GLOBAL TERRORISM AND ITS TANGIBLE EFFECTS

Data Science and Knowledge Discovery

EM623

Deepanshu Tyagi
dtyagi@stevene.edu

INTRODUCTION

- Over history, terrorism as a negative phenomenon has evolved in its methodologies and conduct. Only in the past few decades have we been learning more about it in order to try and stop the violence that it brings.

- A terrorist attack as the threatened or actual use of illegal force and violence by a non-state actor to attain a political, economic, religious, or social goal through fear, coercion or intimidation.

- In the following sections we try to understand what are the factors that make a terrorist attack successful.

Project Goals and Conditions

- What makes a terrorist attack Successful?

- Can knowing the attributes help government defence agencies plan ahead?

- As I was looking for dataset that could be used for both Network analysis and Decision tree modelling, this was the ideal dataset as it had a lot of columns that had summary and notes about the terrorist attack and also had variables that can be used for decision tree modelling.

- CRISP-DM approach will be used to extract useful information from the dataset.

BUSINESS UNDERSATNDING

- It is estimated that over 200,000 people have been killed in the acts of terrorism around the world

- Terrorism has been deployed as a tactic by some of the rebel forces to bring about a political, economic, religious, or social goal rather than purely military objectives.
- As of September 2015, there are 4.1 million Syrian refugees and 6.5 million people displaced within Syria. Many have fled to nearby countries, with a growing number fleeing to Europe, underlining the worldwide spill-over effects of the Syrian civil war.
- The rise of different terrorist groups time after time has seen an endless need economic resources pumped in by various countries.

Data Understanding

- **Source: https://www.kaggle.com/START-UMD/gtd**

- **Dataset Description:**
  Number of Observations:1,56,673
  Number of variables: 140

- **The Dataset covers terrorists incidents from 1970-2015 and contains huge amount of information on terrorist groups involved and their tactics and weapon used etc.**

- **There Dataset contains both categorical and numeric variables**

Data Understanding

- **The various columns like summary and notes column in the data contains extensive description of the incident and the texts from these columns will be used for Network Analysis.**

- **The success variable which is categorical with Yes and No will be used as Target Variable for decision tree analysis**

Data Preparation

- Two different approaches were used to prepare dataset for the purpose of Network Analysis and Decision tree modelling separately

- Most of the Project work time was spent on preparing the data for Decision tree modelling and network analysis.

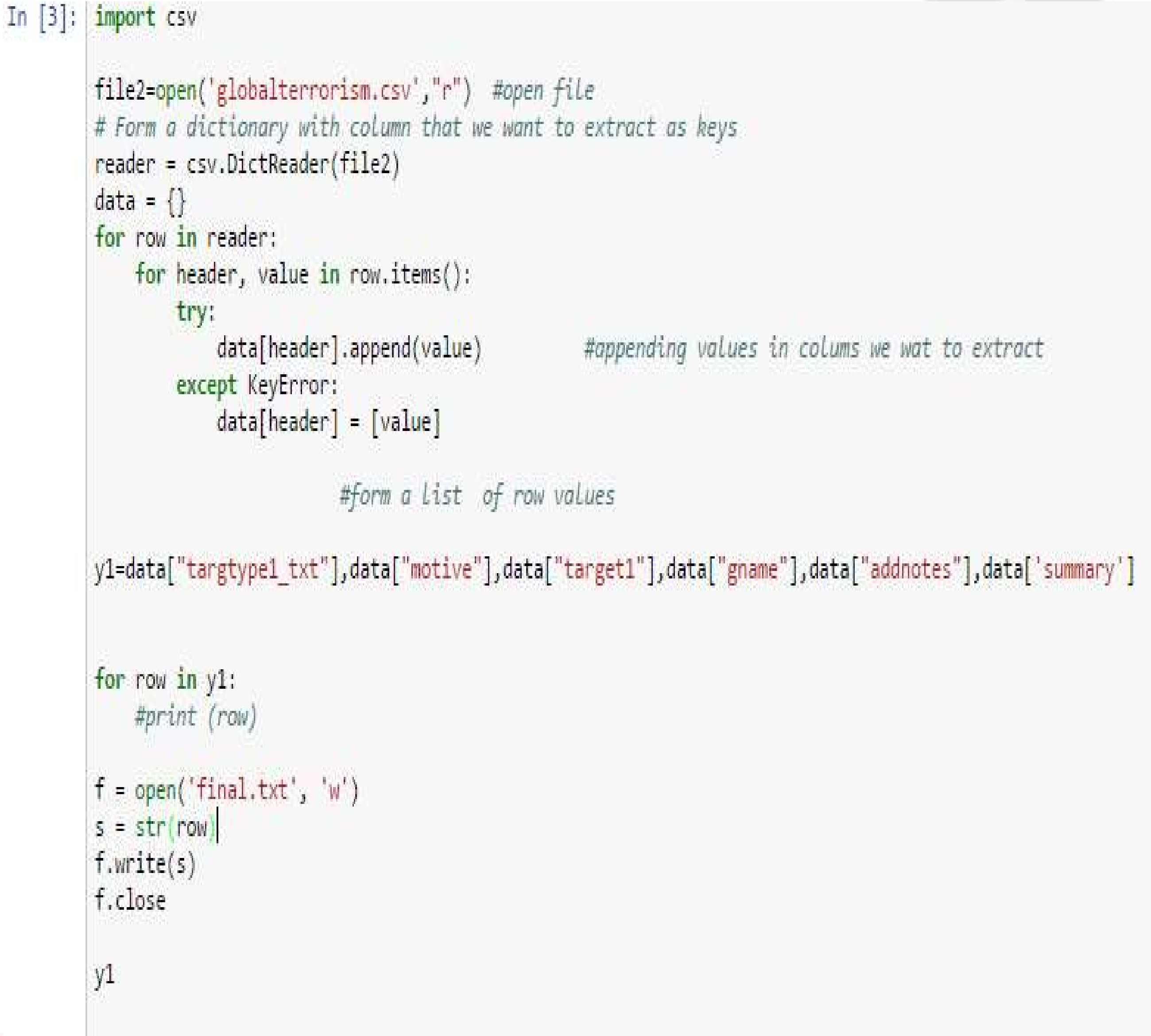
- A lot of research was put in matching the dataset columns with codebook and picking up the right columns for network analysis and decision tree modelling

- The dataset had columns with missing values, the columns that had over 40% of values as missing or NA were deleted using rattle

- In the following sections there is a detailed description of what methods were used to prepare data fro final analysis

Data Preparation

- **Data Preparation for Network Analysis**
  - **As mentioned earlier some of the columns had detailed text description of the terrorist event mainly the summary, notes and motive columns**
  - **These columns were used to extract the detailed description of the event using ipython jupyter notebook**



```
In [3]: import csv

        file2=open('globalterrorism.csv',"r")  #open file
        # Form a dictionary with column that we want to extract as keys
        reader = csv.DictReader(file2)
        data = {}
        for row in reader:
            for header, value in row.items():
                try:
                    data[header].append(value)          #appending values in colums we wat to extract
                except KeyError:
                    data[header] = [value]

                                    #form a list  of row values

        y1=data["targtype1_txt"],data["motive"],data["target1"],data["gname"],data["addnotes"],data['summary']


        for row in y1:
            #print (row)

        f = open('final.txt', 'w')
        s = str(row)
        f.write(s)
        f.close

        y1
```

Data Preparation

- Data preparation for Network analysis
- The file was downloaded as .txt file using python script
- The .txt file was than processed in wordij to remove the stopwords, removing the words and pairs that occurred less than 3 times in the .txt file.
- The .net file abtained from wordij was processed in gephi for network analysis.
- The node .csv file from the datalaboratory was than exported to rattle for further analysis.

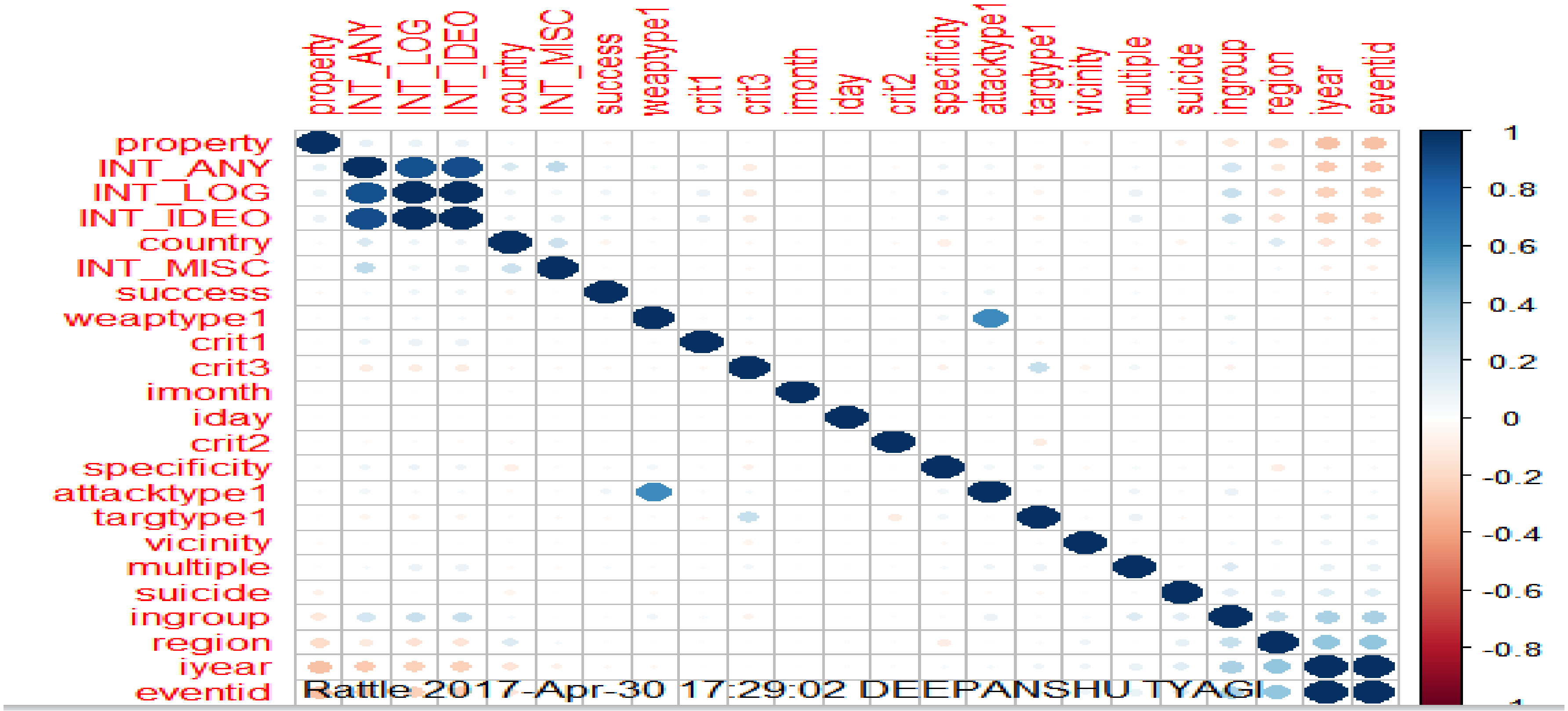
Data Preparation

- **Data Preparation for Decision tree modelling**
  - **Data had a mix of categorical and numerical variables**
  - **Columns with 40% missing values were deleted using rattle**
  - **Columns like eventid , date and columns with a lot of text description which we used for network analysis were removed as they were insignificant for decision tree modelling.**
  - **The number of variables after removing reduced to 45 from 140**
  - **Many categorical variables that were represented as numeric in rattle were recoded as categoric**

Data Preparation

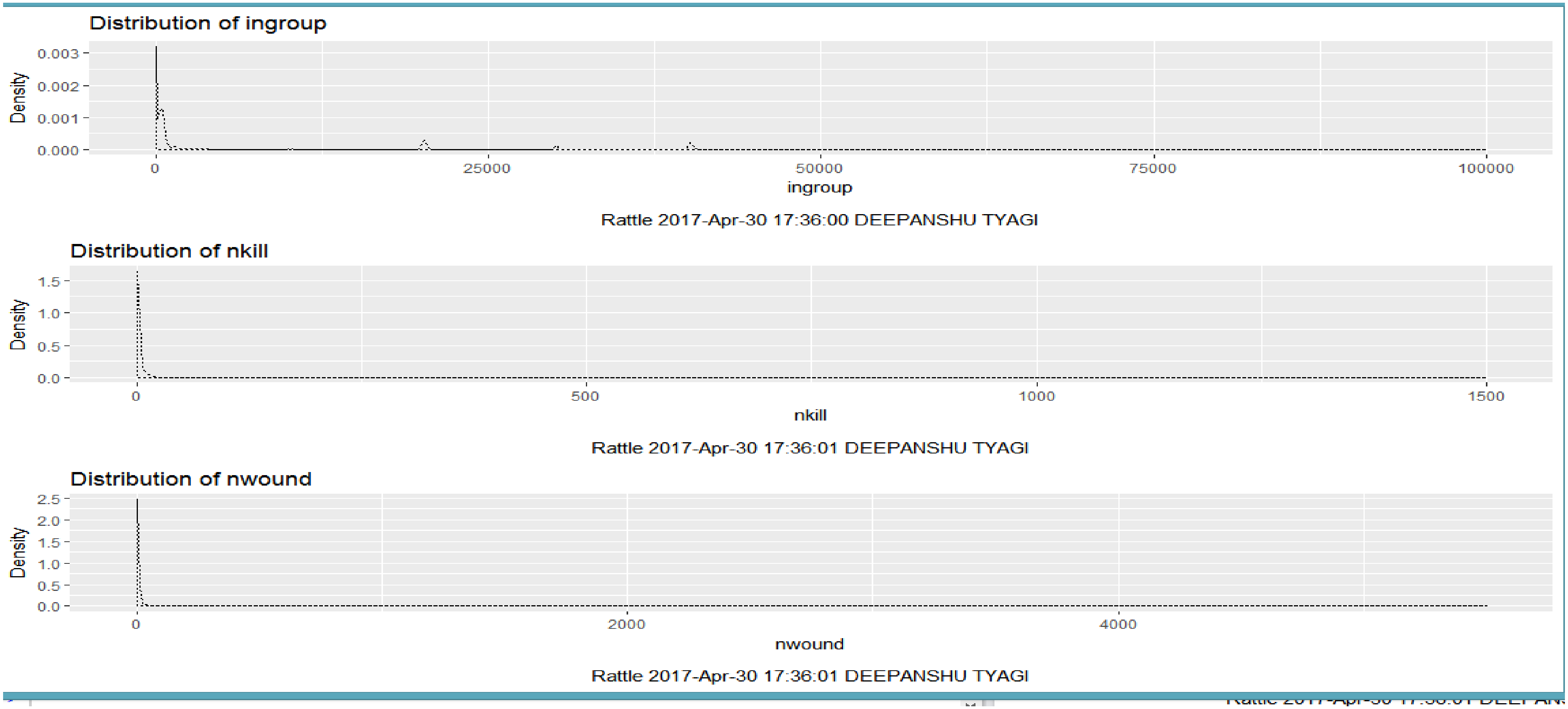
- Data Preparation for decision tree modelling
  - Below is the initial correlation matrix for the dataset that was used for removing highly correlated variables

Correlation globalterrorism.csv using Pearson



Data Preparation

- **Data Preparation for Decision tree modelling**
  - **After cleaning the data the histograms of numeric variables, Below graphs show distributions of number of kills and wounded after a terrorist attack.**



Distribution of ingroup

Rattle 2017-Apr-30 17:36:00 DEEPANSHU TYAGI

Distribution of nkill

Rattle 2017-Apr-30 17:36:01 DEEPANSHU TYAGI

Distribution of nwound

Rattle 2017-Apr-30 17:36:01 DEEPANSHU TYAGI

Data Preparation

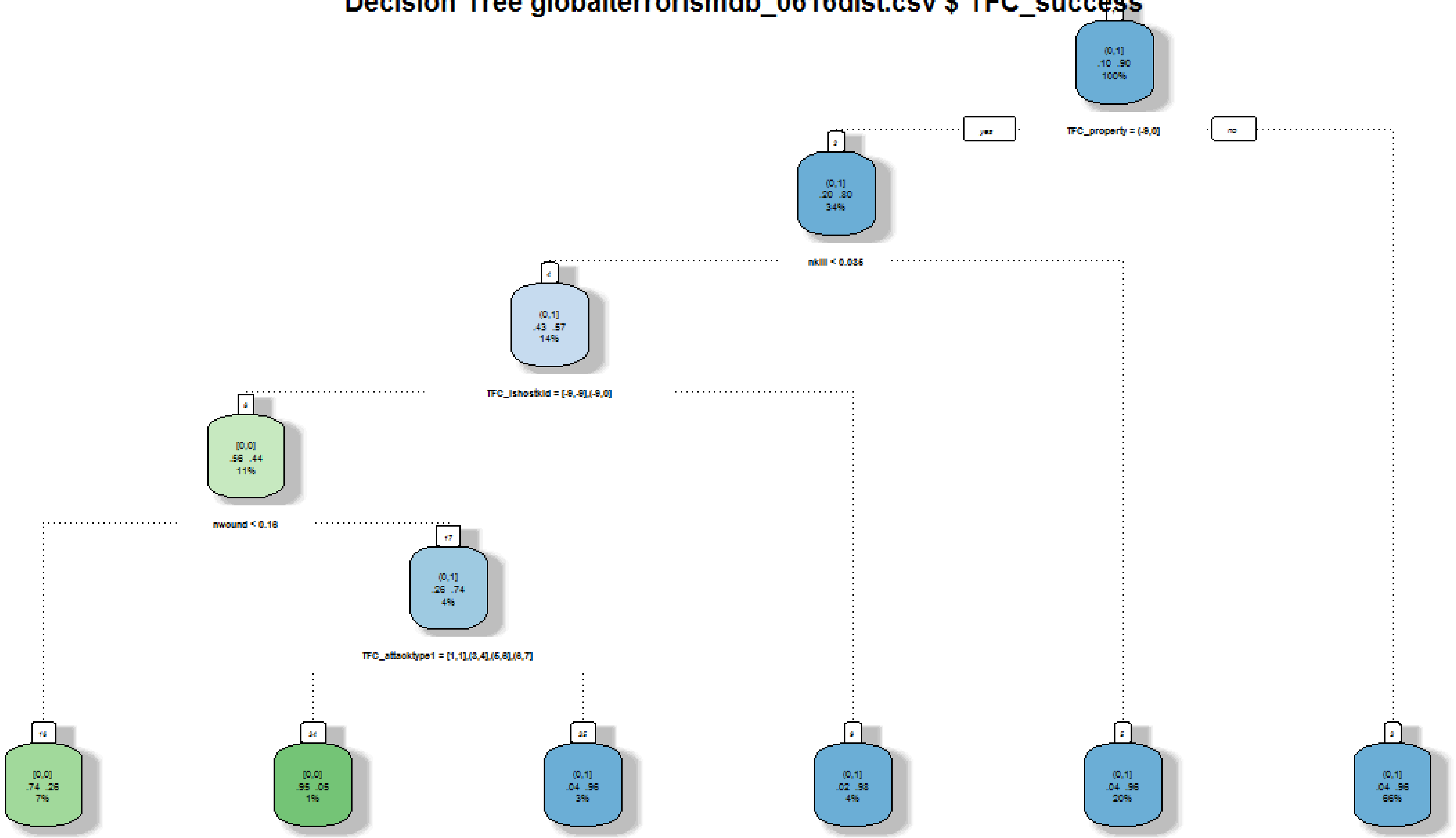
- Data Preparation for Decision tree modelling
  - The final dataset 24 variables with 3 numeric variables and 21 categorical variables
  - The mean for total number and kills and wounded were around 2 and 3 respectively.
  - The most active group was Taliban followed by shining path
  - The maximum number of kills and wounded in any attack 1500 and 5500 respectively
  - In the following sections by decision tree modelling we will try to analyse what are the most significant factor for a successful terrorist attack with success as target variable

Decision tree modelling

- The Dataset was partitioned into 70/30 for training and testing:
- Identified target variable: Success
- Parameter:
  - Minimum split:500
  - Minimum:bucket:200
  - Max Depth:30
  - Complexity:0.0400

Decision tree modelling

Decision Tree globalterrorismdb_0616dist.csv $ TFC_success



EVALUATION

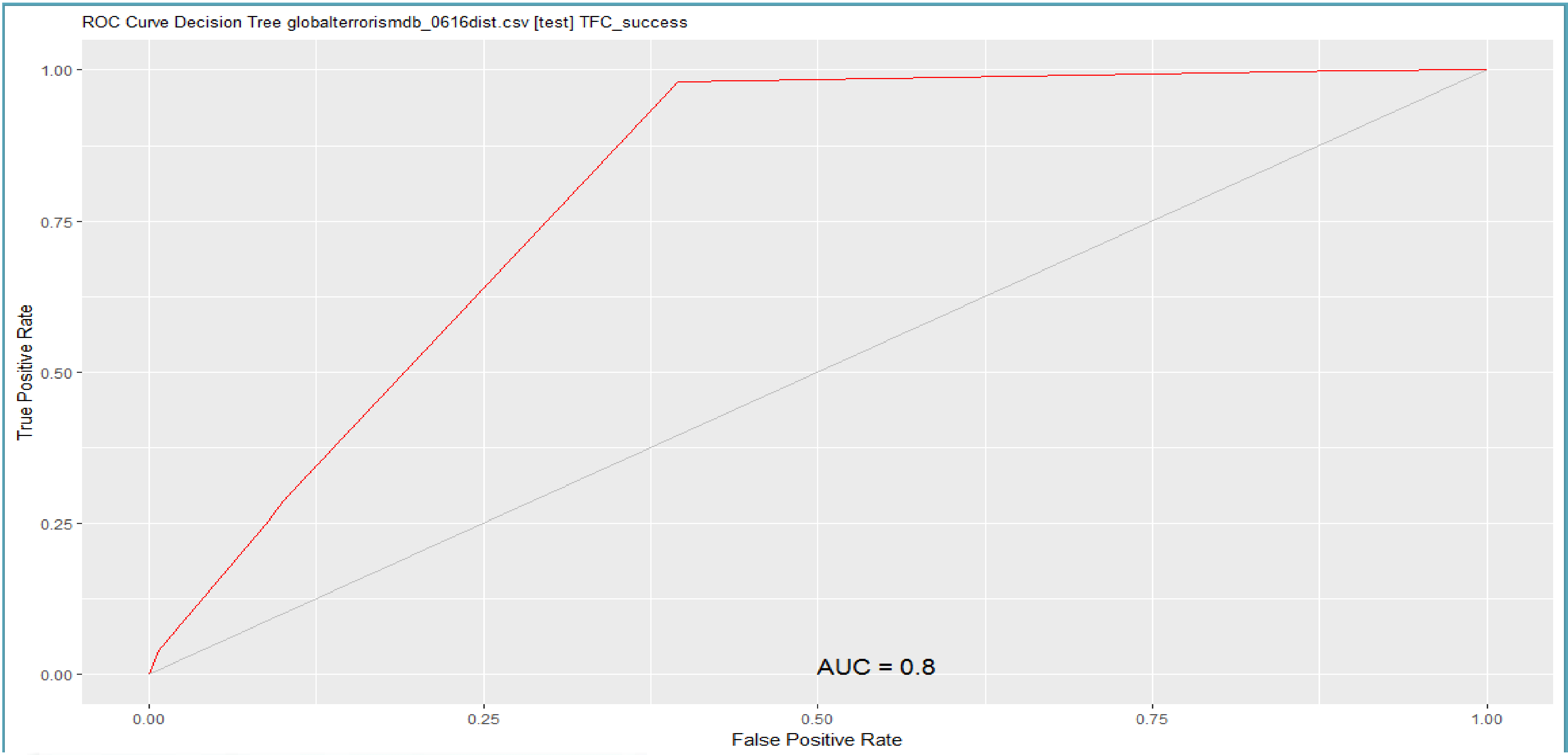
- **According to our Decision tree model 30% of the attacks were successful and variables actually used in our model are Number of people killed and wounded and attack type whether it was an assassination, bombing, kidnapping etc., Other variables used were, was anyone taken hostage and was there a damage to property. So these were the main features of our successful attack according to our model.**

- **Furthermore our decision tree tell us that from 34% successful, in 20% attacks there was higher chance that people got killed whereas in 14% people were left wounded.**

- **The attacks which left more wounded were generally hostage taking or hijacking , the attack type being armed assault, assassination , barricade hijacking etc.**

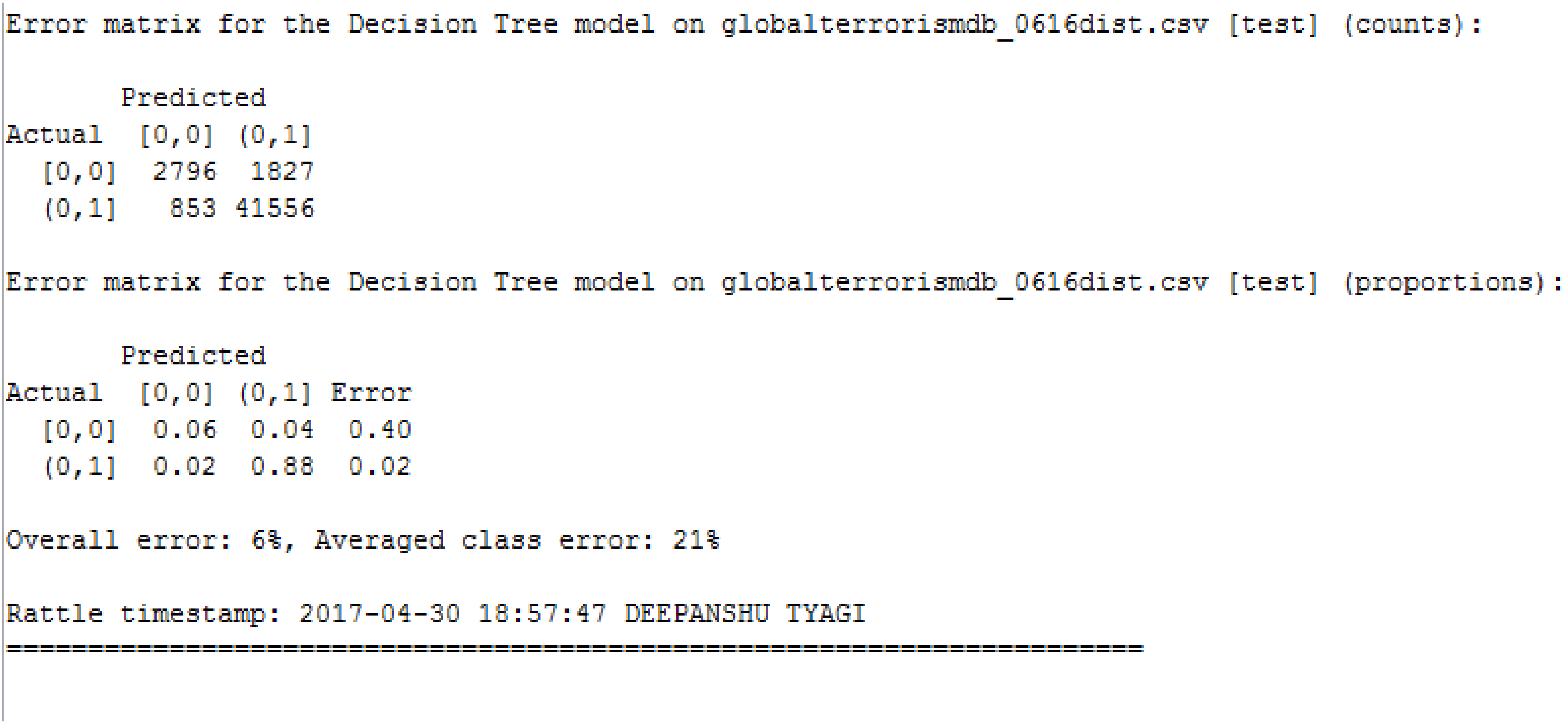
EVALUATION

- Below is the ROC curve which shows that Area under curve the is 0.8 which means that our model has good performance as can be seen in the previous slide that most of the results e got were intuitive enough to believe



EVALUATION

- **Below we can see that our overall error is 0.06% means that our model is relatively good for predicting what are the factors around the success of a terrorist attack.**



```
Error matrix for the Decision Tree model on globalterrorismdb_0616dist.csv [test] (counts):

       Predicted
Actual  [0,0] (0,1]
  [0,0]  2796  1827
  (0,1]   853 41556

Error matrix for the Decision Tree model on globalterrorismdb_0616dist.csv [test] (proportions):

       Predicted
Actual  [0,0] (0,1] Error
  [0,0]  0.06  0.04  0.40
  (0,1]  0.02  0.88  0.02

Overall error: 6%, Averaged class error: 21%

Rattle timestamp: 2017-04-30 18:57:47 DEEPANSHU TYAGI
======================================================================
```

Network analysis

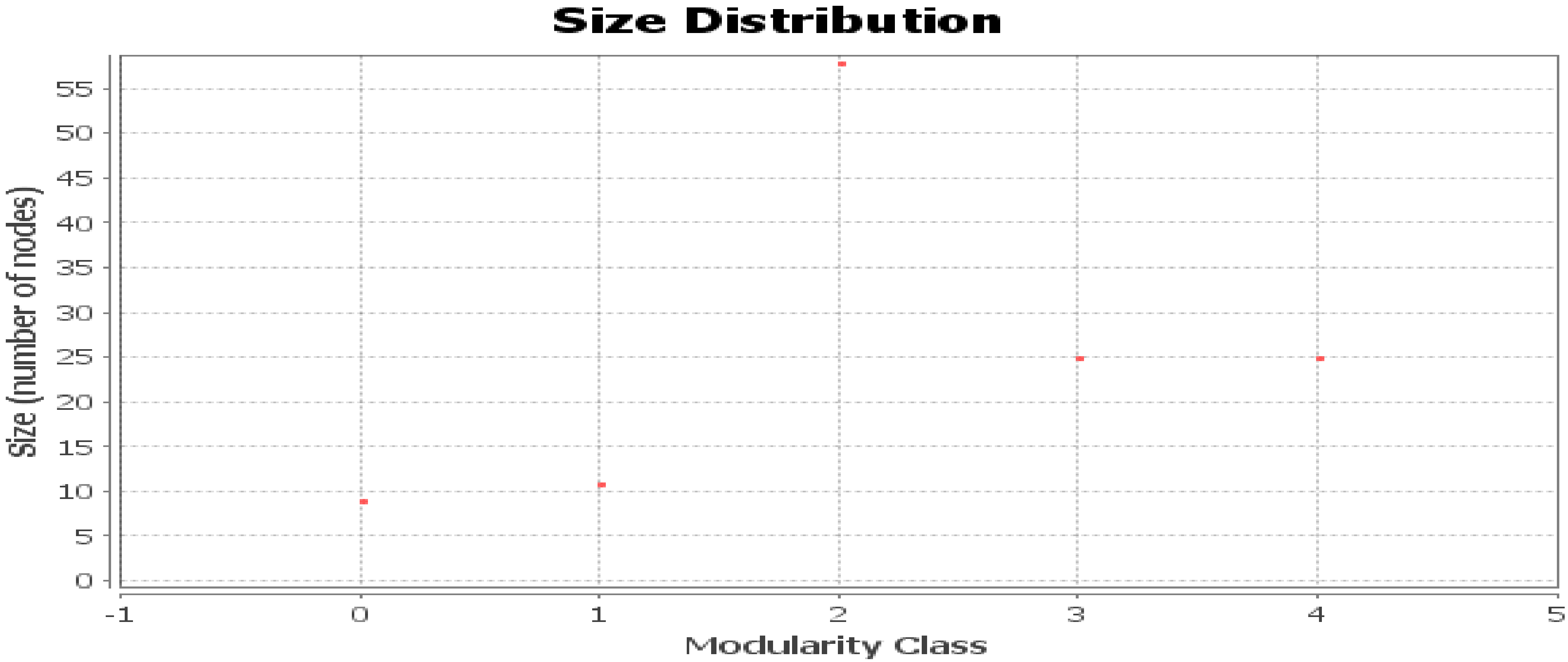
- As mentioned earlier we prepared the data for network analysis as the dataset had lot of informational columns including full summary, note and various other text columns ,so it was good idea to perform network analysis.

- The network initially consisted of 22425 nodes and 203400 edges

- Parameter settings and layouts used:
  - Degree range settings:513-3518
  - Fruchterman Reingold
    - Area:35000
    - Gravity:0.5
    - Speed:4
  - Label Adjust
  - Nonoverlap

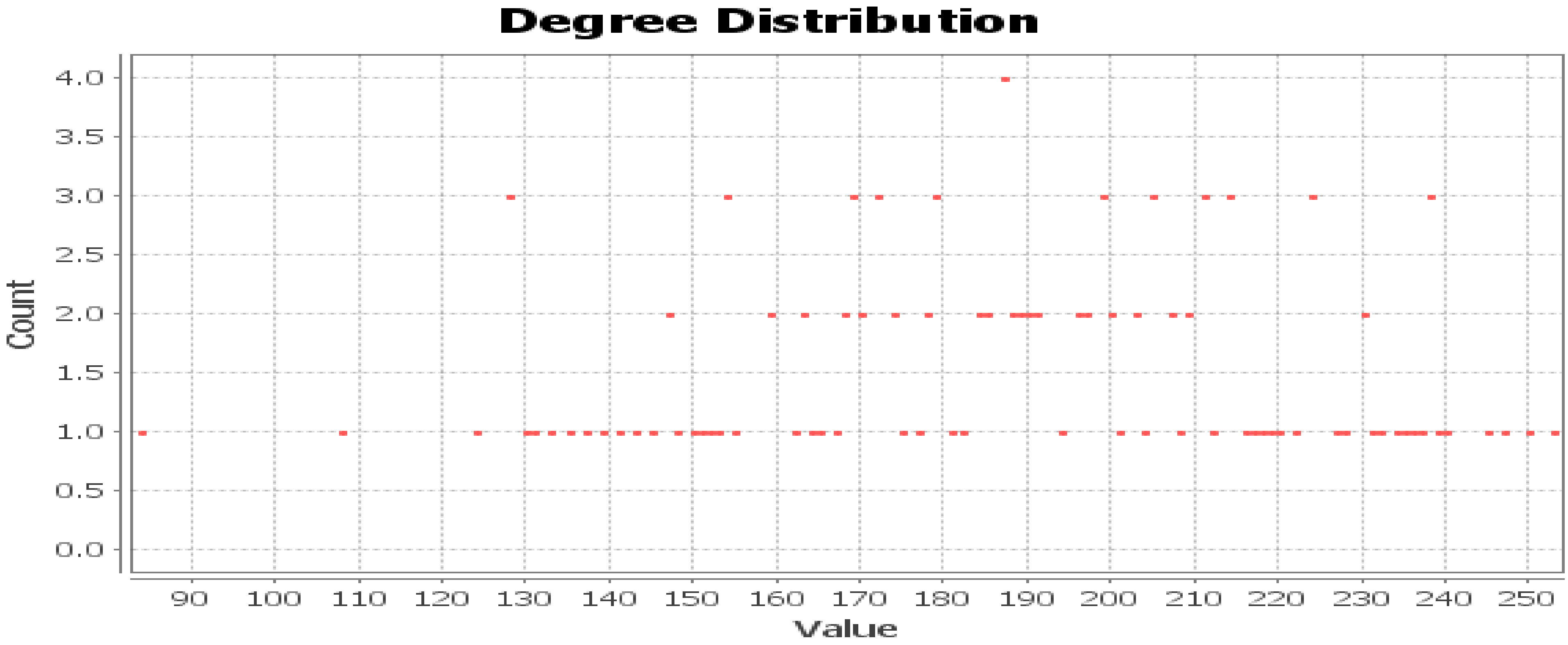
Network analysis

- **Statistics:**
  - **Modularity: 0.074**
  - **Number of communities: 5**
- **The graph below shows number of nodes in each modularity class which tells us the strength of the network, here we have 5 groups with group 2 having the largest sized Node at 55**



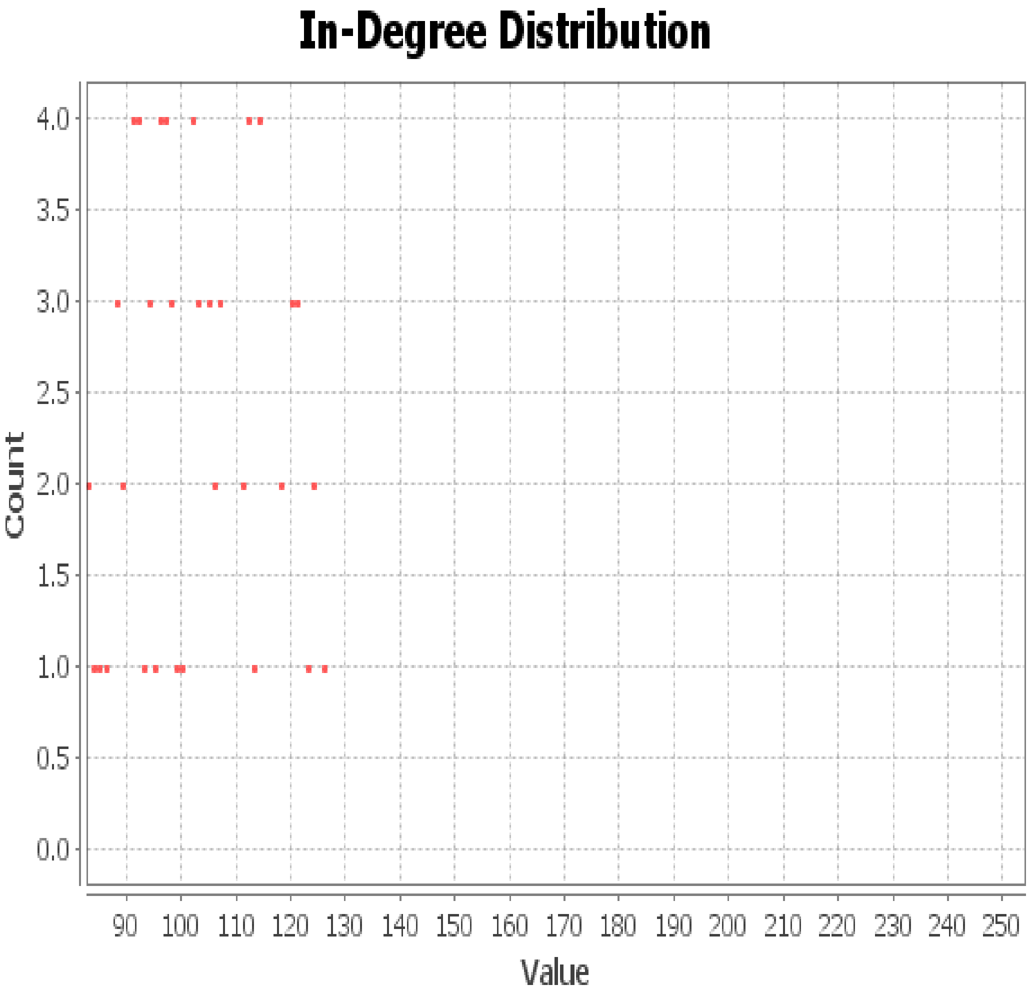
Network analysis

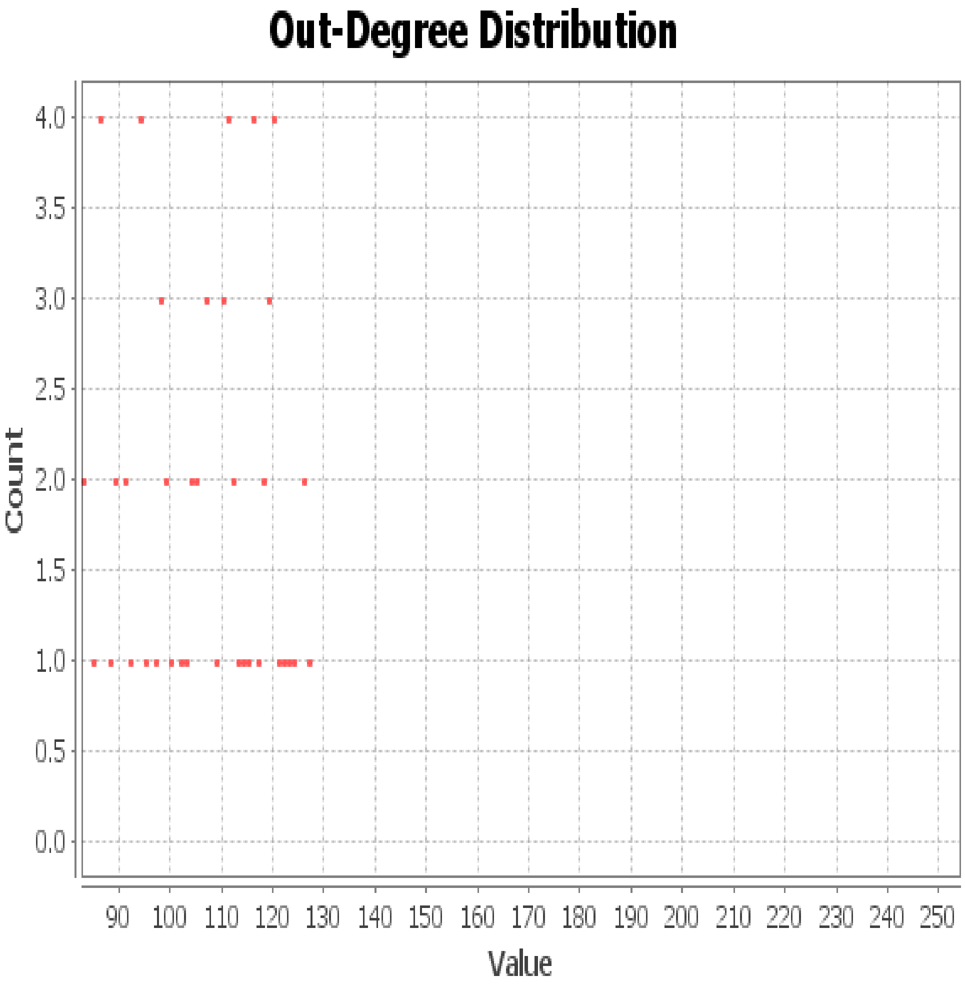
- **Statistics:**
  - **Average Degree: 93.672**
  - **The degree distribution 93.672 means that average connections for each node in our network are around 93 , Below is the graphical representation of in degree and out degrees which gives us an idea about on an average incoming and outgoing connections to a node.**



Network analysis

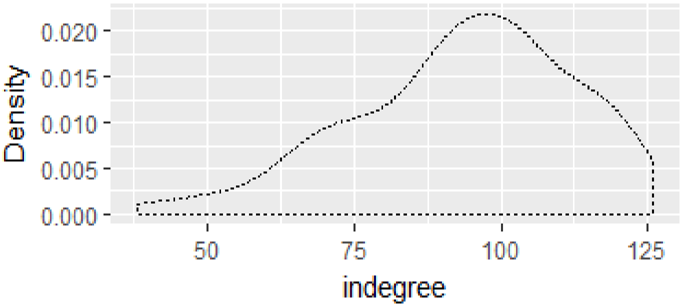
➢ The distributions below show us distribution of In-degree i.e. incoming connection and out-degree which is outgoing connections from a node





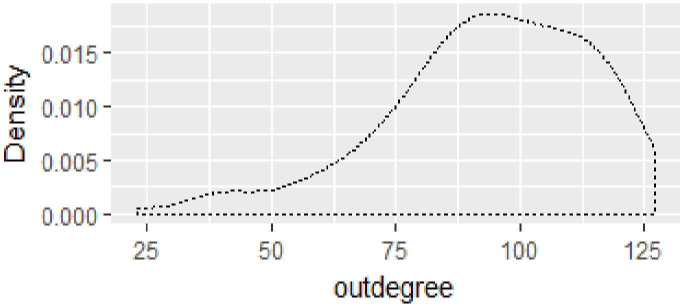
Looking at Distributions of various Statistics Used for Network Analysis

Distribution of indegree



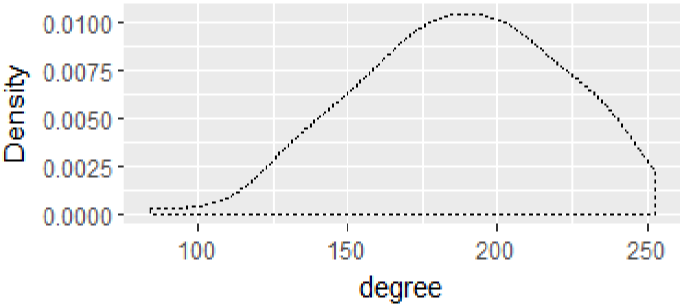
Rattle 2017-Apr-30 20:26:53 DEEPANSHU TYAGI

Distribution of outdegree



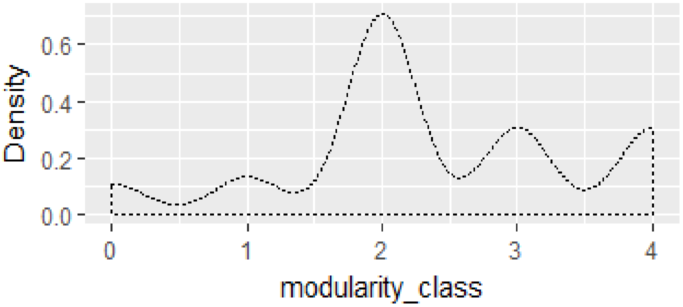
Rattle 2017-Apr-30 20:26:53 DEEPANSHU TYAGI

Distribution of degree



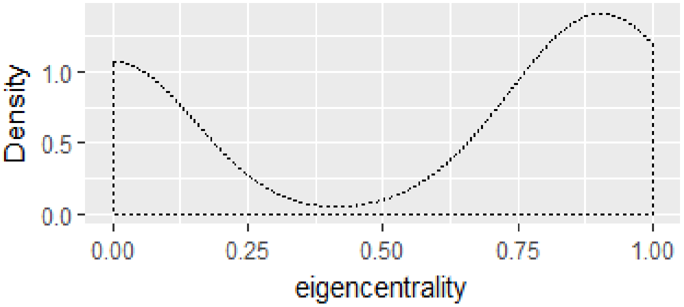
Rattle 2017-Apr-30 20:26:53 DEEPANSHU TYAGI

Distribution of modularity_class



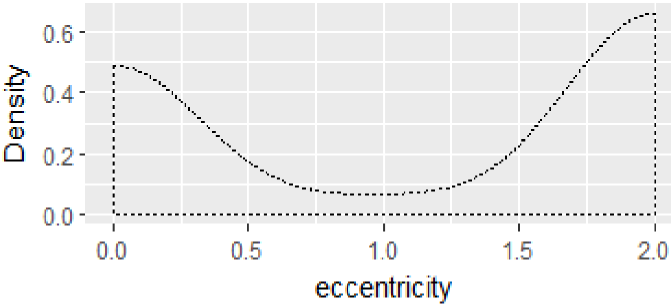
Rattle 2017-Apr-30 20:26:53 DEEPANSHU TYAGI

Distribution of eigencentrality



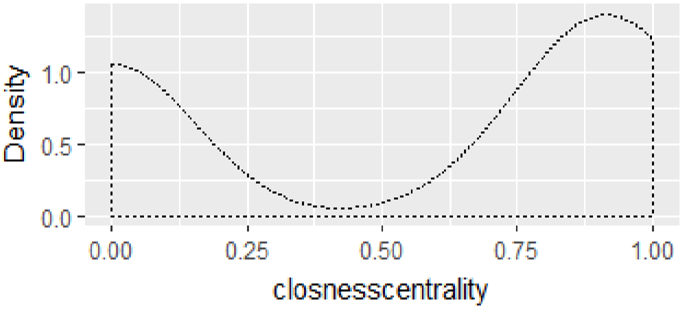
Rattle 2017-Apr-30 20:26:53 DEEPANSHU TYAGI

Distribution of eccentricity



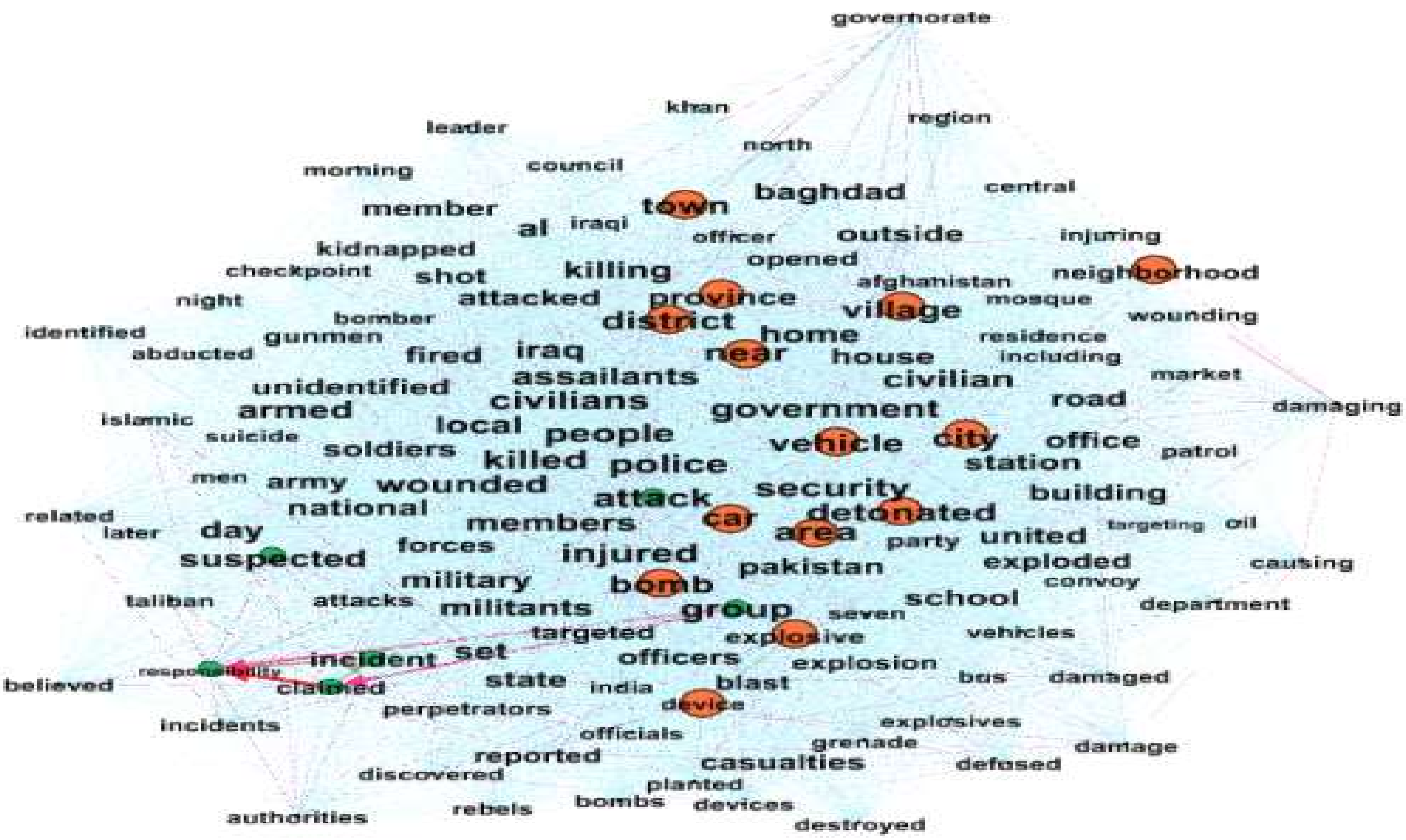
Rattle 2017-Apr-30 20:26:53 DEEPANSHU TYAGI

Distribution of closnesscentrality



Network analysis

- Below is the network showing connections between various nodes



Evaluation of Network analysis

- Network Analysis confirms that Kidnapping and Hostage taking was rampant and one of tactics used by the terrorists between 1970 and 2015 as was seen in our Decision tree model

- It also can be seen that most of the terrorists incidents in our datasets had a terrorist group that claimed the responsibility for he incident

- The Islamic countries were the most who suffered from these activities with Taliban being the most active terrorist outfit among others.

- The most attack types being bombings, explosions and kidnappings which is intuitive result and our decision tree model had these attributes in successful terrorist attacks

Conclusion

- The Dataset selected was ideal as it covered Both network analysis and supervised learning model decision trees giving us some useful insights

- The decision tree model with 80% accuracy confirmed that successful terrorist attack mostly left people killed or injured or damaged property with kidnapping and hijacking being rampant .

- Again Network analysis in a way confirmed our decision tree analysis as it showed that most influential nodes in our network were in line with our decision tree model.