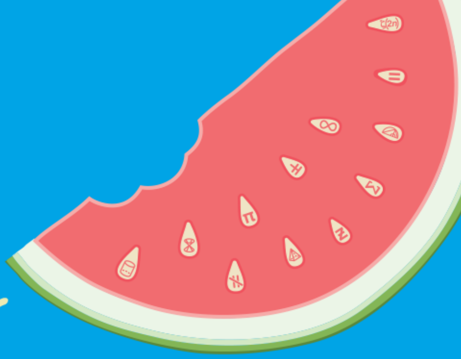
AMSI VACATION RESEARCH SCHOLARSHIPS 2021–22

Get a taste for Research this Summer



Introduction to Social Network Analysis

Nayani Ranasinghe

Supervised by Prof.Asha Rao and Dr.Sevvandi Kandanaarachchi

RMIT University

AMSI AUSTRALIAN MATHEMATICAL SCIENCES INSTITUTE

Acknowledgement

Firstly, I would like to express my gratitude to Australian Mathematical Sciences Institute (AMSI) for giving this opportunity to do a summer research project by granting a vacation research scholarship. This initial research work provided some valuable experience which I will require in my future studies.

I would like to express my deepest appreciation to the two supervisors Prof Asha Rao (Associate Dean, Mathematical Sciences, School of Sciences, RMIT University) and Dr Sevvandi Kandanaarachchi (Lecturer, Statistics – Mathematics, School of Sciences, RMIT University) for their exceptional support in my summer research project. Their enthusiasm, knowledge and exacting attention have been an inspiration and kept my work on track from the first encounter.

I owe thanks to my family members for their continuous encouragement, patience, and compassion shown during the time I undertook the research project.

Contents

1 Introduction **3**

1.1 Overview 3

1.2 Objectives of the study: 3

1.3 Statement of Authorship 4

2 Preliminaries **4**

2.1 Basic concepts in networks 4

2.1.1 What is a network? 4

2.1.2 Adjacency Matrix: 5

2.2 Centrality Measures. 5

3 Social Network Analysis with Enron Email dataset **8**

3.1 Enron Email dataset: 8

3.2 Extracting data from emails- The algorithm 8

3.3 Converting extracted data into network data 9

4 Findings **9**

4.1 Visualisation of the network using R tools [Boj21]: 9

4.2 Centrality measures and inferences 9

5 Discussion **13**

6 Conclusion **13**

Abstract

This project was conducted to understand the basics in Social network analysis (SNA) as well as graph theory and networks which underpin an SNA process. The theory studied at the beginning of the project was applied on a real data-set related to the Enron Scandal which has been investigated for money laundering in early 2000s. The findings of the project were the visualisations of the email- communication network and some of the measures central to the data obtained using R statistical software tools. These findings were used to understand which nodes could be the important and which nodes need to be removed from further analysis. Since the SNA process was not completed, the necessary steps for the completion has been discussed in the report.

1 Introduction

1.1 Overview

The world can be viewed in terms of different physical systems. These systems may include connected human relationships, transportation systems, communication systems, and computer systems such as Local Area Network (LAN), Metropolitan Area Network (MAN) or Wide Area Network (WAN). Analysis of these networks through social networks can lead to realisations of many other interesting topics such as culture, politics, education or even crime sometimes. Hence, modelling these social networks accurately hugely benefits society.

Social network analysis (SNA) focuses on patterns of relations among people and among groups [UKG16] . The process investigates social structures using networks and graph theory, the two broad fields that underpin SNA. The aim of the following study was to understand the basic concepts in SNA, graph theory and networks and explore necessary statistical tools that can handle Social network analysis. Followings are the objective of the study.

1.2 Objectives of the study:

1. Understanding the basic concepts in graph theory and networks.
2. Exploring statistical methodologies available in the R statistical programming language for network analysis.
3. Developing an algorithm to extract data from email messages in a real database related to business that has been investigated for money laundering.
4. Converting extracted data into network data.
5. Applying the statistical methodologies to analyse network data and visualise the communication network.
6. Understanding the important/ non-important nodes and thereby making inferences that could be important in network detection for money laundering.

1.3 Statement of Authorship

A Rao formulated the project idea based on previous research work done by P Magalingam and many others who have worked on the Enron Email data-set. The analysis and findings mentioned in the report were done by N Ranasinghe. Guidance and supervision was given by A Rao and S Kandanaarachchi. Project funding was provided by AMSI.

2 Preliminaries

2.1 Basic concepts in networks

2.1.1 What is a network?

A network is a graph model of a physical system. A graph can be identified as a collection of vertices and edges, with respect to a network [Che10]. For instance, with the following floor plan (figure 1- [Lyn21]), areas numbered from 1-6, can be represented using nodes whereas a doorway between two areas can be represented using edges as doorways allow the connection (relationship) between two areas (elements).

Example 1:

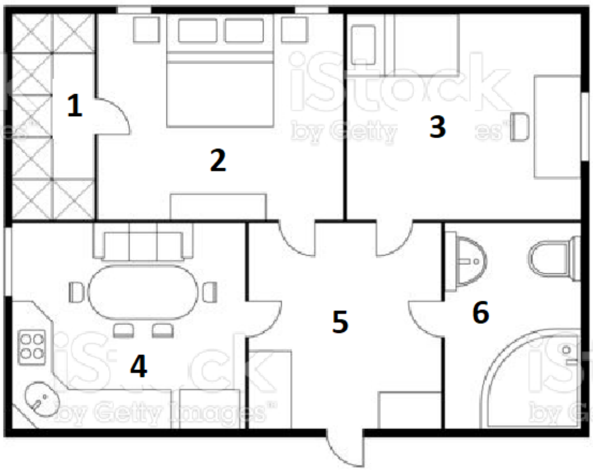


Figure 1: Floor plan

The relevant network for this physical system is shown in figure 2.

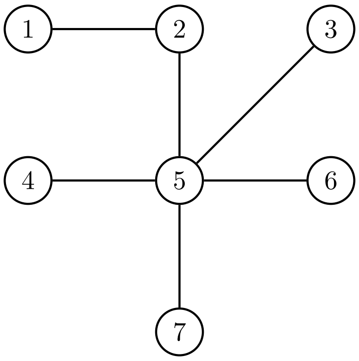


Figure 2: Network diagram1 for the floor plan

Network diagram in Figure 2 is merely about the components (areas) and entrances. It does not state the directions in which dwellers move between areas using doorways. Hence it is an undirected network. However, more information can be added to the graph by adding the number of times (weights) a doorway is used (extent of the relationship) per day.

2.1.2 Adjacency Matrix:

The mathematical way of representing a network is its adjacency matrix [Kev92]. The adjacency matrix (ignoring the edge between node 5 and node 7 which is outside world) for figure 2 is as follows.

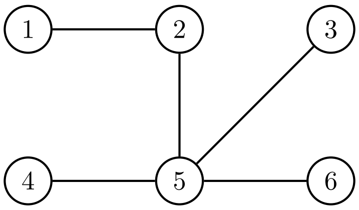


Figure 3: Network diagram2 for the floor plan

$$Adjacency matrix A =$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

2.2 Centrality Measures.

The degree centrality and eigenvector centrality (calculated using power iteration method- after 3 iterations) are as follows:

Node	Degree centrality	Eigenvector centrality
1	1	0.238909231
2	2	0.334472924
3	1	0.238909231
4	1	0.238909231
5	4	0.812291387
6	1	0.238909231

Table 1: Degree centrality's and Eigenvector centralities for the floor plan network

Here, degree centrality is the number of doorways connected to a particular area (have access from a particular room)/ how many nodes a particular node is connected to. For this network, nodes 5 and 2 have highest connections respectively. Therefore, these have higher degree centrality values than the rest of the nodes which have the least (connected with only one other area) connections and least degree centrality values.

Eigenvector centrality tells how important the connected neighbours are of a given node [Che10]. In the above network, node 4 is connected to only node 5 which is the node holds the most important position of the network. The node 2 is connected to node 5 as well as 1 which has a low degree centrality. The eigen vector centrality for node 4 and 2 are 0.238909231 and 0.334472924; node 4 has received a higher eigenvector centrality score.

In relation to the floor area network, area 5 has the highest number of connections with other areas, hence the highest degree centrality. Apart from area 5, area 2 is better connected with other areas. Therefore node 2 has the second highest eigenvector centrality. The eigenvector centrality can be understood better using the following network.

Example 2:

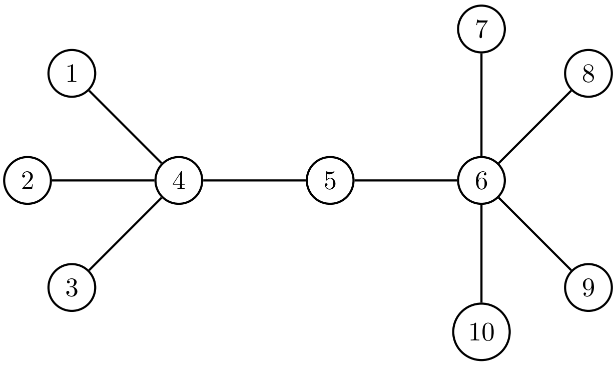


Figure 4: Network diagram 3

Node	DegreeCentrality	EigenvectorCentrality
1	1	0.019230769
2	1	0.019230769
3	1	0.019230769
4	4	0.307692308
5	2	0.076923077
6	5	0.480769231
7	1	0.019230769
8	1	0.019230769
9	1	0.019230769
10	1	0.019230769

Table 2: Degree centralities and Eigen vector centralities for the network 3

With figure 4, node 5 can be considerer a bridge since, if the node is removed the graph disconnects. Nodes 4 and 6 are connected to node 5 and has degree centralises 4 and 5 respectively. Their eigenvector centralities are 0.307692308 and 0.480769231, respectively. Using examples of the two networks, it could be concluded that the eigenvector centrality typically answers the question 'How well is this node/person/object connected to other well connected nodes/persons/objects?'.

Between centrality measures the extent to which a vertex lies on paths between other vertices [Che10]. For instance, with the floor area network, nodes 5 and 2 are located in between many other shortest paths to other nodes. The between centrality for nodes 2 and 5 are 4 and 9, respectively . This confirms that area 5 lies between higher number of paths to other floor areas than area 2 does.

Closeness centrality is the mean number of vertices in shortest paths that need to pass to reach other nodes [Che10]. With the floor plan network, closeness centrality means the number of doorways need to pass to get to another floor area from a given area.It can be expressed as a decimal (the closeness) by taking the inverse of sum of vertices. Higher the closeness centrality, faster the other nodes of the network can be reached [Che10].

node	Sum of edges	Closeness Centrality
1	12	0.08
2	8	0.13
3	9	0.11
4	10	0.10
5	6	0.17
6	10	0.10

Table 3: Closeness centrality measures for the floor plan network

3 Social Network Analysis with Enron Email dataset

3.1 Enron Email dataset:

This is a large Email database published in 2001 by Federal Energy Regulatory Commission (FERC), USA for transparency, historical and academic research purposes. Enron Corporation which had been using accounting loopholes and offshore platforms to conceal billions of dollars of debt in its financial reports for years declared bankruptcy on Dec. 2, 2001. The scandal became known worldwide and the email database containing a wide variety of information has been used in a very large range of studies and research projects worldwide [r521].

The original Enron email contains over 500000 email whereas for this study a fraction of it (1000) has been used.

3.2 Extracting data from emails- The algorithm

One of the difficult tasks of the study was extracting data from the database. For this purpose, an algorithm was developed; The algorithm was capable in extracting senders and receivers from emails, keeping track of each observation while cleaning the data set.

The pseudo-code for the algorithm (appendix 1) developed for this purpose is as follows:

```
Algorithm 1: Extracting senders and receivers
For i=1 to n do
1. Extract senderi from each email$i$
2. Extract receiversi from each email$i$
3. Keep tract of each observation$k$
If sender$i$ in not available and receiver$i$ not available
    Read the next email
Else if @ count is equal to 1
    Add 1 to the observation$k$ array
    Add sender to the sender$i$ array
    Add receiver to the receiver$i$ array
Else
    For $j = 1$ to @count

        Add sender to the sender array
        Add receiver to the receiver array
        Add 1 to the observation array
```

The 1000 emails had been sent to 1228 recipients in total at different occasions. Hence the initial data frame had 1228 entries. However, between a pair of email accounts, more than one email had been exchanged and there for it was necessary to find out the number of distinct entries. The 285 distinct number of relationships (edges) were 285 and this involved 265 distinct email addresses.

3.3 Converting extracted data into network data

For the construction of the network, email addresses (nodes) were labelled with distinct identity numbers. Thereby, edges were defined accordingly, and edge list and the adjacency matrix were obtained using the network package (appendix 2).

To coerce the edge list into an igraph object, graph from data frame function in igraph was used. The igraph objects were used to find centrality measures (appendix 3).

Network function which return a network class object was used to visualise the network (appendix 4).

4 Findings

4.1 Visualisation of the network using R tools [Boj21]:

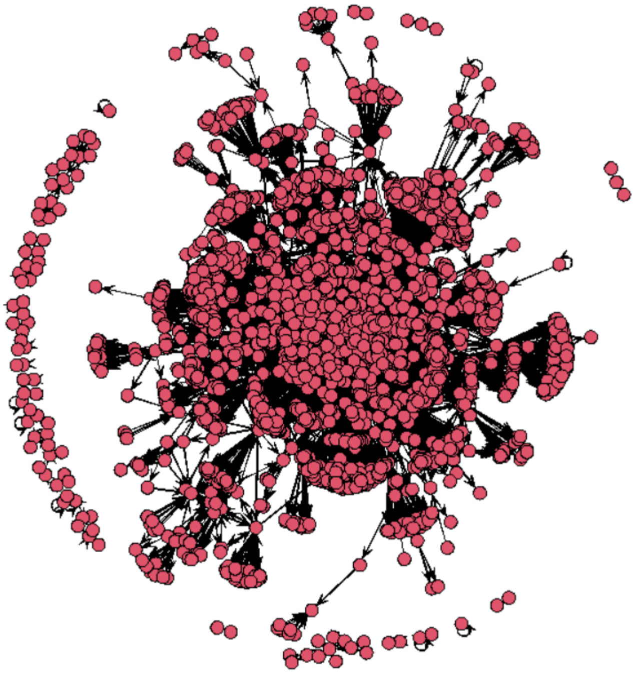
Figure 4: Network diagrams for first 10000 and 1000 emails (See page 10).

4.2 Centrality measures and inferences

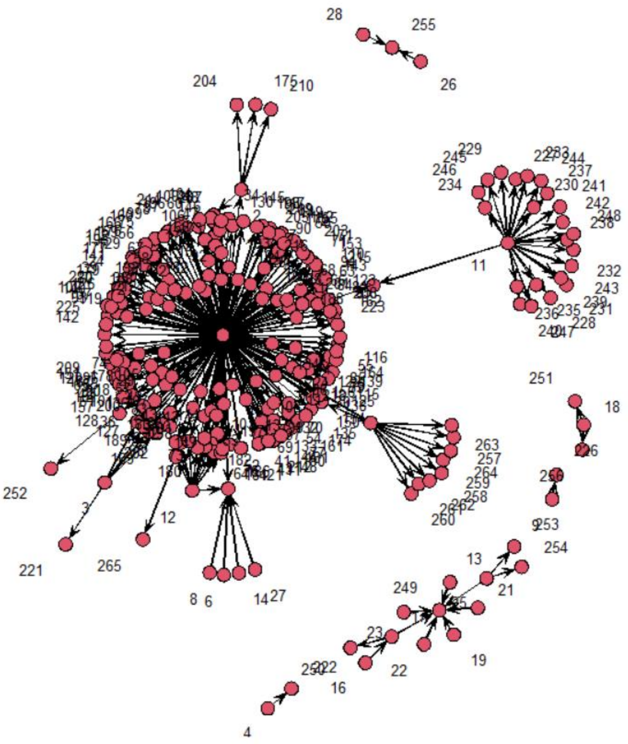
Using above codes, different centrality measures, and other measures such as graph density and reciprocity of the network could be obtained. However, to accomplish the purpose of the study, centrality measures were more important than others.

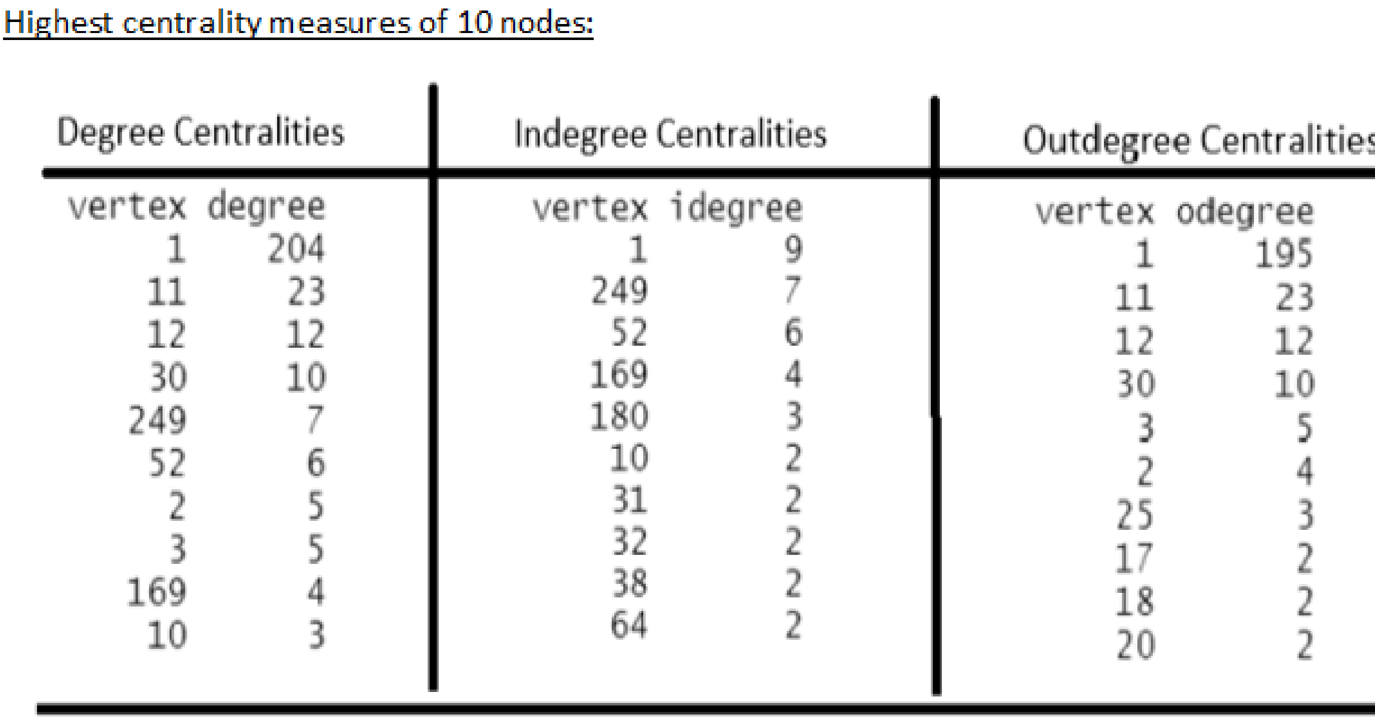
In figure 5, the first visualisation (with first 10000 emails) showed that there are many disconnected nodes from the main network. Therefore, they can be removed from the analysis. The second visualisation (with first 1000 emails) clearly tells a high volume of emails has been sent through certain email accounts where as the number of emails received into the accounts were few or negligible. This kind of nodes can be considered as bots. In other words computer programs that perform automated, repetitive tasks or pre-defined tasks. These are other type of nodes that could be removed from further analysis. In deciding what nodes need to be removed, in-degree centrality, out-degree centrality and degree centrality values play an important role.

Enron emails(first 10000)- network daiagram



Enron emails(first 1000)- network daiagram





Highest centrality measures of 10 nodes:

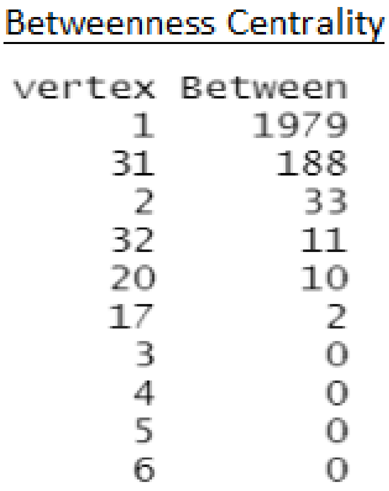
Degree Centralities		Indegree Centralities		Outdegree Centralitie	
vertex	degree	vertex	idegree	vertex	odegree
1	204	1	9	1	195
11	23	249	7	11	23
12	12	52	6	12	12
30	10	169	4	30	10
249	7	180	3	3	5
52	6	10	2	2	4
2	5	31	2	25	3
3	5	32	2	17	2
169	4	38	2	18	2
10	3	64	2	20	2

Table 4: In-degree, Out-degree and Degree centrality measures for first 1000 emails

As per table 4, Node 1 has the highest degree centrality, This indicates that the person/email account related to the node has had the highest number of relationships with other email accounts. However, the number of emails received is comparatively fewer to the number sent. Emails relate to nodes 11,12 and 30 have also sent out a high volume of emails compared to the number that has been received. The email account that has both sent and received emails were *phillip.allen@enron.com* The second highest number of emails and been sent through *tracy.arthur@enron.com* and received by *strawbale@crest.org* which is an outside account to Enron Corporation.

However, in identifying nodes involved in the scandal, the most useful centrality measure is the betweenness centrality. As betweenness centrality indicates the extent to which a vertex lies on the paths between other vertices. With Enron cooporation, these nodes could be considered as those with relay information, such a gateway nodes, or admin officers. Hence, in reconstructing the network for further analysis, these nodes need to be included and considered important in the analysis.

The ten nodes that have highest betweenness centralities are as follows:



Betweenness Centrality

```
vertex Between
     1    1979
    31     188
     2      33
    32      11
    20      10
    17       2
     3       0
     4       0
     5       0
     6       0
```

Table 5: Betweenness centrality values for first 1000 emails

Using the above values, it could be detected that following list of emails accounts and the connected email accounts to them need further investigation.

Node id	email account
1	*phillip.allen@enron.com*
31	*christi.nicolay@enron.com*
2	*ina.rangel@enron.com*
32	*christi.nicolay@enron.com*
20	*frank.hayden@enron.com*
17	*calxa@aol.com*

Table 6: Emails that has highest betweenness centrality (within first 1000 emails)

5 Discussion

The amount of work undertaken within the six weeks was sufficient to understand the nature of the communication network of Enron Corporation. The next step of the study will be the re-construction of the network. For a better analysis:

- the algorithm should be changed in such a manner that it could extract Bcc (Blind Copied) recipients and Cc (copied) recipients.

- the algorithm should be used on the whole database rather than on fractions of the database so that all the necessary data could be extracted for the investigation.

- unimportant nodes should be removed; For this purpose, another algorithm needs to be developed to remove such nodes based on in-degree and out- degree centrality measures.

6 Conclusion

Since the aim of the study was to understand the basics in graph theory, networks and an SNA process, only a fraction of Enron email database was used.

In analysing the Enron email data-set, the network data was obtained by extracting email accounts of senders and the main receivers of first 1000 email accounts. The problem with considering only the main receivers is that some other important receivers such as blind copy receivers are ignored. If an email has been Bcc'd to another party the main recipients are unaware of the fact that contents of the email is visible to a third party and hence, in detecting crime including such recipients in the data is important [Mag16].

Also, in re-constructing the network for further analysis, it is important to identify the important nodes and remove the unimportant ones. For instance, a large volume of emails have been sent out from some of the email accounts, however, two-way communications could be hardly seen. These email accounts could be considered bots, an autonomous program on the internet or another network that can interact with systems or users.

References

[Boj21] Michal Bojanowski. Introduction to network analysis tools in r, 2021.

[Che10] Giorgos Cheliotis. Social network analysis (sna): including a tutorial on concepts and methods, 2010.

[Kev92] Ali Keveh. Structural mechanics: Graph and matrix methods. pages 1–39, 1992.

[Lyn21] Allison Lynch. 2d floor plans with best free software, 2021.

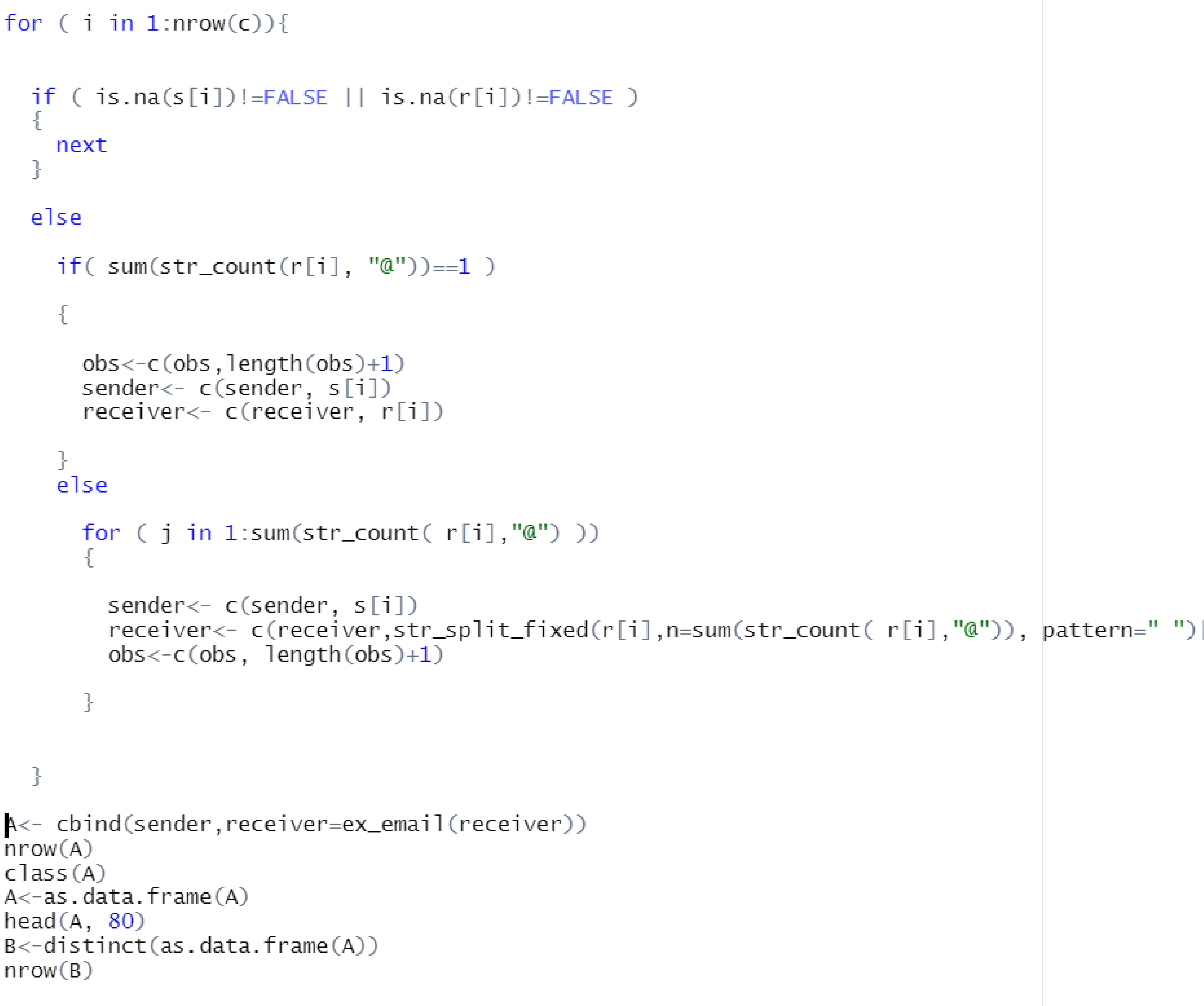
[Mag16] Preethiga Magalingam. Complex network tools to enable identification of a criminal community. *Bulletin of the Australian Mathematical Society*, 94(2):350—352, 2016.

[r521] Investigating enron's email corpus: The trail of tim belden, 2021.

[UKG16] UKGov. Social network analysis: 'how to guide', 2016.

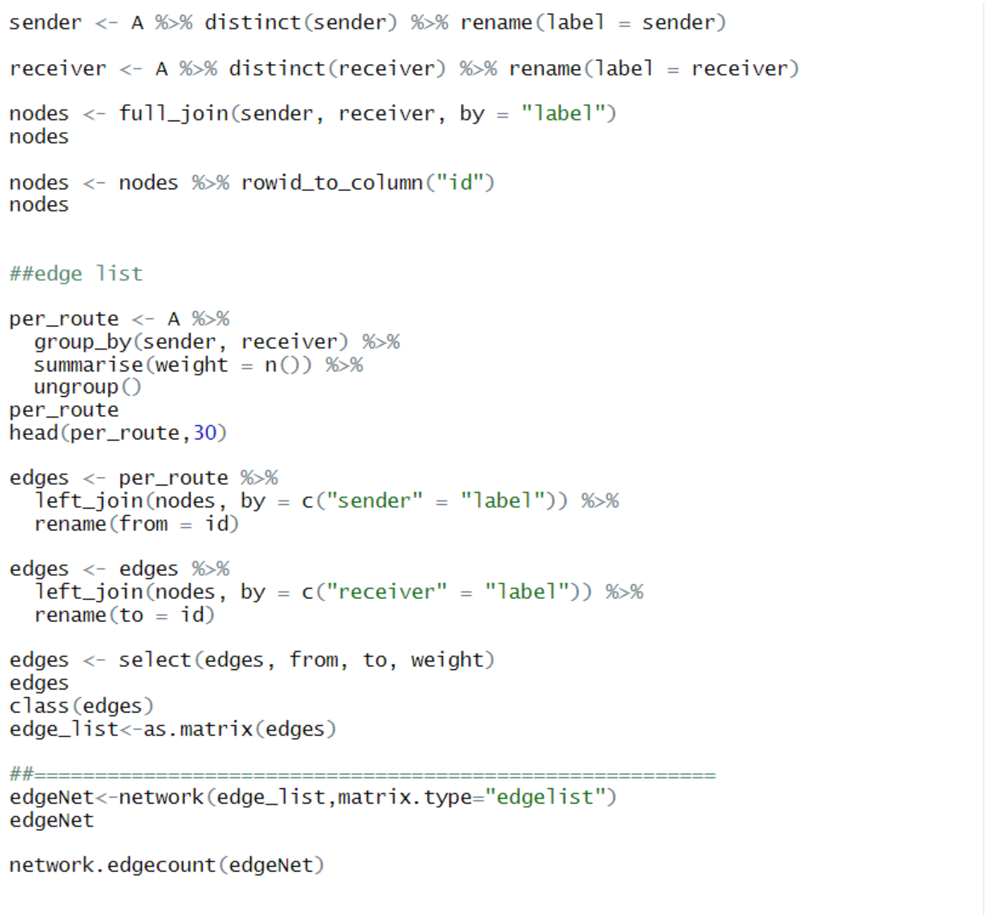
Appendices

Appendix 1



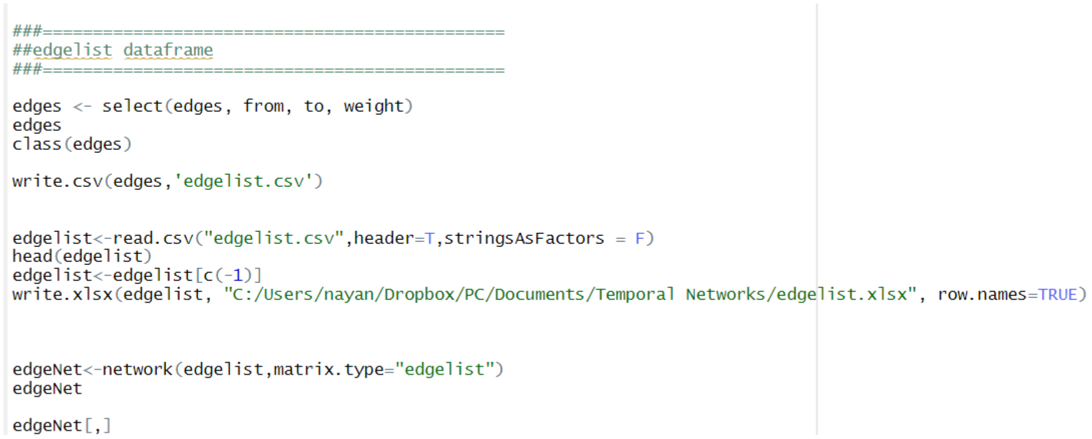
```
for ( i in 1:nrow(c)){


  if ( is.na(s[i])!=FALSE || is.na(r[i])!=FALSE )
  {
    next
  }

  else

    if( sum(str_count(r[i], "@"))==1 )

    {

      obs<-c(obs,length(obs)+1)
      sender<- c(sender, s[i])
      receiver<- c(receiver, r[i])

    }
    else

      for ( j in 1:sum(str_count( r[i],"@") ))
      {

        sender<- c(sender, s[i])
        receiver<- c(receiver,str_split_fixed(r[i],n=sum(str_count( r[i],"@")), pattern=" ")
        obs<-c(obs, length(obs)+1)

      }


  }

A<- cbind(sender,receiver=ex_email(receiver))
nrow(A)
class(A)
A<-as.data.frame(A)
head(A, 80)
B<-distinct(as.data.frame(A))
nrow(B)
```

Appendix 2



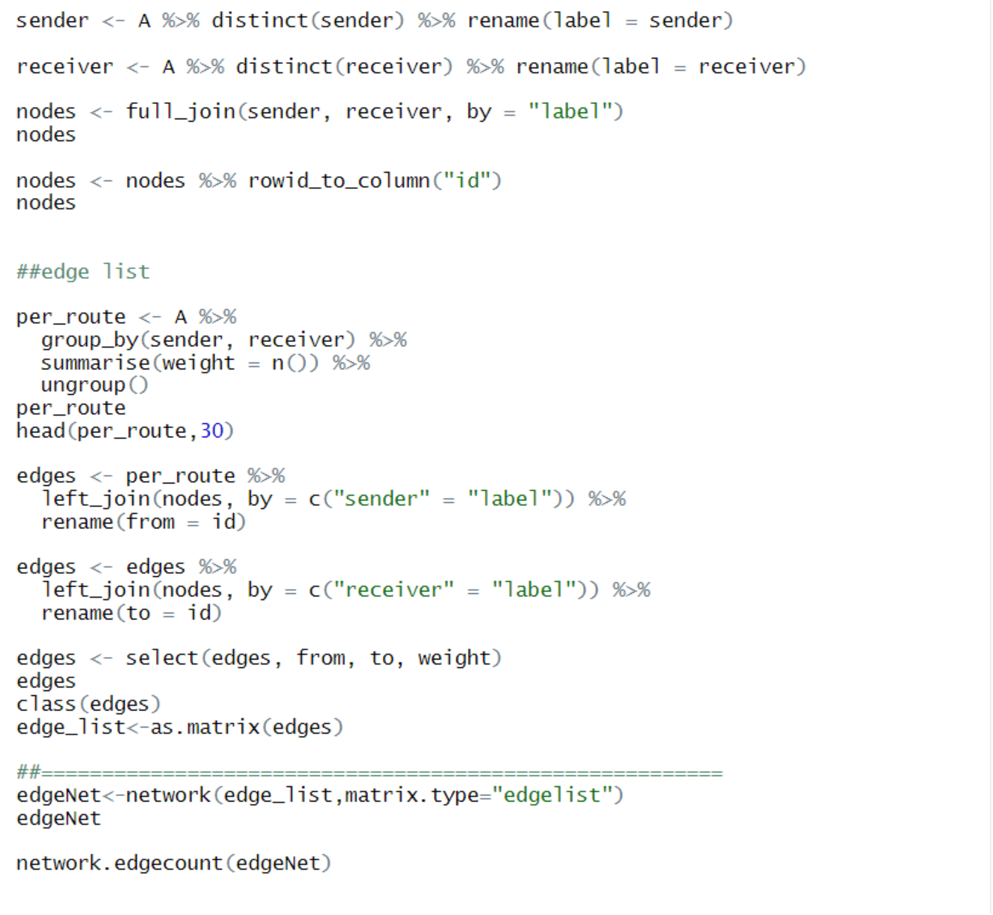
```
sender <- A %>% distinct(sender) %>% rename(label = sender)

receiver <- A %>% distinct(receiver) %>% rename(label = receiver)

nodes <- full_join(sender, receiver, by = "label")
nodes

nodes <- nodes %>% rowid_to_column("id")
nodes


##edge list

per_route <- A %>%
  group_by(sender, receiver) %>%
  summarise(weight = n()) %>%
  ungroup()
per_route
head(per_route,30)

edges <- per_route %>%
  left_join(nodes, by = c("sender" = "label")) %>%
  rename(from = id)

edges <- edges %>%
  left_join(nodes, by = c("receiver" = "label")) %>%
  rename(to = id)

edges <- select(edges, from, to, weight)
edges
class(edges)
edge_list<-as.matrix(edges)

##==========================================================
edgeNet<-network(edge_list,matrix.type="edgelist")
edgeNet

network.edgecount(edgeNet)
```

Appendix 3



```
###=============================================
##edgelist dataframe
###=============================================

edges <- select(edges, from, to, weight)
edges
class(edges)

write.csv(edges,'edgelist.csv')


edgelist<-read.csv("edgelist.csv",header=T,stringsAsFactors = F)
head(edgelist)
edgelist<-edgelist[c(-1)]
write.xlsx(edgelist, "C:/Users/nayan/Dropbox/PC/Documents/Temporal Networks/edgelist.xlsx", row.names=TRUE)



edgeNet<-network(edgelist,matrix.type="edgelist")
edgeNet

edgeNet[,]
```

Appendix 4



```
sender <- A %>% distinct(sender) %>% rename(label = sender)

receiver <- A %>% distinct(receiver) %>% rename(label = receiver)

nodes <- full_join(sender, receiver, by = "label")
nodes

nodes <- nodes %>% rowid_to_column("id")
nodes


##edge list

per_route <- A %>%
  group_by(sender, receiver) %>%
  summarise(weight = n()) %>%
  ungroup()
per_route
head(per_route,30)

edges <- per_route %>%
  left_join(nodes, by = c("sender" = "label")) %>%
  rename(from = id)

edges <- edges %>%
  left_join(nodes, by = c("receiver" = "label")) %>%
  rename(to = id)

edges <- select(edges, from, to, weight)
edges
class(edges)
edge_list<-as.matrix(edges)

##=========================================================
edgeNet<-network(edge_list,matrix.type="edgelist")
edgeNet

network.edgecount(edgeNet)
```