . 6
  - 3.1.3 Web Scraping . . . 8
- 3.2 Temporal Network Modeling . . . 9
  - 3.2.1 Modeling Framework . . . 9
- 3.3 Analysis . . . 10
  - 3.3.1 Programming Libraries . . . 10
  - 3.3.2 Analysis Metrics . . . 10

4 Results **12**
- 4.1 UEFA Champions League Data . . . 12
  - 4.1.1 Data Description . . . 12
  - 4.1.2 Temporal Network Construction and Analysis . . . 14
- 4.2 Celebrity Private Jets . . . 16
  - 4.2.1 Data Description . . . 16
  - 4.2.2 Temporal Network Construction and Analysis . . . 18

5 Discussion & Limitations **21**

6 Conclusion **23**

A Appendix **24**

Abstract

This thesis explores the application of temporal network analysis to two distinct datasets: the UEFA Champions League and celebrity private jet travel. By leveraging temporal network modeling techniques, the aim is to contribute to the scientific community by providing temporal network datasets and insights into patterns of connectivity, activity, and influence within these networks. The analysis focuses on key metrics such as temporal degree centrality, betweenness centrality, and closeness centrality to identify the most influential nodes and frequent interactions over time. The findings reveal differences in the construction of networks and highlight the potential of temporal networks in uncovering dynamic relationships in varied contexts. The results emphasises the importance of data quality and appropriate metric selection in temporal network analysis, offering a foundation for future research and practical applications in diverse fields.

1 Introduction

Temporal networks have emerged as a pivotal framework for modeling dynamic interactions across numerous domains, ranging from social networks to financial systems and beyond. Unlike static networks, temporal networks capture the unseen dynamics of connections evolving allowing individuals to understand trends and patterns over time.

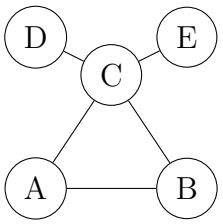
Understanding temporal networks is crucial for addressing a plethora of real-world problems, such as patterns in crime incidents, route optimization for transportation, etc.; however, despite the growing recognition of this field, empirical studies detailing the structural properties and dynamic behaviours of temporal networks remain relatively scarce. Further, there are few high-quality temporal networks available in the scientific community for study.

This thesis seeks to fill the gap by contributing to the collection and analysis of temporal aspects of network dynamics in a real-world context. By combining meticulous data curation and innovative modeling techniques, the thesis seeks to reveal hidden patterns, temporal dependencies, and emergent behaviours within dynamic systems while creating well-documented datasets.

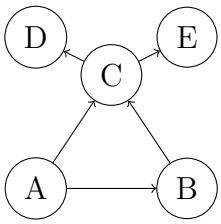
1.1 Background

1.1.1 Networks and Temporal Networks

Networks are defined as $G = \{V, E\}$ where V is the set of nodes (vertices) and E is the set of links (edges) whether directed or not. Given the number of N nodes, the network is uniquely defined by the $N \times N$ adjacency matrix A_{ij} indicating that there is a link from i to j: $A_{ij} = 1$ or $A_{ij} = 0$ otherwise for non-weighed networks.



(a) Undirected Graph



(b) Directed Graph

Temporal networks on the other hand are a subfield of network theory. where one treats the timing of when two vertices are in contact explicitly. A temporal network is any system that can be modeled, mathematically and computationally, as a graph of vertices with explicit timing of the contacts along edges [1].

The adjacency matrix would then be:

$$A(i, j, t) = \begin{cases} 1 & \text{if } i \to j \text{ connected at } t \\ 0 & \text{Otherwise} \end{cases}$$

Where t is continuous or discrete [2].

1.1.2 State of the Art Network Analysis

Until recently, most network studies have the time dimension projected out by aggregating the contacts between nodes to edges, even in cases when detailed information on the temporal sequences of contacts or interactions would have been available. A common approach to finding a solution is to divide the data into consecutive time intervals. During each interval, contacts between nodes are aggregated to form edges, and the network structure is analysed within these intervals. While this method allows for studying the evolution of the network structure over time, it may not capture all nuances of temporal contact patterns. For example, the relationships between nodes may not follow transitivity in temporal networks. Unlike static networks where indirect connections between nodes are established through intermediate nodes, temporal networks introduce a temporal dimension where the timing of the connections becomes crucial.

For example, temporal networks could have a reachability issue which refers to the fact that the ability of a node to reach another node depends not only on the existence of a path but also on the timing of the edges along the path. This is usually in the cases of information, disease, or influence spreading from one node to another. This also involves the factor that temporal networks are not transitive which means that even if A can reach B and B can reach C, it does not guarantee that A can reach C unless the temporal order of the interactions allows it. This emphasises the significance of temporal ordering, as the timings of connections and their correlations can have implications beyond what is captured in static networks. Consequently, this primarily focuses on methodologies that consider the temporal ordering of connections rather than simplifying the analysis by disregarding interaction so it depends on the analysis at hand as well [2].

1.2 Problem Statement

Coming back to the situation at hand, as stated before, there are not many high-quality datasets available for study. The collection of said data is a crucial step that needs to be taken to continue research. That is why there need to be ways to systematically assess the quality of the data to ensure that it is fit for use. Further, the goal of the thesis is also the analysis of the temporal networks. The details of networks were explained above in 1.1 and it is necessary to figure out the best way to model a temporal network from the high-quality data collected to be used for analysis. As such a research question and its sub-questions were identified:

"How can we systematically assess and ensure the quality of base datasets and model the temporal evolution of edges to accurately capture the timing of interactions across various domains?"

1. What criteria and metrics can be established to evaluate the quality of base datasets, and how can these be standardized across different domains to eventually be formed into temporal networks?

2. What modeling techniques best capture the temporal dynamics of edge formation and dissolution, incorporating the timing of interactions in temporal networks?

The research question and its sub-questions encapsulate the individual sections that need to be tackled to be successful. By answering these research questions, the goal of the thesis can be achieved and the descriptions of the datasets as well as results from the

analysis would be shared through an open-source repository to further contribute to the scientific community.

This thesis is structured as follows: first, an overview of relevant literature on temporal networks and data quality will be provided, highlighting key concepts and findings. Subsequently, The methodology will be addressed, defining data collection methods, the data quality and temporal network frameworks as well as analysis metrics. Following this, the empirical results are presented, accompanied by comprehensive discussions and limitations. Finally, a summary of the findings, implications for theory and practice, and avenues for future research will be concluded.

2 Related Work

2.1 Data Quality

Data collection involves many steps and the quality of the data can vary often. It can be frustrating when starting analysis on a particular dataset and issues start to occur since the quality of the data was not properly assessed and it is not fit for use. Thus, A framework to systematically assess if a dataset meets standards is necessary to avoid such issues. Extensive research has been done in the field of managing data quality. A paper by Wang and Guarascio in 1991 [3] elaborated on what data quality means; furthermore, the paper emphasises that quality should be defined concerning the consumers' needs and desires, not the producers. In the case of this thesis, the consumers would be the researchers who would use the data for further research. The focus of the paper was then to identify the dimensions of data quality which led to 20 data quality dimensions being identified. They include Accuracy, Timeliness, Completeness, etc.

In newer articles, the identification of the dimensions tends to vary depending on the use case of the researchers. They tend to also include fewer dimensions than what was proposed by Wang and Guarascio. In the use case of Fatimah et al., they showed how there is no one true set of dimensions [4, 5]. They used several sources to determine the dimensions they would use and it was 14 dimensions which included the ones mentioned above but as well as some others that Wang and Guarascio did not mention or excluded some that they did.

Finally, Cai and Zhu pointed out in their article the role of data quality as well as its assessment in the big data era. They have set up their framework consisting of big data quality dimensions, quality characteristics, and quality indexes. Further, they construct a dynamic assessment process for data quality which, according to them, has good expansibility and adaptability and meets the needs of big data quality assessment [6]. While the data they are referring to is big data there are still insights learned from the paper that could be useful in the context of the thesis. Nine dimensions were identified by the authors and this is fewer than the previous authors but there is also a lot of overlap with dimensions like accuracy, completeness, and timeliness where they repeat in all. While there is no one true set of dimensions for data quality, research does show that there is at least a commonly accepted and widely used quality standard that would then be redefined from their basic concepts based on actual needs.

2.2 Quality Metrics

The idea of the framework above relates mainly to qualitative assessments and there could be room for an even more thorough assessment of how high-quality a dataset is. That is where the quantitative assessments come in to, in hopes, of assisting in enhancing the assessment; however, in the paper by Heinrich et al., they state that even though both research and practice have realised the high relevance of well-thought-out data quality metrics, many of them still lack proper methodical foundation as they are developed for specific situations or exhibits subjectivity [7]. They propose a set of 5 requirements that data quality metrics should uphold to ensure reliable decision-making. This includes (R1) the existence of minimum and maximum metric values, (R2) the interval scaling of the metric values, (R3) the quality of the configuration parameters and the determination of the metric values, (R4) the sound aggregation of the metric values, and (R5) the economic efficiency of the metric. The study was focused on applying their requirements

to five metrics from the literature, specifically metrics covering timeliness, completeness, reliability, correctness, and consistency, to show that the presented requirements can be applied to various dimensions of data views and data values stored in an information system.

To point out some notable results, the authors use a metric for completeness by Blake and Mangiameli [8] and it managed to fulfill all requirements proposed. On the other hand, Hinrichs' [9] metric for correctness failed to fulfill but one requirement. Then the other three metrics would fall somewhere in between the two in terms of fulfillment. Heinrich et al. did succeed in showing that their requirements are not trivial nor impossible to fulfill; further, they pointed out which practical situations where specific requirements would be of particular relevance. Other researchers, however, claim a more general approach should be taken to assess the usefulness and validity of a data quality metric [10] [11].

2.3 Temporal Network Modeling

Collecting data for the network is only the beginning. The next challenge is understanding how to model the datasets into a proper temporal network. In conjunction with section 1.1.2, a common way of simplifying a temporal network is to turn it into a static network. The time factor is sometimes ignored by solely examining either the overall network of combined contacts or individual snapshot graphs that depict different time points [12]. Further, an alternative route is instead of reducing temporal network data to static networks, one can try to retain some but not all of the temporal features. Statistics of time between contacts is an example given in the research with the reason being that the networks are "bursty". This refers to the observation that the intervals between contacts or events in networks often follow heavy-tailed distributions. In simpler terms, it means that there are instances where there are long periods between contacts, followed by sudden bursts of activity where contacts occur more frequently. This phenomenon is commonly referred to as "bursty." So, instead of contacts being evenly spaced out over time, they tend to happen in irregular patterns, with some periods of inactivity followed by intense bursts of activity. Another method of simplifying temporal networks involves disregarding the dynamics of contacts and viewing links as existing between the initial and final observations of a contact in the data, without considering the exact timing of the contacts [12, 13].

Some research on static networks has been proposed to discover mesoscopic structures which are clusters, communities, or modules. These are loosely defined as groups of nodes more densely connected within than between each other. Most methods that integrate the time aspect into community detection typically operate on aggregated time intervals of contact sequences or networks of links that have occurred and are likely to recur. Additionally, reducing the network to a network of clusters that split and merge over time could be the most promising path in this direction; however, this reduction would then overlook any non-transitive features of the original structure, especially when time-slices or aggregation is involved. It would remove the effects of all temporal structures associated with shorter time scales rather than the time windows used [14].

3 Methodology

3.1 Data Collection

3.1.1 Data Quality

To begin with the methods section of the thesis, it would be natural to explain the data collection step. To answer the first research question some definitions of what high-quality would mean based on this thesis' criteria would need to be set. Firstly, based on an IBM article and what is most commonly used, data quality measures how well a dataset meets criteria for accuracy, completeness, validity, consistency, uniqueness, timeliness, and fitness for purpose. There are also some interrelated categories of criteria on how to view data like data integrity and data profiling. Data quality is the broader category that captures the criteria mentioned while data integrity involves a subset of these attributes, specifically accuracy, consistency, and completeness. Data profiling, on the other hand, focuses on the process of reviewing and cleansing data to maintain data quality standards [15].

The assessment of data quality involves considering various criteria, which can vary depending on the data source. These criteria mentioned above are used to classify metrics for evaluating data quality. Their details are as follows:

1. Completeness: This gauges the extent to which data is whole and usable. High levels of missing data may skew analysis results, making them unrepresentative.

2. Uniqueness: It measures the presence of duplicate data within a dataset. For example, each customer should have a unique identifier in customer data.

3. Validity: This assesses where data adheres to specified formats and comes from credible sources. Formatting usually includes metadata, like valid data types, ranges, patterns, etc.

4. Timeliness: Refers to the promptness of data availability within expected time frames. This would be important in, for example, real-time processes like order processing.

5. Accuracy: This focuses on the correctness of data values based on the agreed upon "source of truth," Since multiple sources report on the same metric, it is important to designate the primary data source.

6. Fitness for purpose: Lastly, this evaluates whether the data asset meets the intended requirements. This is particularly difficult to evaluate with new emerging datasets.

As stated before, there are many variations of the list of dimensions as there is no one fully agreed criteria but they are fairly adaptable to whatever situation it would be in. The selection of these 6 dimensions is due to it being one of the most common approaches to data quality standards and would encapsulate the scope and needs of the thesis.

3.1.2 Assessing Quality

With the use of the existing definitions of data quality metrics, a quality metric as the quantified measure of data quality dimension that gives relevant information about the

lack of quality regarding a certain information aspect was defined [16]. The quality metrics defined that would be used relate specifically to Completeness, Validity, and Uniqueness. The following quality metrics implemented constitute a set of widely applicable measures for profiling raw tabular data [16]. Timeliness would be defined differently based on the context of the thesis.

QM1: Completeness. The implementation utilised calculates the completeness of columns by assessing missing values within individual columns. A column entry, represented as $v_{col,row}$, is deemed "dirty" if it is either missing or marked as empty, typically indicated by a specific identifier like NaN.

$$Q_{comp}(v_{col,row}) = \begin{cases} 0 & \text{if } v_{col,row} = null \text{ or } v_{col,row} \in \{NaN, -, ...\} \\ 1 & \text{Otherwise} \end{cases}$$

QM2: Uniqueness. This is a simpler one. It allows for modularity where one can specify one or more columns that are expected to contain a unique combination of entries to check the dataset for duplicate entries.

$$Q_{unique}(col_m, ..., col_y) = \begin{cases} 1 & \text{if } \forall x \in M : M(x) = 1, \text{ for } M = \{\{x_i | x_i = (v_{col_{m,i}}, ..., v_{col_{y,i}}) \\ & \quad \text{for } i = 1...n\}\} \\ 0 & \text{Otherwise} \end{cases}$$

QM3: Validity. The default validity metric includes checking to evaluate if a data entry complies with the automatically detected or manually specified data type of the column. Additional domain-specific validity criteria can be gradually incorporated and improved based on existing knowledge of the dataset. This could also include the source of the data.

$$Q_{valid}(v_{col,row}) = \begin{cases} 1 & \text{if } typeof(v_{col,row}) = type, \text{ for } type \in \{integer, date, string, ...\} \\ 0 & \text{Otherwise} \end{cases}$$

QM4: Timeliness. This metric evaluates a specified time range and if they are in the correct format. The time range and format type would differ between datasets and it is to be determined based on what is available and what would best capture the granularity of a temporal network.

$$Q_{time}(v_{col,row}) = \begin{cases} 1 & \text{if } v_{col,row} \text{ is in format } F \text{ and } v_{col,row} \in D \\ 0 & \text{Otherwise} \end{cases}$$

Where F represents the set of valid timestamp formats, defined based on specific criteria such as YYYY-MM, etc.

D represents the date range or time range, defined based on specific criteria. For example, the date range could be January 2020 to January 2024 and the time range could be set from 9 am to 5 pm.

QM5: Accuracy. This metric can vary from dataset to dataset depending on the source and how it was obtained. For example, if the source was from comes from an archive of a reputable source, there might not be a particular way to measure the accuracy so the best approach could be to assume that the data collected is accurate and make

sure it fulfills the other dimensions. However, if the data is collected primarily then a simple metric could be used to measure the record level of accuracy based on what the reality value states [10, 17, 18].

$$\text{Record Level Accuracy} = \frac{\text{Number of Records Judged "Completely Correct"}}{\text{Number of Records Tested}}$$

QM6: Fitness for purpose. Finally, for fitness, there is no particular quantitative metric to use to assess this. However, within the context of the thesis, it could be the most important metric. The metrics above generally measure the quality of the data collected and ensure a proper analysis but this metric is the one that would determine if the data can be used for the analysis in the first place since if it does not meet the criteria to model a temporal network then there is no purpose for the thesis. This will be further elaborated in the next section as they correlate since the successful modeling of the temporal network from the data can determine its fitness.

3.1.3 Web Scraping

In addition to the quality metrics that need to be assessed, some tools are needed to collect the necessary data which in this case are web scraping tools. Web scraping is the automated process of extracting data from websites. It involves writing code to access and gather information from web pages, usually in HTML format, and then parsing that data to extract the specific pieces of information needed. While web scraping is not illegal, its legality can depend on factors like the terms of service of the website being scraped and the manner in which the data is used. So, it is also necessary to keep these in mind and is good practice to review the terms of service as well as respect the guidelines provided by the website that is being scraped.

The main tools that will be used are **BeautifulSoup** and **Selenium**.

BeautifulSoup is a Python library that allows parsing and scraping HTML and XML documents from the web, providing us with more options while navigating a hierarchical data tree. Once the scraping data is identified and the web structure is understood then BeautifulSoup can be used to quickly obtain that data. It is a very fast and easy tool to use and debug when trying to collect web data; However, when it comes to a web page that is more complex like a dynamic web page then that is where Selenium is more appropriate.

Selenium is an open-source framework commonly used for automating web applications for testing purposes. Selenium is also widely used for automating web scraping and other tasks involving browsers. As mentioned before Selenium is more robust and flexible so it can handle dynamic web pages which allows for access on websites where BeautifulSoup would not do as well. Selenium does have a steeper learning curve than BeautifulSoup but it is necessary when scraping certain web pages. It is also possible to combine the two to streamline work.

3.2 Temporal Network Modeling

As previously mentioned above in 1.1 about the definition of a temporal network, it is a system that could be modeled as a graph with additional information about when contacts happen, or the representation itself [1]. To simplify it further:

- **Node, Vertex**: One unit that interacts with others to form a temporal network.
- **Contact**: One interaction event, limited in time, between a pair of nodes.
- **Edge, Link**: A pair of nodes that at some point are in contact.

3.2.1 Modeling Framework

The modeling framework between domains can vary heavily. There is no one size fits all solution to modeling a temporal network and the nuances of each domain must be taken into consideration. This is the reason why trying to create temporal networks from data collected can be a challenging task. This ties in with section 3.1.2, specifically **QM6**, where one would need to clearly and carefully define nodes, edges, and contact points from the data to best capture the temporal attributes of the network. As such, the potential applications of temporal network modeling can come from many systems [2].

Person-to-person communication. Recordings of one-to-one communication like instant messaging platforms or e-mail messages are particularly suitable for a temporal network approach, especially in the context of the spreading dynamics of information or digital viruses. Such data can come in the form of lists of messages from one person to another at a point in time or dialog between two people within a time interval.

Travel and transport networks. Networks of transportation systems do work well with a temporal network modeling framework. The idea of mapping out networks of all sorts of modes of transport is a fairly common domain due to the nature of the type of data collected. Some work has been done in analysing networks of airline connections and another involves a study of a temporal network of subway travel in London. It is most common to use have nodes as people but in theory, the nodes can also represent vehicles too [19].

The systems above are by no means all the possible temporal networks that are studied. Thus, a general modeling framework is laid out to indicate a step-by-step method for temporal network modeling from the base data to the final product. The framework is as follows:

1. **Data Collection**: Start with collecting and curating a high-quality dataset from a particular domain using the metrics and dimensions defined in 3.1.
2. **Defining Nodes and Edges**: Identify the entities that can link to each other where the contact points are limited in time. This should align with state-of-the-art domain modeling but if there is a lack of that, a logical definition should suffice.
3. **Data Preparation**: Depending on the tool at hand some preparation might be necessary before construction. Attribute types or column names might need to be changed to fit the tool used.

4. **Temporal Network Construction**: Use a network tool to construct the network based on the data. It can come directly from a tabular file or an edge list can be created to be fed into the tool.

3.3 Analysis

3.3.1 Programming Libraries

For the scope of this thesis, the programming language Python would be used to perform any data cleaning, preprocessing, and analysis of the datasets collected. Python has access to a plethora of libraries that allow for easier data manipulation and network tools for analysis. The main libraries that would be used are as follows:

1. **PathpyG**. This is an open-source package that facilitates GPU-accelerated next-generation network analytics and graph learning for time series data on graphs. The package is tailored to analyse time-stamped network data as well as sequential data that capture multiple short walks or paths observed in graphs or networks. Examples of data that can be analysed include high-resolution time-stamped network data, dynamic social networks, passenger trajectories in transportation networks, etc. Pathpy is fully integrated with Jupyter, providing rich interactive visualisations of networks, temporal networks, and higher-order models. Visualisations can be exported to HTML5 files that can be shared and published on the Web. The theoretical foundation of this package was developed in several peer-reviewed research articles [20].

2. **NetworkX**. This library stands out as a great choice for network analysis due to its versatility, extensive functionality, and user-friendly interface. Offering a comprehensive suite of algorithms and tools for analysing graph data. It allows users to efficiently manipulate, visualise, and study complex networks across diverse domains. While it lacks built-in support for temporal aspects, its adaptable nature allows one to represent temporal features through extensions like timestamped edges or node attributes. Despite the primary focus on static networks, NetworkX's versatility makes it a viable option for analysing temporal networks with some additional customization and effort.

3.3.2 Analysis Metrics

Network analysis involves studying the structure, connectivity, and properties of networks to understand their behaviour and properties. Some key metrics that are commonly used for network analysis are centrality measures. They are usually comparatively easy to adapt from static networks. One can simply let the dynamic system evolve following the contacts rather than the links. This means that instead of considering fixed connections between nodes, the network structure changes over time depending on the interactions or contacts among nodes at various points in time. It allows for modeling dynamic systems where connections are temporary or changing, like social networks, epidemiological networks, or communication networks. By concentrating on contacts instead of fixed links, this approach captures the temporal dynamics of interactions, providing a more precise depiction of real-world network phenomena. The centrality measures that would be used are:

- **Temporal Degree Centrality.** The measure is an extension of the traditional degree centrality to temporal networks. This measures the number of connections a node has in the network over time, indicating its importance or prominence in the network. This equation is defined as [21]:

$$D_i^T = \sum_{j=1}^{N} \sum_{t=1}^{T} A_{i,j}^t$$

Where T is the number of time points, N is the number of nodes, and $A_{i,j}^t$ is the graph in that time point.

As stated in the paper, while this metric provides an estimate of how central or active a node is in a temporal network, the metric does not quantify the temporal order of the connections.

- **Temporal Betweenness Centrality**. This quantifies the extent to which a node lies on the shortest temporal paths between other nodes, considering the sequence and timing of interactions. This highlights the nodes that act as bridges or intermediaries in the network. The main difference between this and the static measure is that the static metric measures the network as a whole without considering any temporal aspects. This equation is defined as [22]:

$$B_i^t = \frac{1}{(N-1)(N-2)} \sum_{j=1, j\neq i}^{N} \sum_{k=1, k\neq i,j}^{N} \frac{\sigma_{jk}^i(t)}{\sigma_{jk}(t)}$$

Where if a shortest temporal path from j to k starts at t and passes through node i, then σ_{jkt}^i equals 1; otherwise, it is 0. The term σ_{jk} represents the total number of paths from j to k. The rest of the equation normalizes this value by the number of nodes.

- **Temporal Closeness Centrality**. This is an extension of the concept of closeness centrality but applied to temporal networks. It adapts this concept to consider the timing of interactions. It measures how quickly a node can reach all other nodes through temporal paths in the network. It takes into account the order and timing of connections, recognizing that interactions are not simultaneous and may only be possible at certain times. The equation is defined as [23]:

$$C_{T_i} = \frac{1}{N-1} \sum_{j} \frac{1}{\tau_{ij}}$$

Where τ_{ij} is the average temporal distance between i and j and N the number of nodes.

These are the main metrics that will be used for analysis with each dataset collected for the thesis.

4 Results

4.1 UEFA Champions League Data

4.1.1 Data Description

The first data set collected was the UEFA Champions League fixtures throughout the years of 1992-2024. This dataset was obtained by scraping Transfermarkt.com under the Champions League web page [24]. BeautifulSoup was used here to navigate the HTML code to extract the necessary features. It was important to understand how the structure of the HTML code is laid out as BeautifulSoup requires the classes and tags of the elements one would want to extract. Once the elements are identified and collected, cleaning the text is a must. The raw text collected by BeautifulSoup is filled with unwanted HTML code or unnecessary text so one must strip the text of all the junk to get the final desired description. After initial processing, the collected lists are then put into a pandas data frame. The function used to scrape the data would be looped through the years and the rounds of the competition which would then concatenate all the data frames to make the final data frame containing all the years and rounds. Once concatenated, the final data frame is further processed to clean some formatting issues and add additional columns that could provide more in-depth analysis. The final description of the data frame is as follows:

- **Date**: The date represents the day the games were played. The initial state of this column was a string but it has been changed to a DateTime object in the format of YYYY-MM-DD.
- **Home Team**: This represents the team playing on their home ground. It is in string format.
- **Away Team**: This represents the team traveling to play at the home team's ground. It is in string format.
- **Result**: Represents the final result of the game played. The format is a string but can be manipulated further.
- **Stage**: This is at what stage the games were played. It begins with the Group stage, then the knockout rounds which include the Last 16, Quarter Finals, Semi-Finals, and Finals.
- **Scoresheet Home**: This includes the minutes within the game when goals were scored on the home side and are in a list. If it is empty then that means no goals were scored by the home team. A penalty and an opposition's own goal count towards this too.
- **Scorer Home**: This includes the names of the people who score goals on the home side and are in a list. If the list is empty then no one scored by the home team. Names of the players in the opposition that scored an own goal will also be included here.
- **Scoresheet Away**: This includes the times within the game when goals were scored on the away side and are in a list. If the list is empty then that means no

goals were scored by the away team. A penalty and an opposition's own goal count towards this too.

- **Scorer Away**: This includes the names of the people who score goals on the away side and are in a list. If the list is empty then no one scored by the away team. Names of the players in the opposition that scored an own goal will also be included here.

- **Home Outcome**: This indicates whether the outcome on the home ground is a win, loss, or draw. It is in string format.

Since the description of the data has been clarified, the quality metrics previously mentioned in 3.1.2 need to also be addressed and see if the dataset is considered high quality.

- **QM1**: The dataset contains no null values and can be considered as complete as per collection period.

- **QM2**: The dataset contains no duplicate values and thus is unique.

- **QM3**: As stated in the data description, the format of the columns is appropriate to what is described in the column. The data is also collected from a reputable source.

- **QM4**: The formats of the dates is uniform to YYYY-MM-DD and in the DateTime object. The range of years is also complete as collected from the website.

- **QM5**: Initially, it was discovered there were missing rows in the data. After manually cycling through the years to check if the number of games played in that season matched the number of games in the dataset for that season, it was found that the seasons (97/98), (10/11), and (22/23) were problematic and were missing certain matches. Further, the naming of the rounds changed in some years and that was also mistakenly missed but all was fixed once updating the web scraping code. 50 random samples were tested by cross-referencing the website and 100% of the values match.

- **QM6**: The dataset collected has all the necessary attributes that are needed to create a temporal network. These can include the Date the games were played, the Home Team, and the Away team.

Based on the metrics, the dataset collected is of high quality and can be used in analysis. The final dataset contains 3712 rows and 11 columns.

	Teams	**Result**	**Stage**	**HomeOutcome**
Count	3712	3712	3712	3623
Unique	175	58	26	3

Table 1: Summary Statistics of Main Columns in the Football Dataset

4.1.2 Temporal Network Construction and Analysis

For the construction of the temporal network, NetworkX was adapted to be able to be fed a temporal network formulation. In the most basic sense of the network, the attributes can include the Date, Home Team, and Away Team. Adding more attributes can enhance the analysis and understanding further by providing more context and details of the nodes and edges. Hence, any combination of the entire dataset's columns can be used to provide additional context, depending on the use case. Some use cases and attribute combinations can include:

1. **Match Analysis and Outcome Prediction**:
    - Attributes: Date, Home Team, Away Team, Result, Stage, Home Outcome
    - Analysis: Predict the outcome of upcoming matches based on historical data. The temporal trends can be analysed to understand team performance over time.
    - Temporal Aspect: Analyse how teams perform in different stages and how performance changes over time.
2. **Team Dynamics and Rivalries**:
    - Attributes: Date, Home Team, Away Team, Result, Stage
    - Analysis: Investigate the dynamics between specific teams, identifying patterns in rivalries and outcomes. Understand which teams tend to wind against specific opponents.
    - Temporal Aspect: Examine the temporal aspect of rivalries, how often teams meet, and the outcomes over different seasons.
3. **Goal Timing and Impact**:
    - Attributes: Date, Scoresheet Home, Scoresheet Away, Result, Stage, Home Outcome
    - Analysis: Analyse the timing of the goals and their impact on the match outcomes. Determine if there are critical periods in a match where goals are more influential.
    - Temporal Aspect: Study the distribution of goals within a match and across different stages of the competition.

Some exploratory visualisations about the distribution of attributes related to these examples are in Appendix A.

The statistics of the basic undirected network of just Home and Away teams as nodes and the Date as the edge are:

Table 2: Basic Network Statistics for Football

Statistics	**Value**
Nodes	175
Time-stamped links	3712
Observation period	[16-09-1992, 01-06-2024]

Below is also an illustration of a subset of the network:

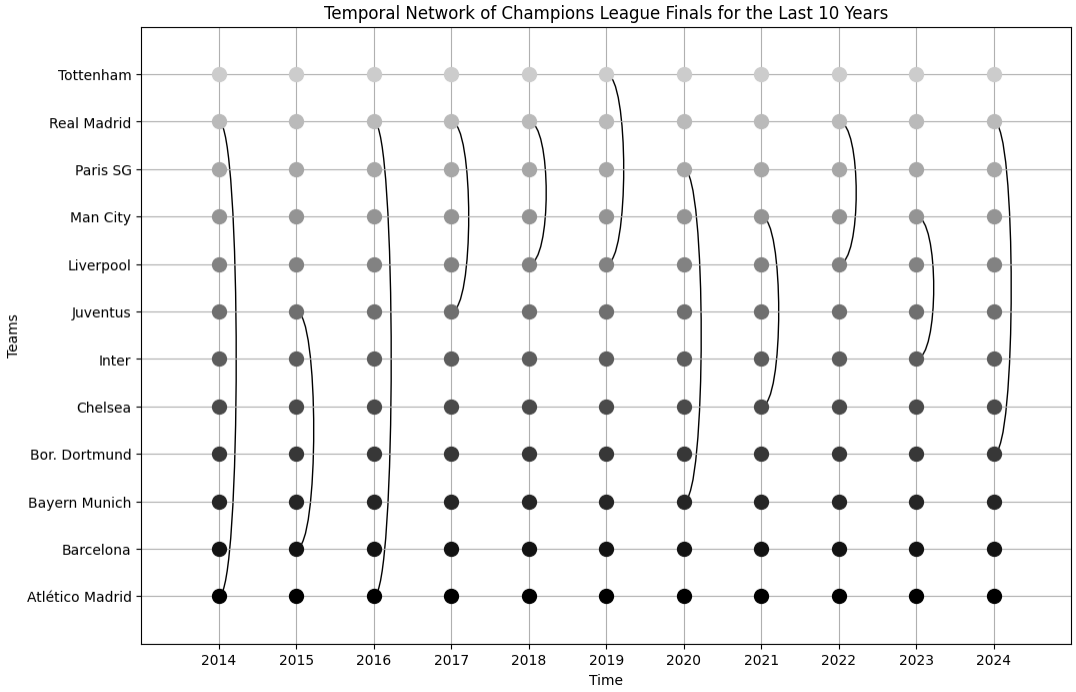


Figure 2: Connections of the finals for the last 10 years

This shows the interaction between teams and it is possible to infer which team is more consistent in their ability to reach the final.

As for the analysis, the metrics in 3.3.2 were applied to this definition of a temporal network to gain some basic insight into the network.

Table 3: Basic Network Analysis for Football

Metrics	**Value (Avg)**
Temporal Degree Centrality	28.27
Temporal Betweenness Centrality	0.011
Temporal Closeness Centrality	5.096131e-10

Table 3 shows the average values of the degree, betweenness, and closeness centralities. The metrics were calculated over 10 years, 2014-2024, rather than the whole dataset due to the running time. The average degree of the dataset is 28.27 with Real Madrid being the top node with 117 connections over time and Bayern Munich coming second with 107 connections. A high degree value indicates that a club has been more actively involved or engaged in temporal interactions compared to others. In this case, it implies a higher frequency of matched played which makes sense since Real Madrid has been one of the most successful clubs in the last ten years. Further, average the betweenness centrality is fairly low with a value of 0.011. A higher betweenness value indicates that a club acts as a critical bridge between other clubs in terms of temporal interactions. Once again, Real Madrid is the leader in this with a value of 0.089 and it suggests that the club plays a significant role in connecting the paths so it seems to be that Real Madrid

appears between nodes as an opponent more often than others. The top 10 teams for the betweenness and degree generally stay the same but they do shift in positions amongst each other. Finally, the average closeness centrality is a very small number and with little variation in general. This mainly refers to how quickly a node can reach other nodes so a higher closeness centrality indicates that a node can reach other nodes more quickly on average. This might not be as applicable to this interpretation of a network but it is still worth analysing. Unlike the previous centralities, the team with the highest value is Liverpool with a value of 6.599599e-10 but do share the value with other teams. The difference between teams is almost negligible which also makes sense since many teams would directly interact with each other. So, the teams that have the same value as Liverpool suggest they have a similar average temporal distance to all other teams. The difference between values would indicate either a marginally better or worse average temporal reachability compared to others. The full tables of the top 10 teams are below in the Appendix A.

4.2 Celebrity Private Jets

4.2.1 Data Description

The second dataset collected was a few celebrity private jet activity data scraped from a website called CelebrityPrivateJetTracker.com [25]. Unlike the UEFA Champions League dataset, Selenium was used in addition to BeautifulSoup due to the more dynamic nature of the web page. It uses the Selenium WebDriver component to access the URL then BeautifulSoup is used to parse the source code. It follows the same procedure as the UEFA Champions League data where it needs the tags and classes to be able to search the elements. It also has the same procedures for obtaining the final data frame but it would search for 10 private jets which include two of Taylor Swift's Jets (Classified as 1), two of Michael Bloomberg's Jets (Classified as 1), Donald Trump's, the Nike Corporation's, Elon Musk's, Drake's, Kylie Jenner's, and Kim Kardashian's jet. The description of the data is as follows:

- **Date**: This represents the date on which the flight took place. This is a DateTime object in the format of YYYY-MM-DD.
- **Departure**: This is the airport where the flight took off. The values are in string format.
- **Arrival**: This represents the destination airport of the flight. It is in string format.
- **Distance**: This is the distance of the flight in miles. It is represented in integer format.
- **Flight Time**: This is the duration of the flight in minutes. This was first converted from a string to an HH:MM format to have a uniform format and later converted to strictly minutes representation in an integer format.
- **Fuel**: This represents the fuel used throughout the flight in gallons. This was converted from a string to an integer format.
- **Carbon Emissions**: The column represents the carbon emissions produced by the flight. This was converted from a string to an integer format.

- **Celebrity**: This is the owner of the jet that flew and is in string format.
- **Departure Code**: This represents the airport code of the departure location. It is in string format and was split from the Departure column.
- **Arrival Code**: This represents the airport code of the arrival location and is in string format and was split from the Arrival column.

As in the previous dataset, the quality metrics need to be checked.

- **QM1**: The dataset was missing one value in the Departure column on 13-04-2023 and also missing one value in the Arrival column on the same date. These rows were dropped as a result since without one or the other analysis could not be performed. The resulting dataset does not have any null values and is now complete.
- **QM2**: There are no duplicate rows in the dataset making it comply with uniqueness.
- **QM3**: As mentioned in the description of the data, certain columns needed to be converted and the resulting data set has appropriate data types for its category. The source of the data is also reputable as the data collected on the website was taken from three websites known for tracking flights. These websites include airplanes.live, ADSB.fi and ADSB Exchange.
- **QM4**: The dates are in a uniform format and a DateTime object. There is no particular date range as the data collected are whatever dates that are stored for each private jet tracking from a different start time each.
- **QM5**: Other than the NaN values in some columns due to the website itself, there are no missing values after iterating over the number of flights from each celebrity and cross-referenced with the website. A random sample of 50 flights was taken and checked if the values matched and 100% of the cases matched the website's values.
- **QM6**: The data collected the necessary attributes to create a temporal network and to perform analysis. This includes the Date, Departure, and Arrival mainly. It is also possible to slice the dataset by celebrities to monitor their flights specifically.

Based on the metrics, the dataset can be considered high quality and can be used to perform analysis. The final dataset contains 1263 rows and 10 columns.

	Departure	**Arrival**	**Celebrity**
Count	1263	1263	1263
Unique	239	242	10

Table 4: Summary Statistics of the Main Columns in the Flights Dataset

4.2.2 Temporal Network Construction and Analysis

Just as with the previous dataset, NetworkX was adapted to create a temporal network. The most basic form of the network is to include the Date, Departure, and Arrival as attributes where Departure and Arrival are the nodes with the Date linking them. As stated prior, adding more attributes can enhance the analysis and understanding of the network. Thus, some use cases and attribute combinations are:

1. **Flight Pattern Analysis**:
    - Attributes: Date, Departure, Arrival, Celebrity
    - Analysis: Investigate flight patterns of celebrity flight routes over time.
    - Temporal Aspect: Examine temporal trends in the destinations they frequent, the airports they prefer, and any seasonal variation in their travel behaviour.
2. **Optimization of Flight Routes**:
    - Attributes: Date, Departure, Arrival, Flight Time, Distance, Celebrity
    - Analysis: Optimizing flight routes to minimize travel time for a celebrity
    - Temporal Aspect: Analysing temporal patterns of past flight data to understand how flight durations have changed over time.
3. **Fuel Efficiency Study**:
    - Attributes: Date, Departure, Arrival, Distance, Fuel, Carbon Emissions
    - Analysis: Investigating the relationship between flight distance, fuel consumption, and carbon emissions over time.
    - Temporal Aspect: Analysing temporal trends of past flight data to reveal insights of how fuel consumption and carbon emissions changed over time.

Some exploratory visualisations about the distribution of attributes related to these examples are in Appendix A.

The statistics of the basic directed graph of the Departure and Arrival as nodes and the Date as the edge attribute are:

Table 5: Basic Network Statistics for Private Jets

Statistics	**Value**
Nodes	252
Time-stamped links	1263
Observation period	[01-02-2023, 30-05-2024]

Below is also an illustration of a subset of the network:

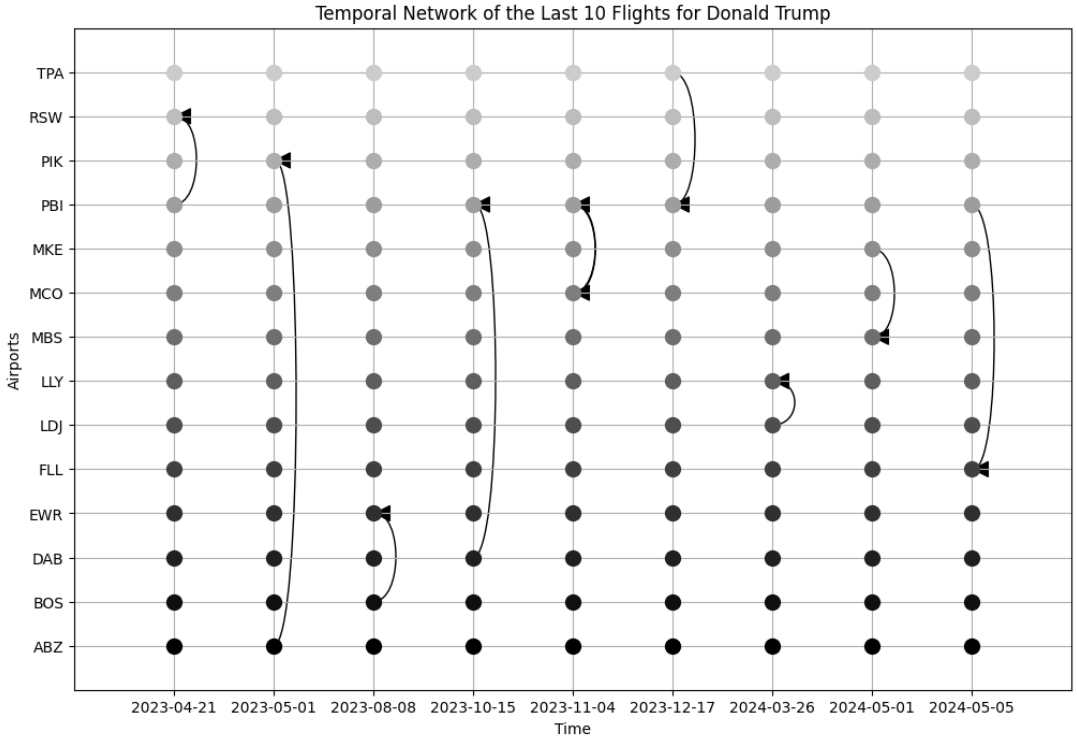


Figure 3: Connections of Donald Trump's Last 10 Flights

This shows the directed graph of Donald Trump's last 10 flights. It can be seen where his origin was and where he traveled on a given date.

As in the previous section, the temporal metrics are also applied here for some basic insight into the network.

Table 6: Basic Network Analysis for Private Jets

Metrics	**Value (Avg)**
Temporal Degree Centrality	5.28
Temporal Betweenness Centrality	0.0051
Temporal Closeness Centrality	3.029812e-10

Table 6 shows the average values of the three metrics applied. The metrics were calculated over the entire dataset. The average degree is 5.28 with La Guardia Airport (LGA) being the airport with the highest degree of 126. In this case, it implies that La Guardia Airport has had a higher volume of activity during that period whether it be arrivals or departures. The average betweenness is a small value of 0.0051 where the betweenness in this context represents the importance of each airport in facilitating temporal connections between other airports within the period. Camarillo Airport (CMA) has the highest betweenness of 0.143 and it indicates that CMA plays a crucial role in connecting airports within the temporal network. It frequently lies on the shortest paths between other airports, making it an important intermediary in facilitating temporal interactions.

The small average value is likely due to the many infrequent airports that these celebrities travel to. Finally, the average closeness is 3.029812e-10 with Rifle Garfield County Airport (RIL) having the highest closeness of 5.370116e-10. In this context, it represents how quickly an airport can reach all other airports in the network through temporal paths. This measures the average temporal distance from an airport to all other airports, taking into account the timing and sequence of interactions. RIL having the highest closeness centrality indicates it has high efficiency, relatively, in reaching other airports in the temporal network. It means that on average RIL is close to all other airports in terms of temporal paths. The full tables of the top 10 highest centralities airports are in the Appendix A.

The link to access the code and datasets behind these results for reproducibility is here: https://github.com/izzfaris/BEP-Empirical-Study-of-Temporal-Networks.

5 Discussion & Limitations

The results in the previous section touch upon each aspect of what this thesis wants to achieve. Half of the work in the thesis was to collect high-quality datasets from various domains to encourage further research in this field which was successful as the two datasets collected contained the core attributes of a temporal network while also including other metadata for researchers to use in any way they can think of. The quality metrics established indicated that it was possible to qualitatively and quantitatively determine whether a dataset is considered high-quality or not. It was also made general enough to be able to fit the criteria to almost any dataset for a holistic evaluation. With the data passing through quality checks, the construction of a network from it is fairly simple. The general modeling framework defined provides a fundamental understanding of how to construct a network but it does require some domain knowledge of the dataset at hand as certain identification of nodes, edges, and other attributes can differ depending on the task and goals that one wants to achieve. In the context of the datasets collected, some examples of interesting analyses have been provided but those are by no means all the possible use cases and it is up to users of the data to decide what they want to discover from the dataset. For the analysis section, the temporal versions of the basic centralities were applied and developed from scratch using NetworkX due to the finicky nature of the open-source libraries. An algorithm to calculate the reachability latency was also developed but it is not applicable for these datasets as it is more useful in disease or information-spreading situations, for example. It was developed with literature in mind, as defined in [21], so it could theoretically be put to use but not for this case.

The results of obtaining the datasets, providing simple analysis, and releasing the data onto an open repository could stimulate new research in the field whether it be using the datasets for theoretical work or improving the current analysis tools available.

While the data collection section allows for much experimentation of ways to collect data, the analysis of the temporal networks can be fairly challenging due to how under-researched this topic is. The open-source tools available online for analysis are scarce or limited to specific software. This led to the use of sub-optimal tools like PathpyG which does allow for some form of analysis on a temporal network but is very restrictive on how a temporal network is modeled. It only allows for the connecting nodes and integer time attributes as its inputs so its inability to insert additional attributes might lead to the analysis not being able to fully capture the complexities and dynamics of a temporal network. In the end, PathpyG was not used for network modeling since it was much easier to use NetworkX to construct a temporal network due to its flexible nature of being able to use extensions to represent temporal features. It was also much easier to develop the metrics using NetworkX based on literature rather than using pre-made metric functions from the libraries because the inputs that those functions wanted were unclear and even if it did run it was difficult to interpret. Thus, the use of PathpyG was mainly reduced to a visualisation tool since it showed a decent dynamic plot. However, one drawback with the metrics is that the running time for temporal betweenness is fairly high for large networks with many edges. The function was optimised with parallelism to have a 50% reduction in running time compared to without parallelism and as the graph size increases, in terms of nodes and edges, the time complexity increases significantly. So it highlights the importance of efficient parallelisation and potentially exploring more

scalable algorithms or optimisations but this is currently the best that was developed within the scope of the thesis.

About open-source libraries, while developing and maintaining such a tool requires a lot of time and effort from many contributors even perhaps, it is an area of the field that could be further developed in a future project or by other researchers. As much as providing datasets is important, the availability of more open-source analysis tools will allow for a bigger impact on the field and more innovative research can be performed. An example of such an impact is in the realm of machine learning with Python libraries like scikit-learn allowing a larger set of individuals to develop their predictive models with more ease. It led to a lot of research being put into the library itself and the field as a whole which is why it is crucial to have such tools publicly available to see a boom in research.

6 Conclusion

In conclusion, the thesis successfully demonstrates the application of temporal network analysis in analysing dynamic interactions within two distinct datasets: the UEFA Champions League and Celebrity Private Jets. The analysis highlights how temporal network metrics can uncover key details about node influence, interaction frequency, and network structure over time. For example, Real Madrid is seen as a highly connected node in the football network, reflecting its consistent performance and engagement in the league. Similarly, the private jet analysis identified patterns in celebrity travel behaviour.

These findings have important implications for both theoretical and practical applications. Theoretically, they contribute to a deeper understanding of temporal network dynamics, emphasising the need for robust data quality and tailored metric selection. Practically, the ability to manipulate the data by including different attributes or having different sources and targets can help inform strategies in sports management, celebrity marketing, and beyond.

Future research should expand on this work by incorporating additional datasets and exploring more advanced modeling techniques. Further, addressing limitations such as computational constraints or the lack of well-documented libraries to work with will enhance the accuracy and applicability of temporal network analysis. Overall, the thesis brings more light to the already emerging topic of temporal networks and shows the potential of temporal networks as a powerful tool for analysing dynamic systems in various domains.

A Appendix

Table 7: Temporal Degree Centrality of Football Clubs

Teams	**Value**
Real Madrid	117
Bayern Munich	107
Man City	105
Paris SG	95
Barcelona	94
Atlético Madrid	92
Juventus	84
Bor. Dortmund	77
FC Porto	74
Liverpool	71

Table 8: Temporal Betweenness Centrality of Football Clubs

Teams	**Value**
Real Madrid	0.088534
Bayern Munich	0.070074
Man City	0.063308
Chelsea	0.055341
Atlético Madrid	0.055161
Liverpool	0.050184
FC Porto	0.049964
Barcelona	0.048994
Paris SG	0.047908
Juventus	0.045873

Table 9: Temporal Closeness Centrality of Football Clubs

Teams	**Value**
Liverpool	6.466294e-10
Ludogorets	6.466294e-10
Real Madrid	6.466294e-10
FC Basel	6.466294e-10
Chelsea	6.464728e-10
FC Schalke 04	6.464728e-10
Sporting CP	6.464728e-10
NK Maribor	6.464728e-10
Bayern Munich	6.447396e-10
Man City	6.447396e-10

Table 10: Temporal Degree Centrality of Jet Travel

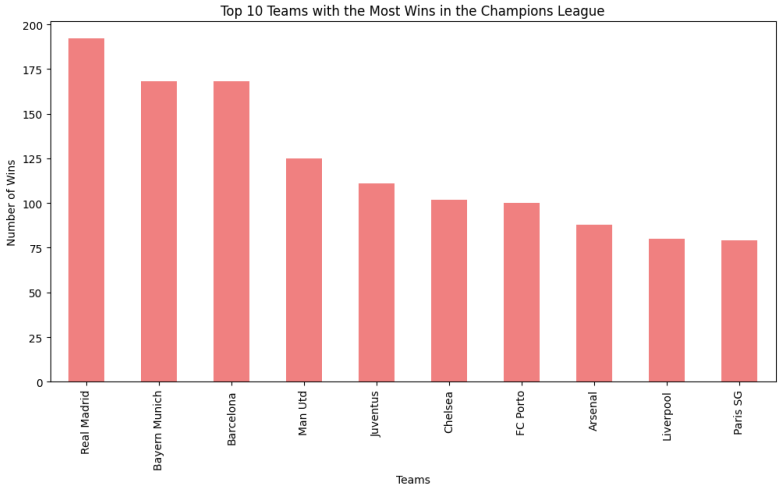
Airports	**Value**
LGA (LaGuardia Airport)	126.0
MMU (Morristown Municipal Airport)	111.0
CMA (Camarillo Airport)	106.0
PBI (Palm Beach International Airport)	69.0
VNY (Van Nuys Airport)	48.0
HPN (Westchester County Airport)	43.0
HIO (Hillsboro Airport)	33.0
TEB (Teterboro Airport)	32.0
AUS (Austin-Bergstrom International Airport)	31.0
TRM (Jacqueline Cochran Regional Airport)	24.0

Table 11: Temporal Betweenness Centrality of Jet Travel

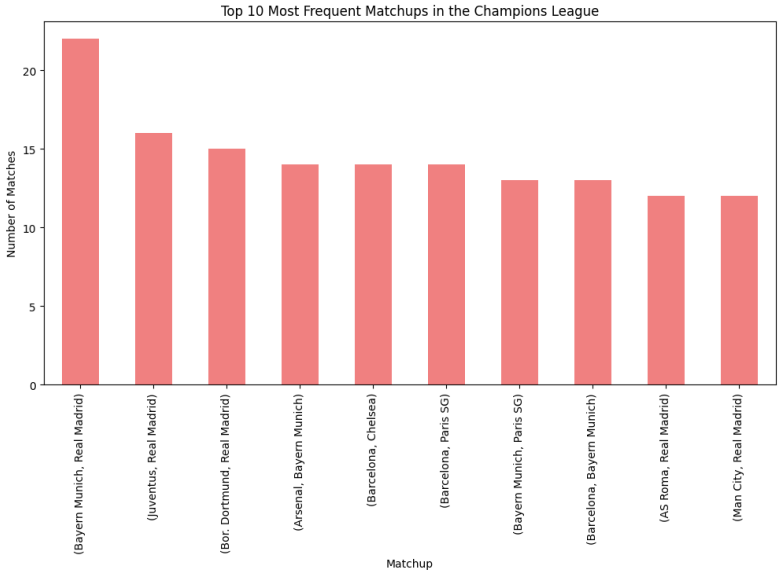
Airports	**Value**
CMA (Camarillo Airport)	0.142745
PBI (Palm Beach International Airport)	0.123678
LGA (LaGuardia Airport)	0.112985
MMU (Morristown Municipal Airport)	0.089380
AUS (Austin-Bergstrom International Airport)	0.059522
BNA (Nashville International Airport)	0.044499
LTN (London Luton Airport)	0.042087
VNY (Van Nuys Airport)	0.041371
LAS (McCarran International Airport)	0.037424
EWR (Newark Liberty International Airport)	0.031909

Table 12: Temporal Closeness Centrality of Jet Travel

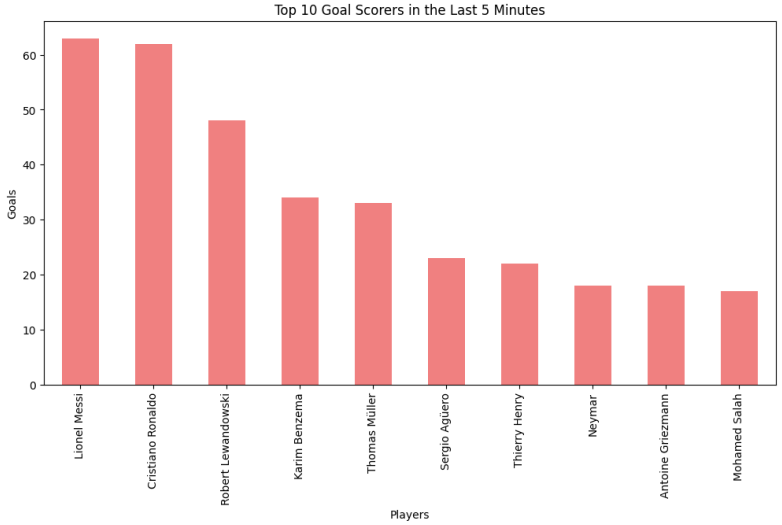
Airports	**Value**
RIL (Rifle Garfield County Airport)	5.370116e-10
CMA (Camarillo Airport)	5.321302e-10
YYZ (Toronto Pearson International Airport)	5.320967e-10
VNY (Van Nuys Airport)	5.320966e-10
CEC (Del Norte County Airport)	5.320887e-10
BUR (Bob Hope Airport (Burbank))	5.296328e-10
AUS (Austin-Bergstrom International Airport)	5.296306e-10
OAK (Oakland International Airport)	5.296301e-10
TEB (Teterboro Airport)	5.225149e-10
PHL (Philadelphia International Airport)	5.225148e-10



(a) Teams with the most wins

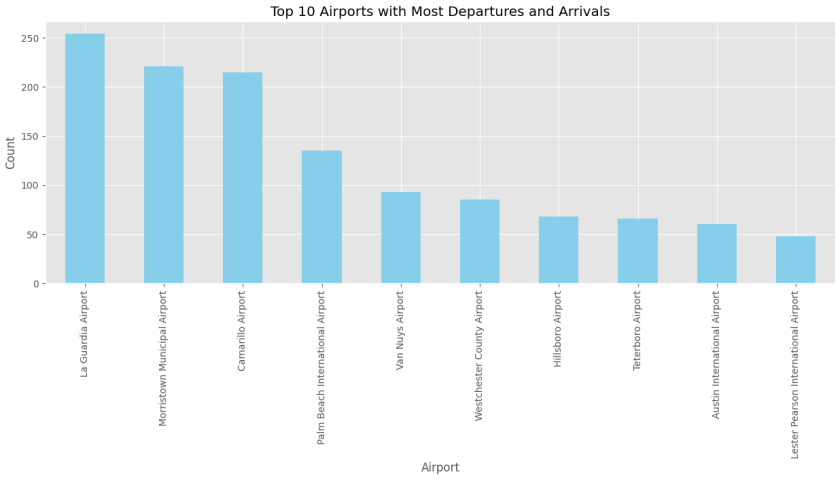


(b) Most common matchups

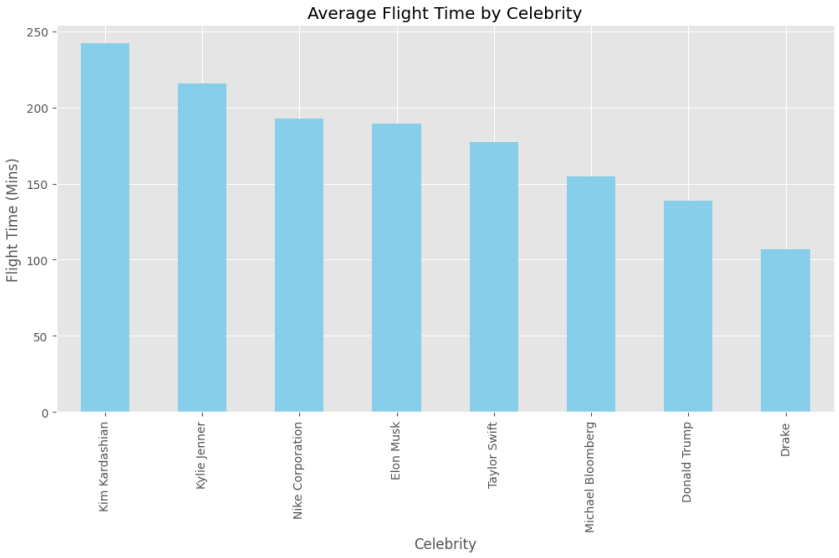


(c) Top scorers for late goals

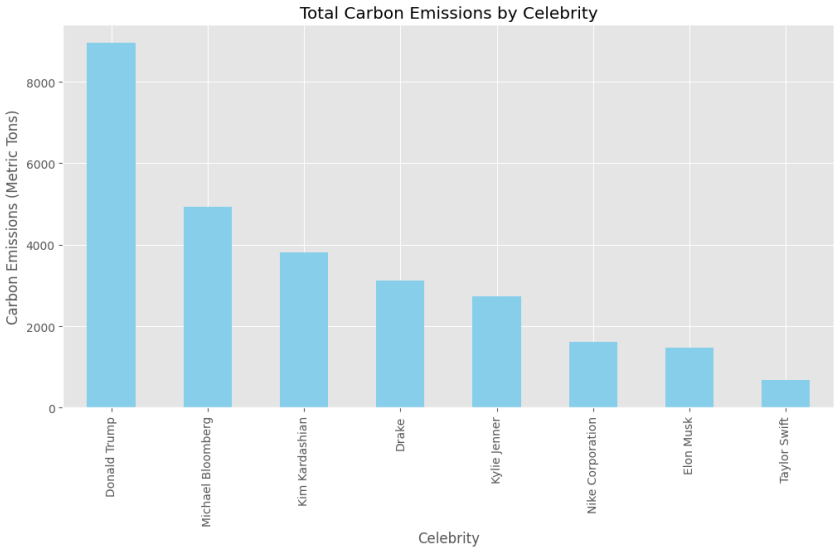
Figure 4: Exploratory plots for Champions League Dataset



(a) Most Frequented Airports



(b) Average Flight Time per Celebrity



(c) Most Carbon Emissions per Celebrity

Figure 5: Exploratory plots for Private Jet Dataset

References

[1] Petter Holme. *Temporal networks*. 1 2014.

[2] Petter Holme and Jari Saramäki. Temporal networks. *Physics Reports*, 519(3):97–125, 2012. Temporal Networks.

[3] Richard Y. Wang and Lisa M. Guarascio. Dimensions of data quality: Toward quality data by design. 1991.

[4] Fatimah Sidi, Payam Hassany Shariat Panahy, Lilly Suriani Affendey, Marzanah A. Jabar, Hamidah Ibrahim, and Aida Mustapha. Data quality: A survey of data quality dimensions. In *2012 International Conference on Information Retrieval Knowledge Management*, pages 300–304, 2012.

[5] Yair Wand and Richard Y. Wang. Anchoring data quality dimensions in ontological foundations. *Commun. ACM*, 39(11):86–95, nov 1996.

[6] Li Cai and Yangyong Zhu. The challenges of data quality and data quality assessment in the big data era. *Data science journal*, 14(0):2, 5 2015.

[7] Bernd Heinrich, Diana Hristova, Mathias Klier, Alexander Schiller, and Michael Szubartowicz. Requirements for data quality metrics. *ACM journal of data and information quality (Online)*, 9(2):1–32, 6 2017.

[8] Roger Blake and Paul Mangiameli. The effects and interactions of data quality and problem complexity on classification. *J. Data and Information Quality*, 2(2), feb 2011.

[9] Holger Hinrichs. *Datenqualitätsmanagement in Data Warehouse-Umgebungen*. 1 2001.

[10] Lisa Ehrlinger and Wolfram Wöß. A survey of data quality measurement and monitoring tools. *Frontiers in big data*, 5, 3 2022.

[11] Antoon Bronselaer, Robin De Mol, and Guy De Tré. A measure-theoretic foundation for data quality. *IEEE Transactions on Fuzzy Systems*, 26(2):627–639, 2018.

[12] Petter Holme and Jari Saramäki. *A Map of Approaches to Temporal Networks*, pages 1–24. Springer International Publishing, Cham, 2019.

[13] Petter Holme and Fredrik Liljeros. Birth and death of links control disease spreading in empirical contact networks. *Scientific reports*, 4(1), 5 2014.

[14] Petter Holme and Jari Saramäki. *Temporal networks as a modeling framework*. 1 2013.

[15] What is data quality? — IBM.

[16] Christian Bors, Theresia Gschwandtner, Simone Kriglstein, Silvia Miksch, and Margit Pohl. Visual interactive creation, customization, and analysis of data quality metrics. *Journal of Data and Information Quality*, 10(1):1–26, 3 2018.

[17] Thomas Redman. Measuring data accuracy: A framework and review. *Information Quality*, pages 21–36, 01 2005.

[18] Carlo Batini and Monica Scannapieco. *Data Quality: Concepts, Methodologies and Techniques*. 01 2006.

[19] Petter Holme. Modern temporal network theory: a colloquium. *The European physical journal. B, Condensed matter physics/European physical journal. B, Condensed matter and complex systems*, 88(9), 9 2015.

[20] Ingo Scholtes and Luca Verginer. PathpyG.

[21] William Hedley Thompson, Per Brantefors, and Peter Fransson. From static to temporal network theory: Applications to functional brain connectivity. *Network neuroscience*, 1(2):69–99, 6 2017.

[22] John Tang, Mirco Musolesi, Cecilia Mascolo, Vito Latora, and Vincenzo Nicosia. Analysing information flows and key mediators through temporal centrality metrics. *Research Gate*, 4 2010.

[23] Raj Kumar Pan and Jari Saramäki. Path lengths, correlations, and centrality in temporal networks. *Physical review. E, Statistical, nonlinear and soft matter physics*, 84(1), 7 2011.

[24] Transfermarkt. Uefa champions league schedule. `https://www.transfermarkt.com/uefa-champions-league/gesamtspielplan/pokalwettbewerb/CL/saison_id/`, 2024. Accessed on: June 12, 2024.

[25] Celebrity Private Jet Tracker. Celebrity private jet tracker. `https://celebrityprivatejettracker.com/`, Year of Access. Accessed on: June 5, 2024.