VIT®

Vellore Institute of Technology

(Deemed to be University under section 3 of UGC Act, 1956)

School of Computing Science and Engineering

VIT Chennai

Vandalur - Kelambakkam Road, Chennai - 600 127

Review III Report

Analysis Of Cricket Data Using Gephi

Programme: M.Tech CSE with specialization in AI and ML

Course: CSE6006 - NoSQL

Slot: E1

Faculty: Dr. A. Bhuvaneswari

Component: J

Title: Analysis of International Cricket Data

Team Member(s): 4

Venkata Sai Satvik 20MAI1005

Khanjan Shah 20MAI1008

Prit M Vasiyani 20MAI1014

Shubham Fuldeore 20MAI1021

ABSTRACT:

As there are tremendous measure of information accessible these days, Data mining is an arising field in sports examination. To pick a best group or to anticipate appropriate arrangement for dominating a match or to examine shortcoming of the rival, information mining assumes an imperative part. In this task we are wanting to break down the global cricket information utilizing Gephi apparatus, which is profoundly ideal instrument utilized in ventures for organizations and diagrams examination. As there are a few highlights accessible in the dataset, among all that highlights we will utilize some of significant highlights that are sufficient for better expectation, significant highlights as well as can take a lot of choice in our examination. Our examination separated into three areas; prior to beginning the game, after one innings played and constant fall of wickets which prompts the likely expectation of the odds of winning and losing even while the game is in progress. We were utilizing some examination and grouping calculation on our dataset to show the precision level of our dataset.

INTRODUCTION

A NoSQL initially alluding to non-SQL or non-social is an information base that gives a system to capacity and recovery of information. This information is demonstrated in implies other than the plain relations utilized in social data sets. Such information bases appeared in the last part of the 1960s, however didn't acquire the NoSQL moniker until a flood of ubiquity in the mid twenty-first century. NoSQL data sets are utilized continuously web applications and enormous information and their utilization are expanding over the long haul. NoSQL frameworks are additionally at times called Not just SQL to stress the way that they may uphold SQL-like question dialects.

OBJECTIVE AND GOAL OF THE PROJECT

It would understate the obvious to communicate that Indians love cricket. The game is played in essentially wherever of India, country or metropolitan, celebrated with the young and the old the equivalent, partner billions in India not at all like some other game. Cricket likes lots of media thought. There is a ton of money and prevalence being referred to. All through the latest very much an extended period of time, advancement has from a genuine perspective been a particular benefit. Groups are spoilt for choice with streaming media, rivalries, sensible permission to flexible based live cricket watching, and anything is possible from that point.

Today, there are rich and for all intents and purposes limitless supplies of cricket coordinate records and experiences available, e.g., ESPN cricinfo and cricsheet. These and a couple of such cricket data bases have been used for cricket examination using the latest AI and judicious showing estimations. Media and redirection stages close by first class games bodies related with the game use development and examination for choosing key estimations for improving game ruling prospects: • batting performance moving average,

- score forecasting,
- gaining insights into fitness and performance of a player against different opposition,
- player contribution to wins and losses for making strategic decisions on team composition

PROBLEM STATEMENT

The thought is to make an organization investigation of complete execution of both the cooperative individuals those are essential for the counterpart for at any rate once. The investigation will be performed on every one of the parts of a player that he/she can do on the field.

KEY DATA ANALYTICS OBJECTIVES

 Sports data assessment are used in cricket just as various games for improving the overall gathering execution and enlarging winning prospects.

 Continuous data assessment can help in procuring pieces of information regardless, during the game for changing techniques by the gathering and by related associations for financial benefits and advancement.

 Other than chronicled assessment, insightful models are handled to choose the possible match results that require basic computing and data science expertise, portrayal devices and capacity to fuse more cutting-edge discernments in the examination.

THE CHALLENGES

Information Cleaning and preprocessing

IPL has extended cricket past the exemplary test match organization to a lot bigger scope. The quantity of matches played each season across different configurations has expanded thus has the information, the calculations, more up to date sports information investigation advances and reenactment models. Cricket information investigation requires field planning, player following, ball following, player shot examination, and a few different perspectives engaged with how the ball is conveyed, its point, twist, speed, and direction. Every one of these variables together have expanded the intricacy of information cleaning and preprocessing.

Dynamic Modeling

In cricket, actually like some other game, there can be an enormous number of factors identified with following different quantities of players on the field, their qualities, the ball, and a few prospects of possible activities. The intricacy of information investigation and demonstrating is straightforwardly corresponding to the sort of prescient inquiries that are advanced during examination and are exceptionally subject to information portrayal and the model. Things get much more testing as far as calculation, information correlations when dynamic cricket play forecasts are looked for, for example, what might have occurred if the batsman had hit the ball at an alternate point or speed.

Prescient Analytics Complexity

A significant part of the dynamic in cricket depends on questions, for example, "how frequently does a batsman play a specific sort of shot if the ball conveyance is of a specific kind", or "how does a bowler change his line and length if the batsman reacts to his conveyance with a particular goal in mind". This sort of prescient investigation inquiry requires profoundly granular dataset

accessibility and the ability to combine information and make generative models that are exceptionally precise.

THE NEED OF ANALYSIS AND STATISTICS

Experiences have reliably had a colossal part in sports. As I referred to above, sports assessment is on the climb and will continue accepting a basic part in how gatherings work, pick their players, how they play the game, etc. Cricket is something similar. The runs scored by a batsman, the wickets were taken by a bowler, or the matches won by a cricket group – these are through and through examples of the principle numbers in the game of cricket. Tracking all such estimations enjoys various benefits. The gatherings and the individual players can plunge significant into this data and find regions of progress. It can similarly be used to overview an opponent's characteristics and deficiencies.

LITERATURE SURVEY

In this paper, an endeavor has been made to consider the exhibition of Cricket Players utilizing Factor Analysis strategy. For this, the dataset of 85 batsmen, 85 bowlers; and 95 batsmen, 95 bowlers have been considered from IPL9, 2016 (20 overs) and ICC World Cup, 2015 (50 overs) individually, and the discoveries of this examination uncovers that batting ability overwhelms over bowling capacity which is in similarity with a prior investigation on same sort of game. [1]

Keywords: Factor Analysis, World Cup, IPL, Cricket.

Winning a One Day International (ODI) cricket match relies upon different elements identified with scoring just as the athletic qualities of the two groups. While a portion of these variables have been very much examined in the writing, others presently can't seem to be explored. In this investigation, measurable importance for a scope of factors that could clarify the result of an ODI cricket match is investigated. Specifically, home field advantage, winning the throw, blueprint (batting first or handling first), match type (day or day and night), and the impact of the Duckworth-Lewis technique for matches abbreviated because of climate interferences will be key interests in our examination. For motivations behind model building, strategic relapse is applied reflectively to information previously got from recently played matches. Some astonishing outcomes arise. [2]

Keywords: cricket, logistic regression, sports statistics

Cricket requires a congruity of execution by one group in batting until an infringement of the playing rules is experienced. That infringement—causing a paired choice, the conceivable excusal of a batsman—may significantly affect the course of the batting side, and at last on the game. A

few excusals are unquestionable, however a huge minority are the subject of choices by umpires. A portion of these are mind boggling circumstances yet need quick dynamic, and TV moderate movement replays have featured the issue of evident miss-decisions. The impact of such choices on batsmen and on innings sums can be reenacted and thus assessed to survey their significance. [3]

The ICC Men's T20 Cricket World Cup 2020 is booked to be facilitated by Australia in the long stretch of October and November 2020. AI in sports investigation is presently a day's effectively applied for forecast of champs. The work introduced in this paper expects to foresee the champ of the impending seventh variant of ICC Men's T20 world cup utilizing Random Forest Classifier, Naïve Bayes, KNN, Logistic Regression, Decision Tree, SVM, Bagging Classifier, Extra Trees Classifier, Voting (HARD and SOFT) preparing. Every one of these methodologies are tried on the distinctive accessible memorable information of worldwide cricket matches played between various nations from 2005 to March 2020. Unstructured memorable cricket insights is picked from ESPN and Cricbuzz sites. Exploratory outcomes demonstrate that all methodologies can soak up the separated examples from the different arrangement of matches performed and henceforth is found appropriate to anticipate the victor of the ICC Men's T20 Cricket World Cup 2020. A near report is likewise introduced for the expectations made through various methodologies. [4]

Keywords: Cricket analytics, Winner Prediction, Classification.

In cricket the exhibition of the players whether bowler or batsman is being investigated with the assistance of basic measurable instruments. For the most part normal scores, strike rate, normal runs per wicket are being utilized. The principle thought of this exploration study is to give the utilization of value control graphs in assessing the presentation of the players. We have made the relative investigation of the two remarkable batsman of this game. We utilized the individual and moving reach control graphs for assessing their exhibitions. We applied essential sharpening rules to break down the non-arbitrary and un-characteristic variety present in their presentation. We will offer the chance to experts to check in the light of their insight for the assignable causes that have made the scores wild so that in future these un-characteristic varieties can be controlled and execution can be improved. [5]

Keywords: batsman, performance, moving range control charts, individual control charts graphical displays, assignable cause, non-normality.

Indian Premier League (IPL) is one of the more well-known cricket world competitions, and its monetary is expanding each season, its viewership has expanded notably and the wagering market for IPL is developing altogether consistently. With cricket being a powerful game, bettors and bookies are boosted to wager on the match results since it is a game that changes ball-by-ball. This paper researches AI innovation to manage the issue of foreseeing cricket match results dependent on recorded match information of the IPL. Powerful highlights of the dataset have been distinguished utilizing channel based strategies including Correlation-based Feature Selection,

Information Gain (IG), Relief and Wrapper. All the more significantly, AI procedures including Naïve Bayes, Random Forest, K-Nearest Neighbor (KNN) and Model Trees (grouping through relapse) have been received to produce prescient models from particular capabilities inferred by the channel based techniques. Two highlighted subsets were planned, one dependent on host group advantage and other dependent on Toss choice. Chosen AI procedures were applied on both capabilities to decide a prescient model. Test tests show that tree-based models especially Random Forest performed better as far as exactness, accuracy and recall measurements when contrasted with probabilistic and factual models. Notwithstanding, on the Toss included subset, none of the considered AI calculations performed well in delivering precise prescient models. [6]

Player grouping in the sport of cricket is vital, as it helps the mentor and the commander of the group to distinguish every player's job in the group and allocate obligations as needs be. The goal of this examination is to characterize all-rounders into one of the four classifications in a single day global (ODI) Cricket design and to precisely foresee the new all-rounders'. This investigation was led utilizing an assortment of 177 players and ten player-related execution pointers. The expectation was led utilizing three AI classifiers, to be specific Naive Bayes (NB), knearest neighbors (kNN), and Random Forest (RF). As per the exploratory results, RF demonstrates altogether better expectation precision of 99.4%, than its partners. [7]

Keywords: Team sport, machine learning, cricket, ODI, player classification

As one-day worldwide (ODI) games ascend in fame, it is imperative to comprehend the potential indicators that influence the game result. The home-field advantage, cointoss result, bat-first or second, and day versus day-night game configuration are such mainstream factors being considered in the cricket writing. This article centers around an extensive investigation of evaluating the meaning of those significant indicators through graphical 'characterization and relapse tree' (CART) and the mainstream strategic relapse draws near. This examination uncovers the significance of the home-field advantage for significant cricket playing countries in one-day worldwide games however questions the consistency of such factors under various playing conditions. Critically, the home-field advantage is examined additionally dependent on the rival's geological area. Decisively, the CART approach gives intriguing and novel translations to well known indicators in ODI games. [8]

This article is worried about the recreation of one-day cricket matches. Given that solitary a limited number of results can happen on each ball that is bowled, a discrete generator on a limited set is created where the result probabilities are assessed from authentic information including one-day worldwide cricket matches. The probabilities rely upon the batsman, the bowler, the quantity of wickets lost, the quantity of balls bowled and the innings. The proposed test system seems to make a sensible showing with creating practical outcomes. The test system permits examiners to resolve complex inquiries including one-day cricket matches. [9]

DATA SET DESCRIPTION:

For our project we took data from cricsheet.org source.

Cricsheet is the Retrosheet for cricket, they provide ball by ball data of each matches played by men's and women's. The data of cricsheet is in zip file, which we need to extract after downloading dataset.

The Dataset consist of :

- match_id
- season
- start_date
- venue
- innings
- ball
- batting_team
- bowling_team
- striker
- non_striker
- bowler
- runs_off_bat
- extras
- wides
- noballs
- byes
- legbyes
- penalty
- wicket_type
- player_dismissed
- other_wicket_type
- other_player_dismissed

This all are the columns in dataset. From this dataset we have extracted the information of partnership, batting and bowling and created a network in the Gephi. The dataset we took is for IPL matches. As it is the most played and famous cricket league and most of the useful data comes from the IPL. The data is of 12 years IPL matches, from 2008 to 2020 we have our data for each entity given above.

METHODOLOGY:

System Architecture diagram

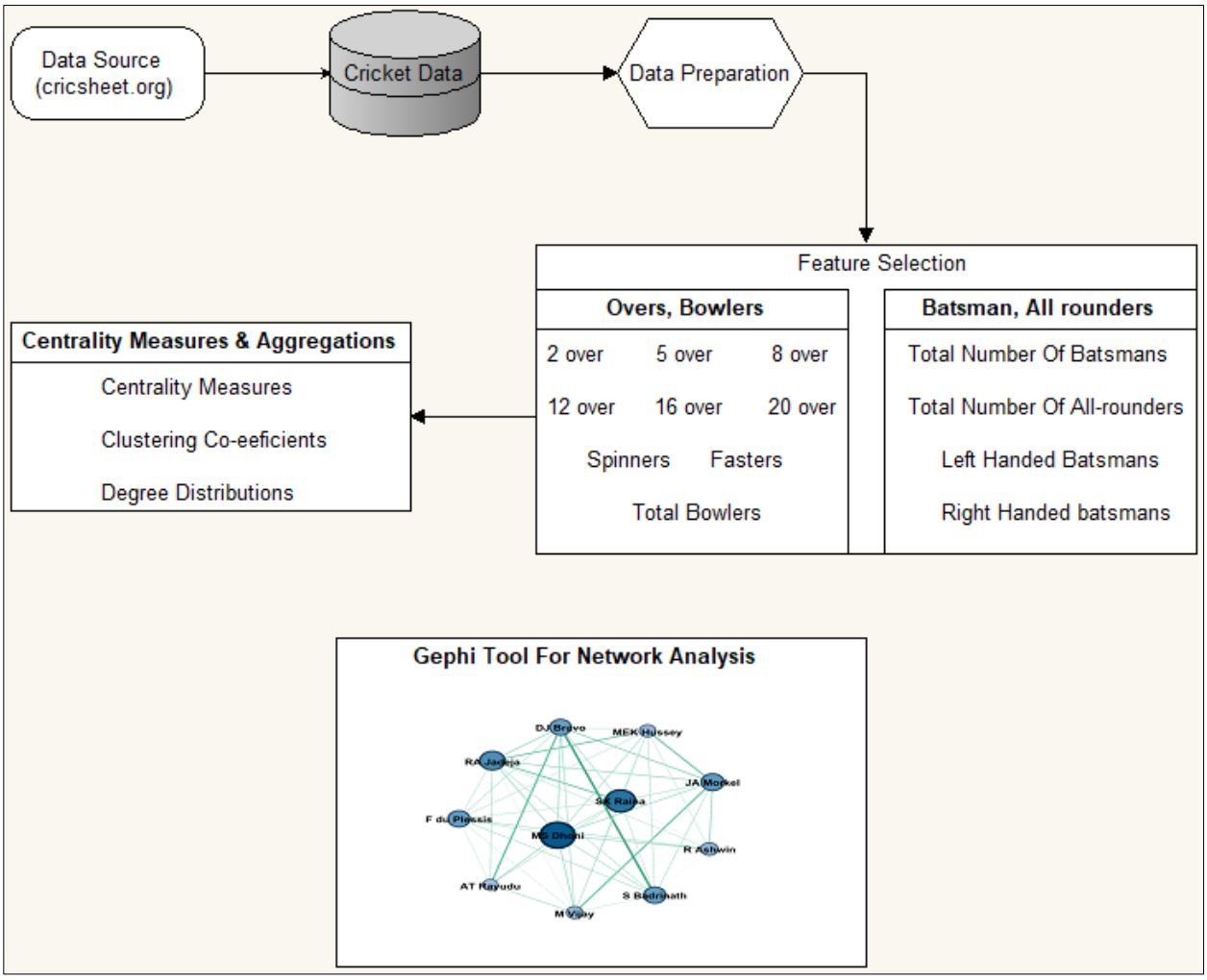


Figure 1: System Architecture

Flowchart

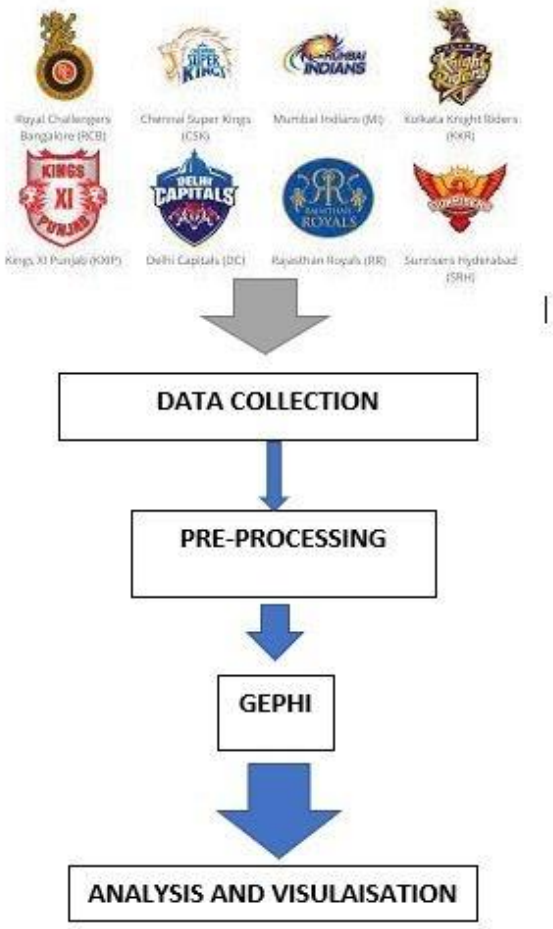


Figure 2: Flowchart

EXPERIMENTAL SETUP:

We have done preprocessing of the data to extract the important information in Jupyter Notebook with Python 3.6 and Analysis in Gephi.

IMPLEMENTATION STEPS:

- Firstly, We have extracted NodeList and EdgeList from the dataset with preprocessing and created different files for different team as follow:

Figure 3: Preprocessed Edgelist Files

- Now, Imported this spreadsheet in the Gephi

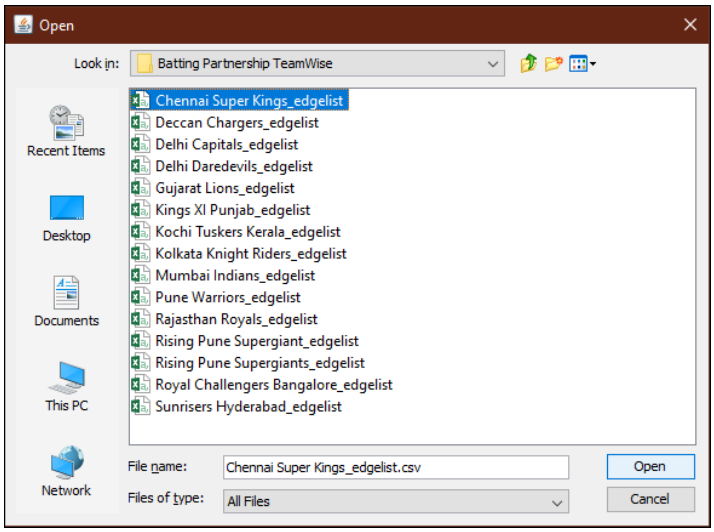


Figure 4: Import Edgelist File

- Now, Select the type of imported spreadsheet. Whether it is edgelist, nodelist or adjacency matrix.

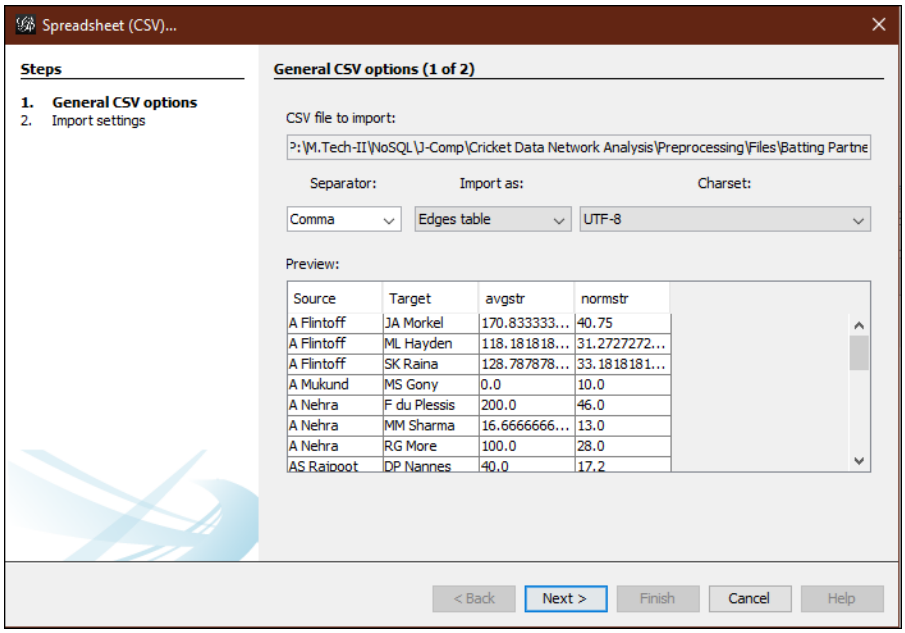


Figure 5: Import as Edges Table

- Now, Selected type of graph we want.

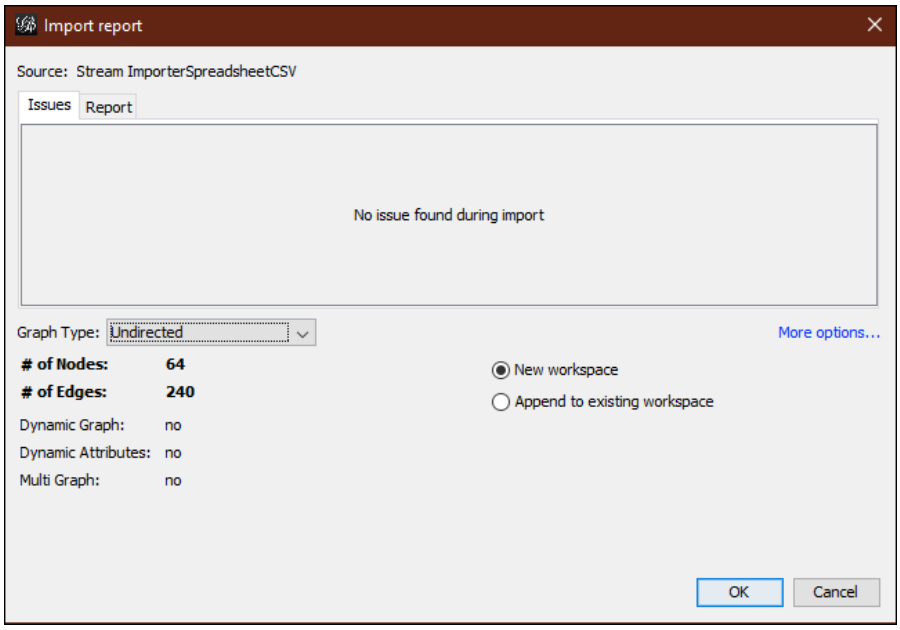


Figure 6: Graph Type Select

- Now, It will initially create the graph as follow:

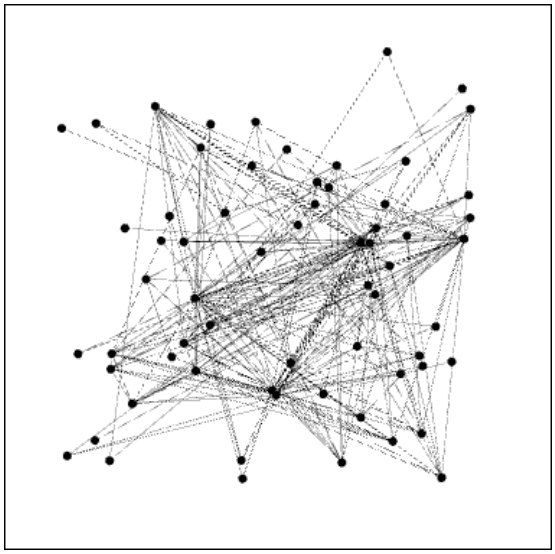


Figure 7: Initial Graph

- Making nodes size based on its degree.

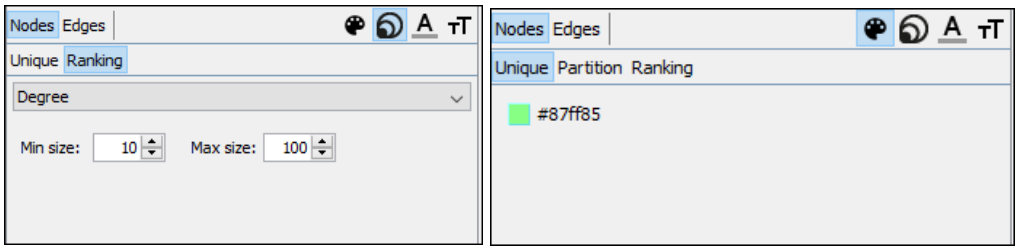


Figure 8: Changing Node Size and Color

- Now, Applied Layout to the graph

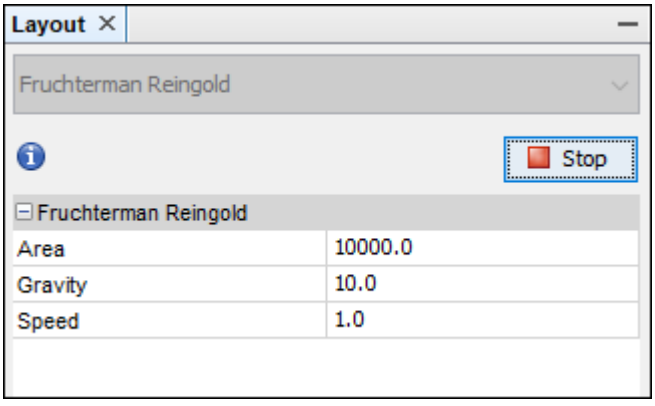


Figure 9: Applying Layout

- Copy the Normalized Strike rate to Weight of Edges, and then check the graph.

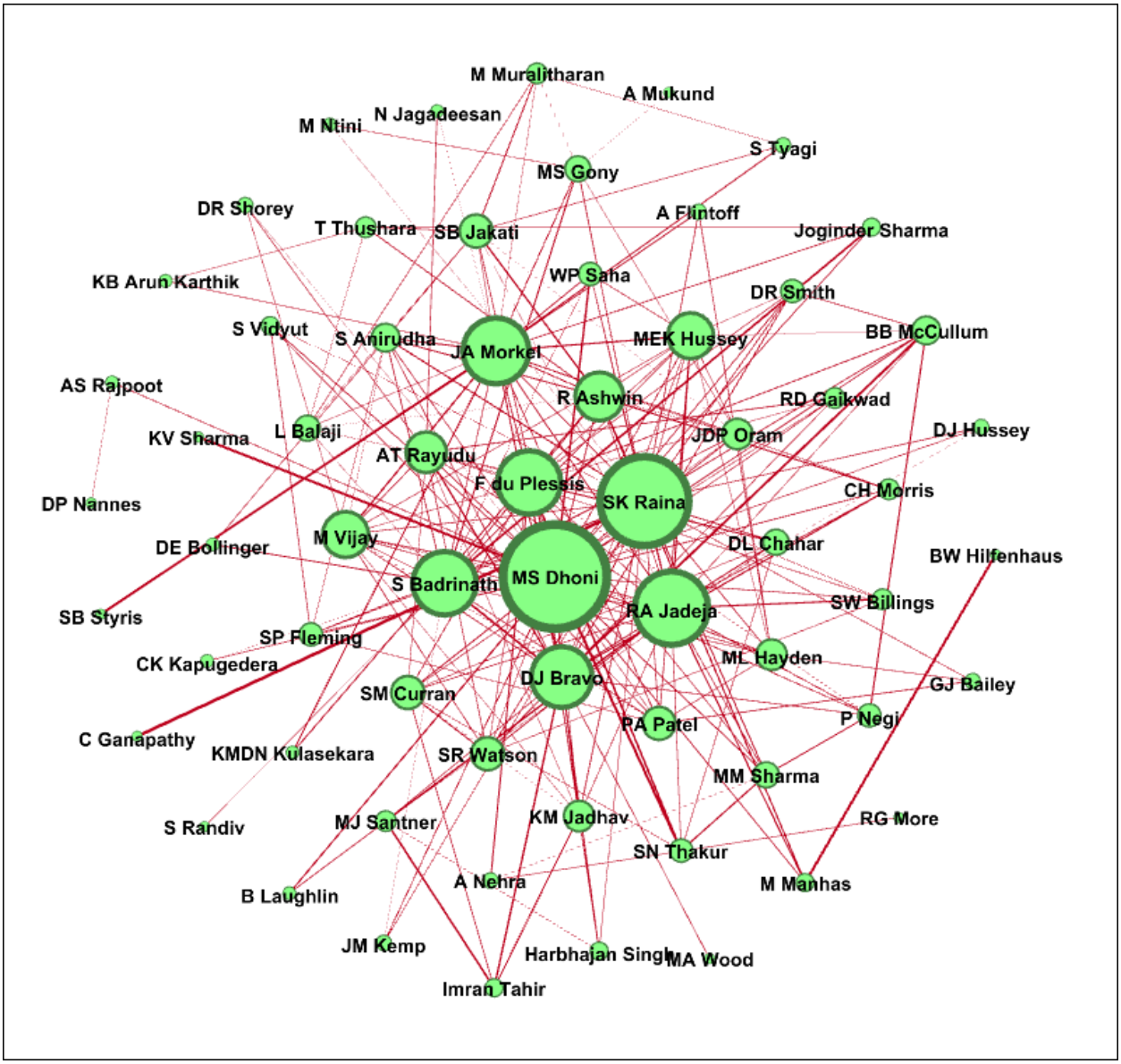


Figure 10: Graph Created

- Now, applying filters to keep the valuable partnerships only.

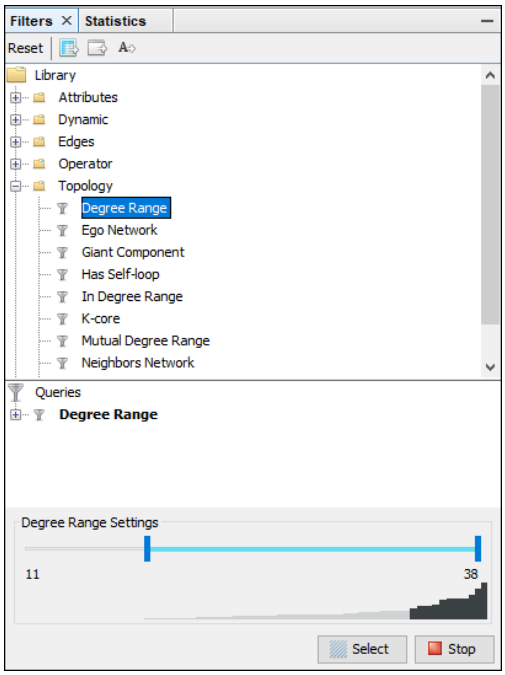


Figure 11: Applying Filters

- Final graph after filtering is as follow:

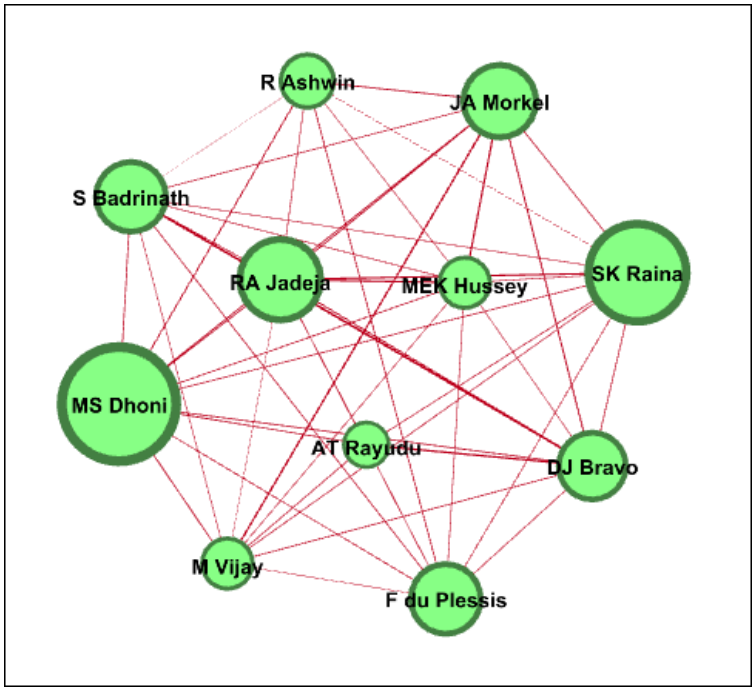
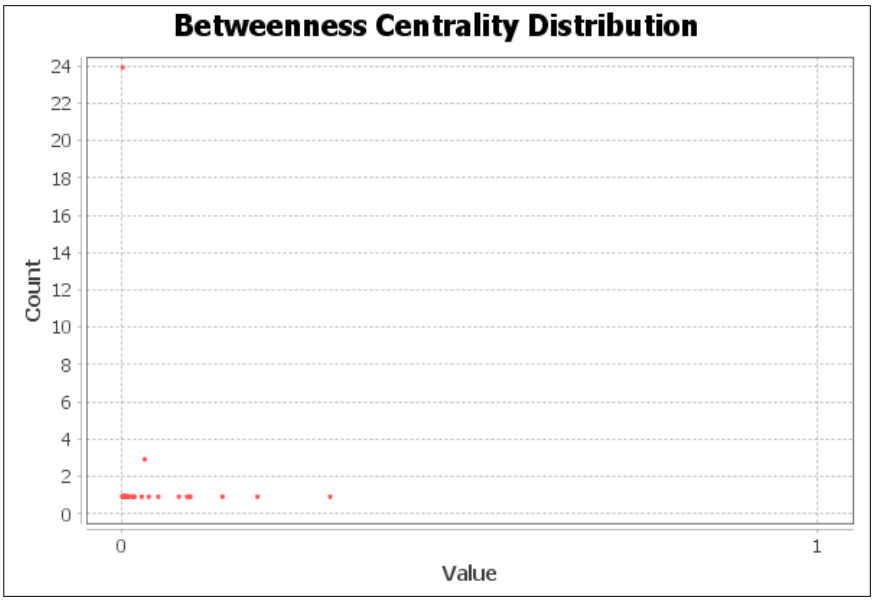
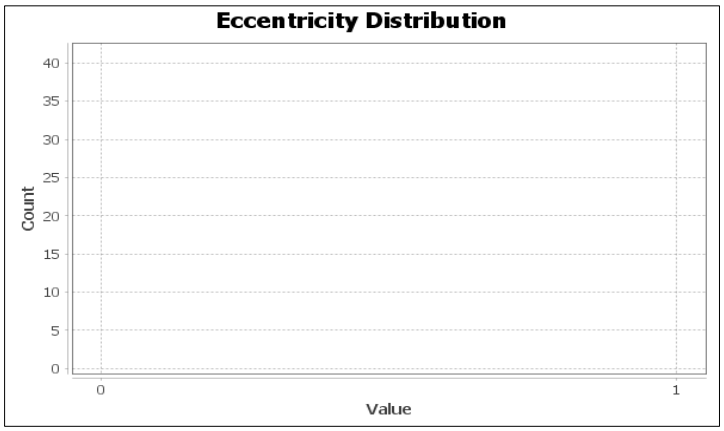


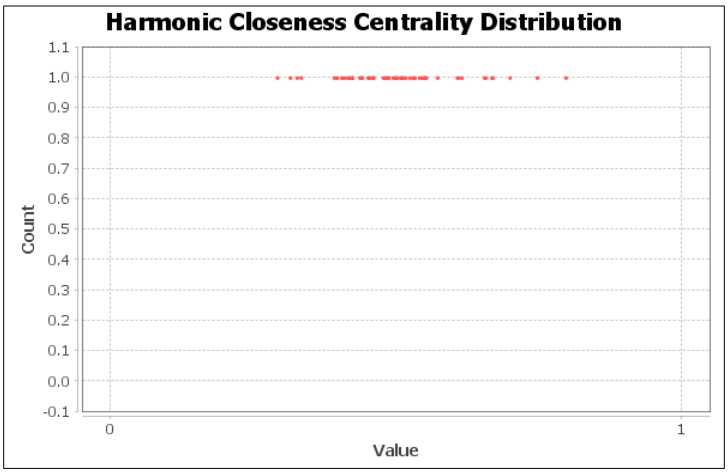
Figure 12: Final Analyzed Network

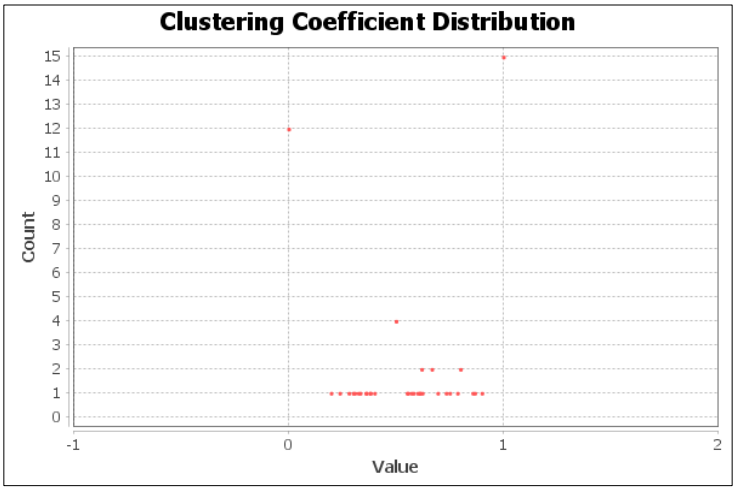
- Other statistics of this graph is discussed in Results. Same steps need to be followed for each team and each team and each mode i.e. batting partnership, batting (team wise), bowling (team wise) and independent batting.

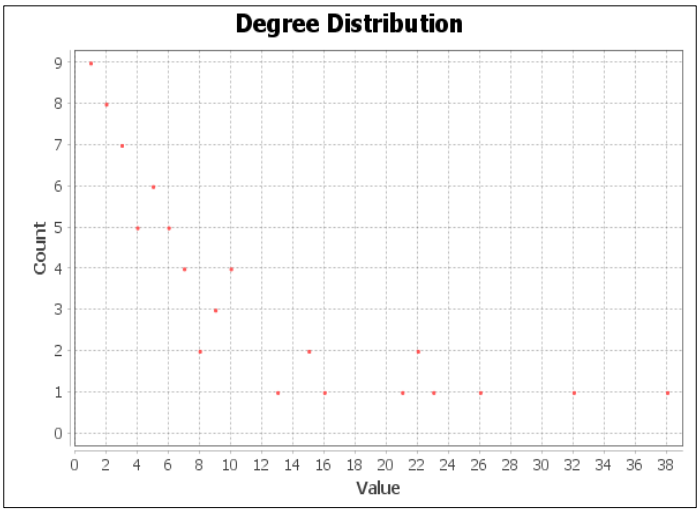
RESULTS AND DISCUSSIONS:




Eccentricity Distribution
Count
Value
0
5
10
15
20
25
30
35
40
0
1


Harmonic Closeness Centrality Distribution
Count
Value
-0.1
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1.1
0
1


Clustering Coefficient Distribution
Count
Value
0
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
-1
0
1
2



PROJECT SCREENSHOTS:

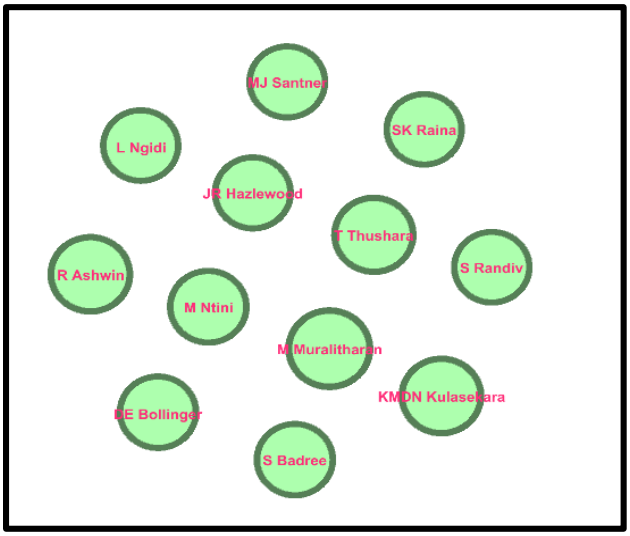


Figure 13: CSK_Bowling

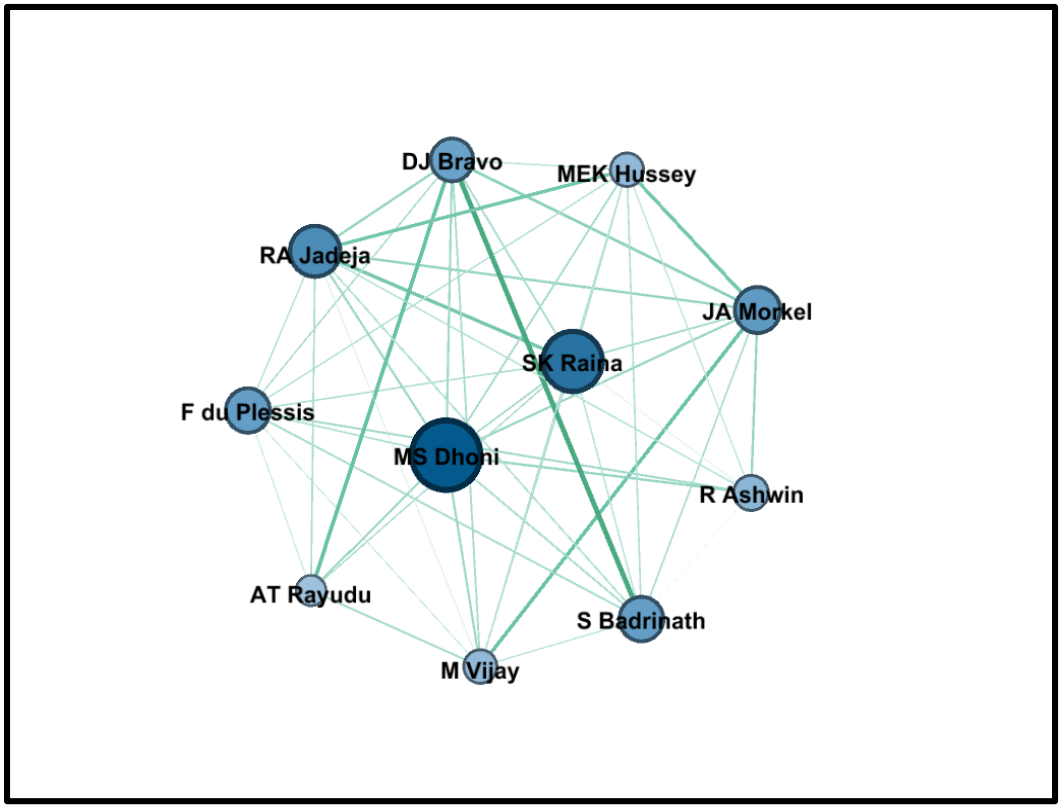


Figure 14: CSK Batting Partnership Filtered

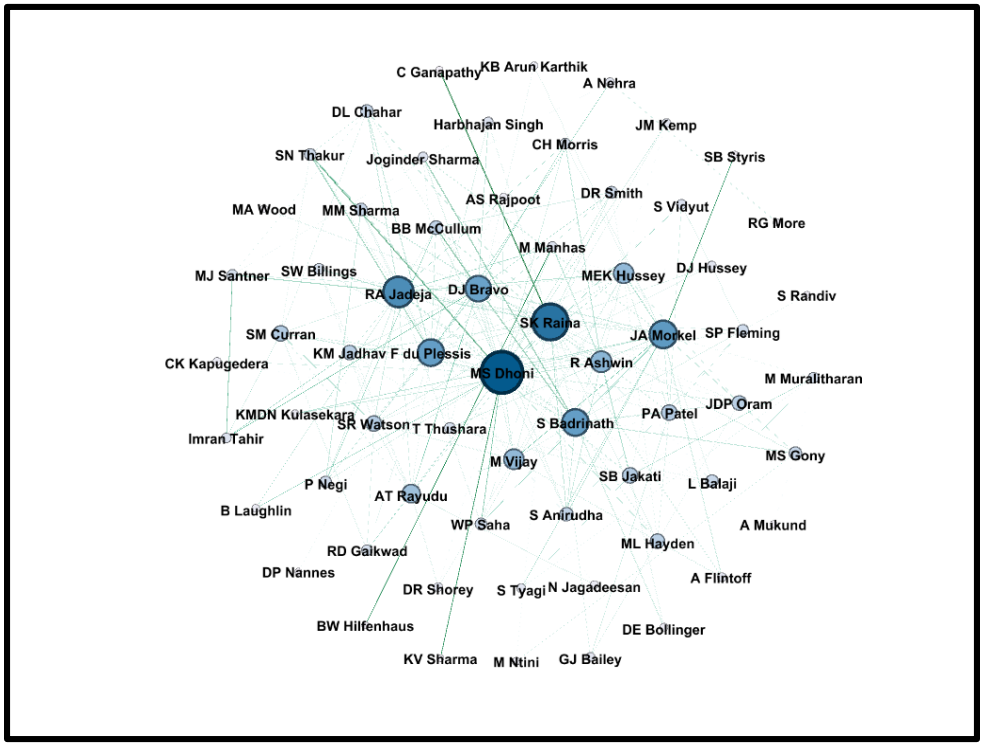


Figure 15: CSK Batting Partnership

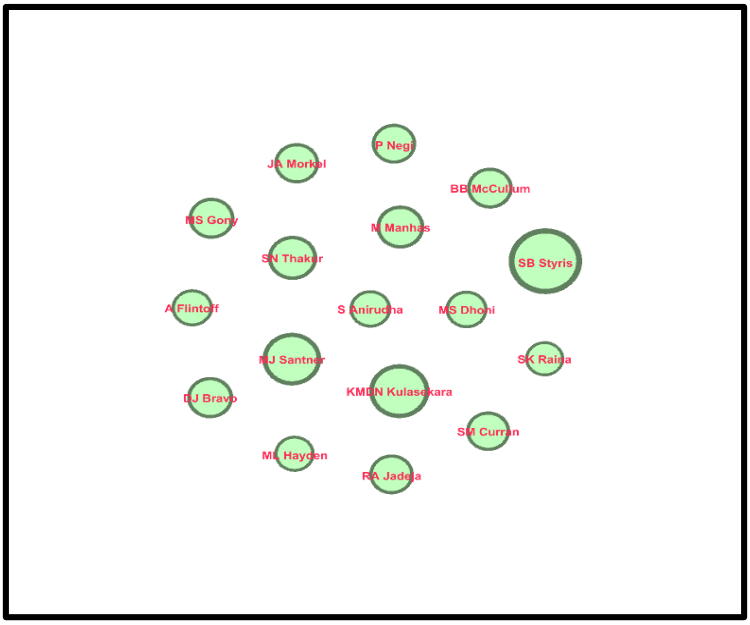


Figure 16: CSK Batting

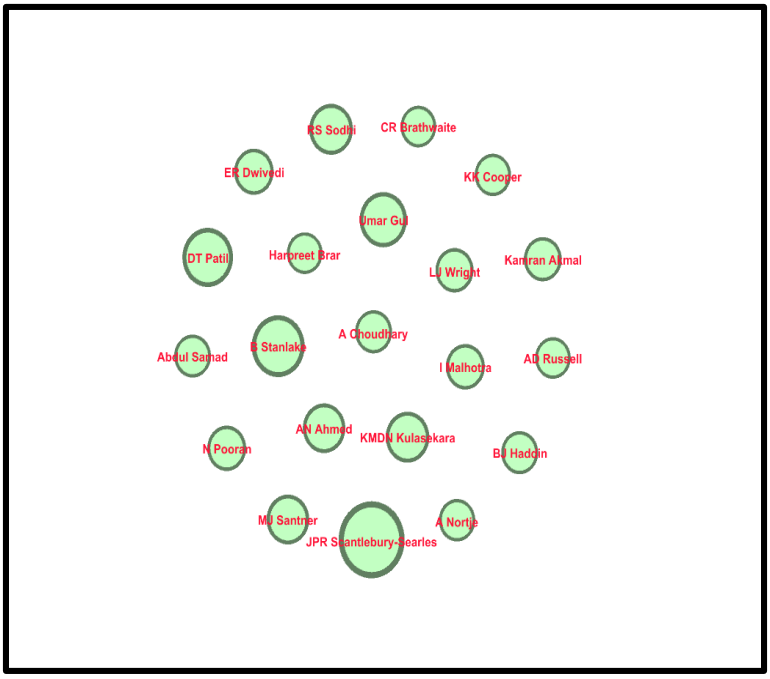


Figure 17: Overall Batsman

CONCLUSION:

From this project we conclude with all our analysis we did using Gephi tool and python. Starting from the data preprocessing then data preparation and segregate our data into different files for analysis purpose. Here we did network analysis which required data in the form of edges and vertices. We use this data and analyze each attribute in detail with all its statistical measures. After giving statistical measures we are done with network analysis of each team and players. So our overall contribution in the project is divided into 4 parts. Research paper reading and collection of ideas, Data collection and preparation, analysis using Gephi tool, Report writing and display the most optimal results. Further, we decide to do prediction on the dataset and display the link prediction graph through Gephi.

REFERENCES:

1. Sricharan Shah, Partha Jyoti Hazarika and Jiten Hazarika: A Study on Performance of Cricket Players utilizing Factor Analysis Approach. Global Journal of Advanced Research in Computer Science, 8 (3), March-April 2017,656-660.

2. Ananda Bandulasiri and Ferry Butar: Statistical Analysis of One Day International Cricket. Diary of sports Science and Medicine, 5, 480-487.

3. Chedzoy, Olaf. (2002). The impact of umpiring mistakes in cricket. Diary of the Royal Statistical Society: Series D (The Statistician). 46. 529 - 540. 10.1111/14679884.00107.

4. Shashank Singh, Yash Aggarwal, Kumud Kundu, Quantitative Analysis of Forthcoming ICC Men's T20 World Cup 2020 Winner Prediction utilizing Machine Learning. Worldwide Journal of Computer Applications (0975 – 8887) Volume 176 – No. 32, June 2020.

5. Muhammad Daniyal, Tahir Nawaz, Iqra mubeen , Muhammad Aleem: Analysis of Batting Performance in Cricket utilizing Individual and Moving Range (MR) Control Charts. ISSN 1750-9823 (print) .International Journal of Sports Science and Engineering Vol. 06 (2012) No. 04, pp. 195-202.

6. Kumash Kapadiaa, Hussein Abdel-Jaberb, Fadi Thabtaha, Wael Hadic : Sport investigation for cricket match-up outcomes utilizing AI: An exploratory examination. Arrangement E (Analytics). 26. 129 - 140. 20.1311/6741-8984.11007.

7. Bandulasiri, A., Brown, T., Wickramasinghe, T. (2016).Classification of Allrounders in the round of ODI cricket, Operation Research and Decisions. 26 (4).

8. Kalanka P. Jayalath, An AI way to deal with investigate ODI cricket indicators.

9. Diary of Sports Analytics xx (20xx) x–xx DOI 10.3233/JSA-17175 IOS Press.

10. Tim B. SWARTZ 1*, Paramjit S. GILL2 and Saman MUTHUKUMARANA1, Modeling and reproduction for one-day cricket. The Canadian Journal of Statistics Vol. 37, No. 2, 2009, Pages 143–160 La revue canadienne de statistique.

CONTRIBUTORS:

1. **Venkata Sai Satvik: 20MAI1005**
   He has contributed for Literature Survey part and report preparation along with presentations. He also Contributed for Overall batting network analysis in Gephi.
2. **Khanjan Jayraj Shah: 20MAI1008**
   He has contributed for Implementation in Gephi for Bowling Network Analysis, along with results and conclusions in report.
3. **Prit M Vasiyani: 20MAI1014**
   He has contributed with Implementation in Gephi and Preprocessing of the dataset for Implementation. Along with that, he also done implementation steps part in report.
4. **Shubham Fuldeore: 20MAI1021**
   He has contributed for Dataset part and report formatting, along with that he also contributed to Batting Analysis part in Gephi.