**CHAPTER 3**

**METHODOLOGY**

## 3.1 DATASET

## 3.1.1 DATASET USED

Twitch is used widely by gamers to live-stream themselves while playing games. The nature of the platform is such that there are few popular gamers with many followers. We obtained the dataset from the Stanford Large Network Dataset Collection of more than 50 large network datasets from tens of thousands of nodes and edges to

tens of millions of nodes and edges. We chose the twitch dataset as there

has not been much link prediction work done on this previously.

## 3.1.2 UNDERSTANDING THE DATA

The dataset was reasonably sized with 7126 nodes. The total number of possible edges in the network is 50,772,750 from which 35,324 are present in the network. Using all the missing edges from the graph would highly skew the dataset so we randomly sampled 35,324 missing edges. While sampling these missing edges, we only consider an edge as missing if the distance between the source and destination was more than 2 as closely connected users are likely to be mutual friends even if an edge does not already exist. This is to ensure the model is able to properly distinguish between present and absent edges, thus improving its performance. We used this presence or absence of an edge as the target class variable for prediction.

## DATA VISUALIZATION

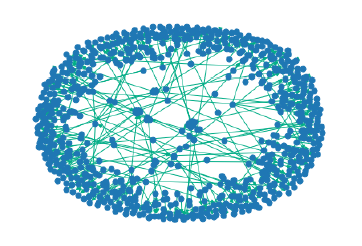


Figure 3.1 visualization of a random sample of 500 edges from the dataset

## 3.2 ASSUMPTIONS

* Using all the missing edges from the graph would highly skew the dataset so we are planning to randomly sampling few missing edges to avoid skewed dataset
* While sampling these missing edges, we have added a condition to only consider an edge as missing if the distance between the source and destination was more than 2 as closely connected users are likely to be mutual friends even if an edge does not already exist.
* The edges that are present as 1 and the missing ones as 0. We have decided to use this presence or absence of an edge as the target class variable for prediction.
* We have taken into six-degree phenomenon while calculating shortest path between 2 nodes.

## 3.3 DATA PREPARATION

Standardisation of Data: Standardisation is required to bring the data to the same scale. In case the data is not standardised, we get skewed results as one or more features might dominate the others. For example, features like number\_of\_followers have a wider distribution than page\_rank because of which number\_of\_followers will dominate over page\_rank. However, after standardisation, all values lie between 0 to 1 and will be given equal weightage while training a machine learning model

Split Data for Training and Testing: In order to test the model on unseen data, it is necessary to split the data and this should be done randomly to avoid bias in the training or the testing phase. Random split ensures that the model is trained on edges which belong to both the classes (1 and 0). We have used 70% of the data for training and the remaining 30% for testing the model.

## 3.4 FEATURE ENGINEERING

## 3.4.1 PAGE RANK

Standardisation PageRank is the algorithm used by Google Search to rank websites in their search engine results. It works by counting the number and quality of links to a page to determine a rough estimate of how important the website is. The underlying assumption is that more important websites are likely to receive more links from other websites. Accordingly, we calculated the page ranks of both source and destination nodes of each edge and these formed two of the features we used.

**3.4.2 SHORTEST PATH**

The shortest path between two nodes is simply the path with the least number of intermediate nodes. In our network, if a direct link already exists between two nodes, we first delete it and then calculate the shortest path between them. The intuition behind this feature is that nodes which are close to each other have shorter path lengths indicating that they are likely to be good recommendation candidates.

## 3.4.3 FOLLOWS BACK

This feature simply indicates whether a reversely directed edge exists in the network for each existing edge, ie whether a user follows back one of his followees.

## 3.4.4 FOLLOWER AND FOLLOWEE COUNT

These features are the number of followers and followees of source and destination nodes. The intuition here is that popular streamers have a large number of followers and are good choices for recommendation candidates.

## 3.4.5 GRAPH THEORY ALGORITHMS

## 3.4.5.1 COMMON NEIGHBOUR

Common Neighbours is a similarity measure used in link prediction that captures the notion that a common friend may introduce two strangers to each other. [It tells us about the total number of common friends a pair of nodes have](http://education.abcom.com/link-prediction-using-similarity-measures/). [The common-neighbours predictor computes the number of common neighbours between two nodes and entities with more neighbours in common are more likely to have a link**.**](https://en.wikipedia.org/wiki/Link_prediction)

## 3.4.5.2 PREFERENTIAL ATTACHMENT

Preferential attachment score will be computed for each pair of nodes given in the iterable. The pairs must be given as 2-tuples (u, v) where u and v are nodes in the graph. If ebunch is None then all non-existent edges in the graph will be used.

## 3.5 HEATMAP OF THE FEATURES

A heatmap of the features uses colours to indicate the strength and direction of the correlation between each pair of features. Correlation is a measure of how much two features vary together. For example, if two features have a positive correlation, it means that they tend to increase or decrease together. If they have a negative correlation, it means that they tend to move in opposite directions. A heatmap of the features can help you understand which features are related to each other and which ones are independent. This can help you select the most relevant features for your analysis or modelling task. You can also use a heatmap of the features to identify any outliers or anomalies in your data.

## 3.4.5.2 EXPLORATORY DATA ANALYSIS

Exploratory data analysis (EDA) is a method used by data scientists to analyze and investigate data sets and summarize their main characteristics, often using data visualization methods. EDA helps determine how best to manipulate data sources to get the answers you need, making it easier for data scientists to discover patterns, spot anomalies, test a hypothesis, or check assumptions.

EDA is primarily used to see what data can reveal beyond the formal modelling or hypothesis testing task and provides a better understanding of data set variables and the relationships between them. It can also help determine if the statistical techniques you are considering for data analysis are appropriate.

EDA is an important first step in any data analysis, as it can help you look at data before making any assumptions. It can help you identify obvious errors, as well as better understand patterns within the data, detect outliers or anomalous events, find interesting relations among the variables.