NETWORK NEIGHBORHOOD ANALYSIS FOR DETECTING ANOMALIES IN TIME SERIES OF GRAPHS

by

Suchismita Goswami

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Computational Science and Informatics

Committee:

_____	Dr. Igor Griva, Committee Chair
_____	Dr. Edward Wegman, Committee Co-Chair
_____	Dr. Jeff Solka, Committee Member
_____	Dr. Dhafer Marzougui, Committee Member
_____	Dr. Jason Kinser, Acting Department Chairperson
_____	Dr. Donna Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science
_____	Dr. Peggy Agouris, Dean, College of Science

Date: _____ Spring Semester 2019
George Mason University
Fairfax, VA

Network Neighborhood Analysis For Detecting Anomalies in Time Series of Graphs

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Suchismita Goswami
Master of Science
George Mason University, 2013
Master of Science
State University of New York, Stony Brook, 2001

Chair: Dr. Igor Griva, Professor
Department of Computational and Data Sciences

Spring Semester 2019
George Mason University
Fairfax, VA

Dedication

I dedicate this dissertation to my parents, Professor Rabindranath Ganguli and Basanti Ganguli.

Acknowledgments

I would not have done this research work without the great teachers in my undergraduate and graduate courses. In particular, I would like to thank Dr. Stephen Finch at SUNY, Stony Brook, Dr. Clifton D. Sutton at GMU, Dr. Bijoy Biswas and Dr. Probodh Bhowmic at Krishnath College, Berhampore, India, who made this possible.

I would like to thank my dissertation director and adviser Dr. Edward J. Wegman for his support and guidance, and for making this work possible. I also thank him for suggesting this work and providing me with the data. I am also thankful to him for providing me the opportunity to present this work at the Symposium on Data Science and Statistics (SDSS) in May 2018 in Reston, VA, and at the International Conference on Computational Social Science (IC^2S^2-2018), Kellogg School of Management at North Western University, Chicago. In addition, I thank my committee members, Dr. Igor Griva, Dr. Jeff Solka and Dr. Dhafer Marzougui, for their encouragements and suggestions.

I would like to thank Dr. Edward Wegman for offering Measure theory and Linear Space, Data Mining, and Text Mining, Dr. Clifton Sutton for Applied Statistics, and Statistical Inference, Dr. James Gentle for Time Series Analysis, and Dr. Yunpeng Zhao for Network Modeling. Those courses helped me enormously in carrying out this research work.

I would like to thank my mother, Basanti Ganguli and my father, Professor Rabindranath Ganguli, who is my friend, philosopher and guide, for encouraging me to pursue a PhD. My father taught me how to be mindful, which helped me enormously to carry out my PhD studies. I remember that he used to recite a Sloka written in Sanskrit, ...'Gururevo Param Brahma, Tasmai Shree Gurave Nama', which can be approximately translated as Gurus or Teachers are the supreme beings or enlightened persons, and I bow down or salute those Gurus.

I would like to thank my friends and classmates at GMU, at SUNY at Stony Brook, and at Krishnath College, Berhampore, India for their encouragements. In addition, I would like to thank my sisters, Mithu and Mita, for their support at various stages of my academic pursuits. Lastly, I would like to give a big THANK to Ramasis for giving me the support and the encouragement, for telling me every morning to stay focused and positive, and always being there during my PhD studies. I couldn't have come this far without you.

Table of Contents

List of Tables

List of Figures

Abstract

NETWORK NEIGHBORHOOD ANALYSIS FOR DETECTING ANOMALIES IN TIME SERIES OF GRAPHS

Suchismita Goswami, PhD

George Mason University, 2019

Dissertation Director: Dr. Igor Griva

Around terabytes of unstructured electronic data are generated every day from twitter networks, scientific collaborations, organizational emails, telephone calls and websites. Excessive communications in communication networks, particularly in organizational e-mail networks, continue to be a major problem. In some cases, for example, Enron e-mails, frequent contact or excessive activities on interconnected networks lead to fraudulent activities. Analyzing the excessive activity in a social network is thus important to understand the behavior of individuals in subregions of a network. In a social network, anomalies can occur as a result of abrupt changes in the interactions among a group of individuals. Therefore, one needs to develop methodologies to analyze and detect excessive communications in dynamic social networks. The motivation of this research work is to investigate the excessive activities and make inferences in dynamic sub networks. In this dissertation work, I implement new methodologies and techniques to detect excessive communications, topic activities and the associated influential individuals in the dynamic networks obtained from organizational emails using scan statistics, multivariate time series models and probabilistic topic modeling. Three major contributions have been presented here to detect anomalies of dynamic networks obtained from organizational emails.

At first, I develop a different approach by invoking the log-likelihood ratio as a scan statistic with overlapping and variable window sizes to rank the clusters, and devise a two-step scan process to detect the excessive activities in an organizations e-mail network as a case study. The initial step is to determine the structural stability of the e-mail count time series and perform differencing and de-seasonalizing operations to make the time series stationary, and obtain a primary cluster using a Poisson process model. I then extract neighborhood ego subnetworks around the observed primary cluster to obtain more refined cluster by invoking the graph invariant betweenness as the locality statistic using the binomial model. I demonstrate that the two-step scan statistics algorithm is more scalable in detecting excessive activity in large dynamic social networks.

Secondly, I implement for the first time the multivariate time series models to detect a group of influential people and their dynamic relationships that are associated with excessive communications, which cannot be assessed using scan statistics models. For the multivariate modeling, a vector auto regressive (VAR) model has been employed in time series of subgraphs in e-mail networks constructed using the graph edit distance, as the nodes or vertices of the subgraphs are interrelated. Anomalies or excessive communications are assessed using the residual thresholds greater than three times the standard deviations, obtained from the fitted time series models.

Finally, I devise a new method of detecting excessive topic activities from the unstructured text obtained from e-mail contents by combining the probabilistic topic modeling and scan statistics algorithms. Initially, I investigate the major topics discussed using the probabilistic modeling, such as latent Dirichlet allocation (LDA) modeling, then employ scan statistics to assess the excessive topic activities, which has the largest log likelihood ratio in the neighborhood of primary cluster.

These analyses provide new ways of detecting the excessive communications and topic flow through the influential vertices in a dynamic network, and can be extended in other dynamic social networks to critically investigate excessive activities.

Chapter 1: Introduction

Anomalies, which are clusters of events or excessive or unusual activities, are common in science and technology. Some of the most commonly used methods for anomaly detection in data mining are density-based techniques such as k-nearest neighbor [KNT00] and local outlier factor [BKNS00], one class support vector machines [SPST$^+$01], neural networks [HHWB00], cluster analysis-based outlier detection [HXD03] and ensemble techniques [LK05]. All these methods used to detect excessive activity, are mostly descriptive in nature, and not effective in making statistical inferences. In other words, these methods do not predict if these observed clusters of events are statistically significant or not [Kul79]. A very powerful statistical inference methodology that has been developed to detect the region of unusual activity in a random process and to infer the statistical significance of the observed excessive activity is scan statistics [Kul79], which is also termed as moving window analysis in the engineering literature and has mostly been used in spatial statistics and image analysis.

Scan statistic is defined as a maximum or minimum of local statistics estimated from the local region of the data. Let $\{X_t,\, t \geq 0\}$ be a Poisson process with rate, λ, where X_t is the number of points (events) occurring in the interval [0, t). In any subinterval of [0, T) of length, w, let Y_t be the number of points (events) in a window of the interval, $[t, t+w)$, such that $Y_t = X_{t+w}$ - X_t. The one-dimensional continuous scan statistic, S_w, is written as [GB99]:

$$S_w = \max_{0<t\leq T-w} \; Y_t(w). \tag{1.1}$$

In other words, the scan statistic, S_w, is the largest number of points that are observed in any subinterval of [0, T) of length w. In this Poisson process, λ is the expected (average)

number of events in any unit interval and the number of points (events) in any interval, $[t, t+w)$, Y_t follows a Poisson distribution with mean λw. The probability mass function for the random variable, Y_t, can be written as:

$$P\left(Y_t(w) = k\right) = \frac{e^{-\lambda w}(\lambda w)^k}{k!}. \tag{1.2}$$

In one-dimensional setting, it has been used by a number of authors to investigate the unusual clusters of events in various fields, for example, in visual perception [Gla79], molecular biology [KB92], epidemiology [WN87], queueing theory [Gla81], material science [New63], and telecommunication [Alm83]. Public health officials are often interested in finding explanations of clusters of cancer cases. Kulldorff et al. [Kul79] have assessed the unusually large number of brain cancer cases using spatial scan statistics. In medical imaging or screening situation, detection of abnormalities in structural images is very important. In this situation, a one-dimensional scan statistic model may not be adequate for cluster detection. It is, therefore, extended to two or three dimensional settings to study the mammography images. Priebe et al. [POH98] have exploited the stochastic scan partitions in the mammography images by studying the texture of the breast, and evaluated clusters of breast calcification using spatial scan statistics, and provided an exact sampling distribution of the spatial scan statistic under the null hypothesis of homogeneity. For the non-homogeneous mammogram, a p value of 0.034 has been reported by Priebe et al [POH98].

Although considerable work has been done to detect clusters of events or anomaly using scan statistics in spatial statistics and image analysis, relatively less attention has been given to detect anomaly in social networks, where lots of interaction take place among individuals or group of individuals. In addition to sharing knowledge and experience with one another in social networks, each individual develops a pattern of interactions. Anomaly in social network occurs when some individuals or group of individuals make sudden changes in their patterns of interactions. Research on social networks includes both static and dynamic social relations. Methods and theorems from graph theory and statistics are used intensively in

analyzing social networks. Despite the fact that the majority of research focuses greatly on static networks, they fail to capture information flow in dynamic networks.

Recently, a number of measures and algorithms have been developed for dichotomous and symmetric relation matrices in the analysis of dynamic networks. As the amount of unstructured electronic data created by various social networks, for example, twitter network, research network, a network of scientific collaboration, organizational e-mails and telephone calls increases enormously to terabyte range day by day, the need for tools and techniques to analyze such unstructured massive data sets has grown. In some cases, one needs to analyze the excessive activity in a social network to understand the behavior of the network. The motivation of this research work is to investigate the excessive activities in the network data from organizational e-mails by implementing statistical models and data mining algorithms, particularly scan statistics, time series models and content analysis. The methodologies developed here can be applied to other dynamic networks to assess excessive activities in the network.

1.1 Prior Work on Network Data Using Scan Statistics

Priebe et al. [PCMP05] first applied temporal scan statistics to network data to detect anomaly in time series of graphs, obtained from Enron email data. The full network was partitioned into disjoint subregions or subnetworks over time, which results in a collection of graphs or a time series of graphs. The large networks have computational difficulties, and the visualization and statistical inference are almost impossible to apply to a global network. An alternating approach to identify interesting features at a specific point of time is to split the global network as subnetworks or subregions, and consider the network neighborhood analysis as opposed to global network analysis. The subregions are modeled subsequently by directed graphs indexed by time, D_t, which are a collection of vertices that are joined by edges. Graph, D_t, can therefore be expressed as $D_t = (V, E_t)$, where each graph has the same set of vertices, V, and different set of edges, E_t. The order of the graph, D_t, is $n = |V|$ = number of vertices, and size of graph, D_t, is $m = |E_t|$ = number of edges. The

adjacency matrix of D_t is defined as $A = (A_{ij})$ such that

$$A_{ij} = \begin{cases} 1, & \text{if there is an edge between vertices i and j,} \\ 0, & \text{otherwise.} \end{cases}$$

From the graph at each time point one can obtain local regions or neighborhoods of vertices. The k^{th} order neighborhood of a vertex, v, of the network, D_t, is defined as:

$$N_k[v; D_t] = \{u \in V : d_t(u, v) \leq k;\ k = 0, 1, 2, ..\}, \tag{1.3}$$

where $d_t(u, v)$ is the geodesic distance, which is in fact the shortest path between u and v within k. A family of sub graphs induced by neighborhood denoted by $\Omega(N_k[v; D_t])$ with a set of vertices, $N_k[v; D_t]$, can then be obtained.

To quantify the characteristics of a node in subgraphs, the graph invariant feature, such as degree, can be used. Priebe et al. [PCMP05] used outdegree as the locality statistics. The degree of a node in a graph is the number of direct connections incident on it, while in the directed network, the degree is defined based on both in-degree and out-degree, where the in-degree is the number of inward edges and the out-degree is the number of outward edges. Thus, a person having high out-degree will be able to send a lot of information to other actors in the network. Priebe et al. [PCMP05] defined the locality statistic,$\Psi_{k,t}(v)$, the standardized locality statistic,$\tilde{\Psi}_{k,t}(v)$, and the scan statistics, $M_{k,t}$, as:

$$\Psi_{k,t}(v) = |E(\Omega(N_k[v, D_t]))|;\ \ k = 0, 1, 2...; \tag{1.4}$$

$$\tilde{\Psi}_{k,t}(v) = \frac{\Psi_{k,t}(v) - \hat{\mu}_{k,t,w}(v)}{\max(1, \hat{\sigma}_{k,t,w}(v))}, \tag{1.5}$$

$$M_{k,t} = \max_{v \in V}\ \ \Psi_{k,t}(v);\ \ k = 0, 1, 2, ..., \tag{1.6}$$

$$\hat{\mu}_{k,t,w}(v) = \frac{1}{w} \sum_{j=t-w}^{t-1} \Psi_{k,j}(v), \tag{1.7}$$

$$\hat{\sigma}^2_{k,t,w}(v) = \frac{1}{w-1} \sum_{j=t-w}^{t-1} \left(\Psi_{k,j}(v) - \hat{\mu}_{k,t,w}(v)\right)^2, \tag{1.8}$$

where $\hat{\mu}_{k,t,w}(v)$ and $\hat{\sigma}^2_{k,t,w}(v)$ are the mean and the variance, respectively, of the local statistic in the window w. The standardized scan statistic at the time t is written as:

$$\tilde{M}_{k,t} = \max_{v \in V} \ \tilde{\Psi}_{k,t}(v); \ \ k = 0, 1, 2, ... \tag{1.9}$$

The standardized scan statistic, $\tilde{M}_{k,t}$, was further normalized to temporally normalized scan statistic, $S_{k,t}$, which is defined as:

$$S_{k,t} = \frac{(\tilde{M}_{k,t} - \tilde{\mu}_{k,t,l})}{\max(1, \tilde{\sigma}_{k,t,l})}, \tag{1.10}$$

where $\tilde{\mu}_{k,t,l}$ and $\tilde{\sigma}_{k,t,l}$ are the estimated mean and standard deviation, respectively, of $\tilde{M}_{k,t}$ corresponding to the lag or time step l. The $\tilde{M}_{k,t}$ have been estimated for $k = 0, 1, 2$ and $t = 1, 2, ..., 189$ for the Enron data. For $k = 2$, $t = 132$ and $l = 20$, the $\tilde{M}_{k,t}$ is greater than 5 times standard deviations above its mean, indicating a clear anomaly. The corresponding temporally-normalized scan statistic, $S_{(2,132)}$, is 7.3. Assuming normality, the observed p-value is $< 10^{-10}$. They also used the extreme value theory, Gumbel model, to estimate the exceedance probability (p-value), which turned out to be $< 10^{-10}$. Therefore, they could infer using scan statistics that there exist anomalies in the Enron network data.

However, to implement scan statistic, a vertex-dependent local stationarity assumption was required. The previous model is an oversimplified statistical inference model, and a short-time near stationary (lag = 20 weeks) for null model was assumed. They did not consider detrending and seasonal adjustment methods on the univariate time series to

make the time series stationary. If x_t is a stationary time series, then the distribution of $(x_t, ..., x_{t+s})$ does not depend on t for all s. As the data are often non-stationary/non-random, they can give rise to trends and non-constant variance over time. It is also very common for dynamic email data to have seasonal effect, which can mask the non-seasonal characteristics of the data. In order to better reveal the features of the data that are of interest, removing trends, non-constant variance, and seasonal effects from time series data are necessary.

To obtain temporal scan statistic, $S_{k,t}$, Priebe et al. [PCMP05] normalized the locality statistic, $\Psi_{k,t}(v)$, twice and obtained the p-value, assuming normality of the scan statistic. In this model, it was assumed that the subgraphs are disjoint. However, the anomaly may split among multiple windows, and the set of subgraphs may not be disjoint, suggesting that the temporally-normalized scan statistic, $S_{k,t}$, is not independent.

They used the outdegree as a locality statistic. The degree, however, is not an effective structural location of a node in a network. The degree of a node in a graph is the number of direct connections that a node has with other nodes. If a node has high degree, the individual will be simply a connector or a hub and will not play a vital role in the social network. As a result, this metric is not very effective in detecting anomaly in a social network. On the other hand, the betweenness of a node in global network measures its influential position (broker, leader or bridge) in the network [PS00]. Thus, the betweenness centrality applied to neighborhood network can be very useful in identifying locally important individuals. Another measure for locality statistics is density applied to neighborhood network, which reveals how tightly-coupled the neighborhood is [PS00]. This will also be effective in detecting anomaly of a network.

1.2 Present Research

The objective of this research is to develop a fundamental understanding of methodologies to discover network patterns, to detect anomalies of dynamic networks obtained from an

organizational email, and to make statistical inferences by implementing statistical models and data mining algorithms. The current research employs temporal scan statistics to detect clusters, where the maximum log likelihood ratio is the test statistic [Kul79], and use the betweenness as the locality statistic. Here the betweenness follows a binomial distribution as it is related to the ratio of number of geodesic paths to the total number of geodesic paths. In addition, this research develops a purely temporal scan statistics for email count data based on the Poisson model.

The alternative approaches, such as the autoregressive moving average (ARMA) process and the vector autoregressive (VAR) process to detect clusters in a point process are implemented for the univariate time series of neighborhood ego subgraphs and the multiple time series of neighborhood ego subgraphs, respectively. In addition, this research employs the latent Dirichtlet allocation modeling on the e-mail content to model the topics associated with the dynamic textual data, and to study any significant topic change associated with the time series of the maximum topic proportion using scan statistics.

One of the scientific challenges of this research includes understanding the distribution of the organizational email subnetworks. As in one dimension, the exact distribution of the scan statistic under the null hypothesis is only available for special cases [PGKS05], this research employs other methodologies, such as Monte Carlo (MC) simulations and the extreme value theory to estimate p-values. Once the sampling distribution of scan statistic is determined, the inference on anomaly can be performed. There are three equivalent ways of performing a hypothesis test, such as the p-value approach, the classical approach, and the confidence interval approach. The extreme value theory is a statistical model that is used to model the extreme data in a given period of time, and is based on the location-scale family. Gumbel distribution is the most well-known distribution that belongs to this family, and has been widely used in engineering. The present work also applies the Gumbel distribution to approximate the p-value.

Another challenge is the choice of local statistic as it provides important structural location of a node and its neighborhood. I have proposed local graph invariant, such as the

betweenness, as a measure to identify local structure in social networks. For building the univariate time series of graphs, the challenges are to compute the graph distance metrics, which are computationally intensive, and to fit time series model to assess anomalies based on residuals.

Technical tasks to meet objectives and scientific challenges of the research are:

1. Scan Statistics: The temporal scan statistics models on email count and network data over a time interval from an organizational email are developed using the maximum log-likelihood ratio as the scan statistic. After applying detrending, variance stabilization, and seasonal adjustment to the time series of email count, anomalies have been assessed. A local statistic for subnetworks, such as betweenness is used. The extreme value theory and the Monte Carlo simulation methods are employed to make inferences, as the sampling distribution of the scan statistics is not known for most of the cases. Also, the Monto Carlo simulations for testing one-change point in mean in time series of count data is conducted.

2. Time Series: A univariate time series has been developed using the graph edit distance (GED) between subgraphs. An ARMA model is fitted to the time series, and the anomalies have been assessed using a residual threshold obtained from the ARMA model fitted to the GED series. In addition, a VAR model for multivariate time series of neighborhood ego subnetworks for each vertex using the GED has been developed to identify anomalies and detect chatter.

3. Content analysis: A vector space model is implemented to construct the term document matrix (TDM) obtained from the corpus extracted from unstructured email content at every week. The probabilistic model, latent Dirichlet allocation (LDA) process is then applied to the document term matrix (DTM) in order to estimate the topic proportions to build a univariate time series. Subsequently, anomalies are assessed using scan statistic model and residual analysis of the fitted time series model. In addition, the multidimensional scaling (MDS) or the singular value decomposition

(SVD) is conducted to reduce dimensionality to compare the optimal number of topics obtained from the LDA. The K-means clustering is used to the reduced dimension obtained from the MDS for the corpus of every vertex at each week to cluster the documents. Further, the k-nearest neighbor method is implemented to classify messages at time, t, for pattern retrieval and to identify chatter.

The remainder of this chapter provides an overview on the distribution of the scan statistics. A brief discussion on the critical region, time series methods, graph properties are also presented. Finally, an overview of the prospective surveillance methods on the social network data is presented.

1.3 Distribution of Scan Statistic

1.3.1 Exact Distribution

In one dimension, the exact distribution of the scan statistic is only available for special cases. Naus [Nau65] has first presented the distribution of the maximum cluster of points on a line. Here N ordered points, $x_1 \leq x_2 \leq \ldots \leq x_N$, with respect to size are considered and independently drawn from the uniform distribution on (0, 1). The $P(k|N;w)$ is the probability that the largest number of points (events) within a sub interval of (0, 1) of length $w \geq k$. Let S_w be the largest number of points within a sun-interval of [0,1) of length w. The right tail probability of the scan statistic, S_w, and the $P(k|N;w)$ are expressed as:

$$p_{value} = P(S_w \geq k|H_0) = P(k|N;w). \tag{1.11}$$

Naus [Nau65] has derived the formulas of $P(k|N;w)$ as:

$$P(k|N;w) = \begin{cases} C(k|N;w) - R(k|N;w), & \text{for } w \geq \frac{1}{2}, k > \frac{(N+1)}{2}, \\ C(k|N;w), & \text{for } w \leq \frac{1}{2}, k > \frac{N}{2}. \end{cases}$$

Here $C(k|N;w)$ is the sum of cumulative binomial probabilities and defined as:

$$C(k|N;w) = (N-k+1)[G_b(k-1|N;w) + G_b(k+1|N;w)] - 2(N-k)G_b(k|N;w), \quad (1.12)$$

and G_b is the cumulative binomial probability such that:

$$G_b(k|N;w) = \sum_{i=k}^{N} b(i|N;w), \quad (1.13)$$

$$b(k|N;w) = \binom{N}{k} w^k (1-w)^{N-k}. \quad (1.14)$$

$R(k|N;w)$ is the sum of the product of binomials and cumulative binomial probabilities defined as:

$$R(k|N;w) = \sum_{i=k}^{N} b(y|N;w)F(N-k|y;v/w) + H(k|N;w)b(k|N;w), \quad (1.15)$$

where

$$F_b(k|N;w) = \sum_{i=0}^{k} b(i|N;w), \quad (1.16)$$

$$H(k|N;w) = \frac{nv}{w} F_b(N-k|k-1;v/w) - (N-k+1)F_b(N-k+1|k;v/w), \quad (1.17)$$

and $v = 1-w$. The exact distribution of S_w has been formulated by Wallenstein and Naus [WN74] and Huntington and Naus [HN75]. Naus [Nau82] has given the following approximation for a Poisson process with mean rate λ per unit time over the interval [0,T). For $\mu = \lambda w$ and $L = \frac{T}{w}$:

$$p_{value} = P_{H0}(S_w \geq k|\mu, L) \approx 1 - Q_2 \left(\frac{Q_3}{Q_2}\right)^{L-2}. \quad (1.18)$$

Q_2 and Q_3 are defined as:

$$Q_2 = (F(k-1,\mu))^2 - (k-1)p(k;\mu)p(k-2;\mu) - (k-1-\mu)p(k;\mu)F(k-3;\mu),$$

$$Q_3 = (F(k-1,\mu))^3 - E_1 + E_2 + E_3 - E_4. \tag{1.19}$$

E_1, E_2, E_3 and E_4 are given by the following equations.

$$E_1 = 2p(k;\mu)F(k-1;\mu)[(k-1)F(k-2;\mu) - \mu F(k-3;\mu)],$$

$$E_2 = 0.5(p(k;\mu))^2[(k-1)(k-2)F(k-3;\mu) - 2(k-2)\mu F(k-4;\mu) + \mu^2 F(k-5;\mu)],$$

$$E_3 = \sum_{i=1}^{k-1} p(2k-i;\mu)(F(i-1;\mu))^2,$$

$$E_4 = \sum_{i=2}^{k-1} p(2k-i;\mu)p(i;\mu)[(i-1)F(i-2;\mu) - \mu F(i-3;\mu)], \tag{1.20}$$

where $p(k;\mu) = \frac{e^{-\mu}\mu^k}{k!}$ and $F(k;\mu) = \sum_{j=0}^{k} p(j;\mu)$.

1.3.2 Order Statistics and Extreme Value Theory

Much of the literature focuses on approximations to the p-value based on the extreme value theory [PCMP05]. The extreme value theory is mainly associated with the maximum or minimum of a sequence of random variables $X_1, X_2, ..., X_n$, and it would be appropriate to discuss order statistics in order to estimate the distribution of the maximum or minimum. Let $X_1, X_2, ..., X_n$ be an independent and identically distributed (iid) random variables with distribution function $F_X(x)$ and density function $f_X(x)$. Given random variables $X_1, X_2, ..., X_n$, the order statistics are $X_{(1)}, X_{(2)}, ..., X_{(n)}$, such that $X_{(1)} < X_{(2)} < ... < X_{(n)}$, where $X_{(1)}$ is called the smallest or the first order statistics and is defined as $X_{(1)} = min\{X_1, ..., X_n\}$. The largest or the nth order statistics is defined as $X_{(n)} = max\{X_1, ..., X_n\}$. Given the continuous iid random variables, $X_1, X_2, ..., X_n$, the

cumulative distribution function of the sample maximum, $X_{(n)}$, is given by [Sut12]:

$$F_{X_{(n)}}(x) = P(X_{(n)} \leq x) = P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x)$$

$$= P(X_1 \leq x)P(X_2 \leq x)...P(X_n \leq x) = [F_X(x)]^n. \tag{1.21}$$

The probability density function of the sample maximum, $X_{(n)}$, can be written as:

$$f_{X_{(n)}}(x) = \frac{d}{dx}[F_X(x)]^n \rightarrow f_{X(n)}(x) = n[F_X(x)]^{n-1} f_X(x). \tag{1.22}$$

Similarly, given the continuous iid random variables, $X_1, X_2..., X_n$, the cumulative distribution function of the sample minimum, $X_{(1)}$, is given by:

$$F_{X(1)}(x) = P(X_{(1)} \leq x) = 1 - P(X_{(1)} > x)$$

$$= 1 - P(X_1 > x, X_2 > x, ..., X_n > x) = 1 - [1 - F_X(x)]^n. \tag{1.23}$$

The probability density function of the sample minimum, $X_{(1)}$, is given by:

$$f_{X_{(1)}} = \frac{d}{dx}[1 - [1 - F_X(x)]^n] = n[1 - F_X(x)]^{n-1} f_X(x). \tag{1.24}$$

The cumulative distribution and the probability density function of $\tilde{M}_{k,t}$ can be derived from this formula. However, the cumulative distribution of X is not known. The alternative approach would be to consider an approximate location-scale family for $[F_X(x)]^n$ [GPW09]. There are only three types of distribution that belong to this location-scale family. They are Frechet, Weibull and Gumbel. The cumulative distribution function of the Frechet

distribution for $x \in R$ is:

$$F(x) = \begin{cases} 0, & x \leq 0, \\ exp(-x^{-a}), & x > 0, \end{cases}$$

where $a > 0$ is the shape parameter. The cumulative distribution function of the Weibull distribution is given by:

$$F(x) = \begin{cases} 1 - exp(-(-x^{-a})), & x \geq 0, \\ 0, & x < 0, \end{cases}$$

where $a > 0$ is the shape parameter. The cumulative distribution function of the Gumbel distribution is:

$$F(x) = 1 - exp(-e^{\frac{x-\alpha}{\beta}}), \qquad x \geq 0, \tag{1.25}$$

where α and β are the location and scale parameter, respectively.

1.4 The Critical Region

Let the data set, X, is partitioned into n disjoint nonempty intervals termed as windows $w_i; i = 1, 2, ..., n$ such that $X_i \in w_i; i = 1, 2, ..., n$, and $X = \cup_{i=1}^{n} X_i \subset \cup_{i=1}^{n} w_i$. A set of local statistics, $M_{w1}(X_1), M_{w2}(X_2), ...M_{wn}(X_n)$ is estimated from data $X_i \in w_i; i = 1, 2, ..., n$. The scan statistic, $M_w(x)$, is defined as the maximum of the set of the locality statistics [Mar12]. The null hypothesis is that there is homogeneity across the subregions, and the alternative hypothesis is that there is inhomogeneity i.e. the existence of local subregions or windows of excessive activity. The null hypothesis (H_0) and the alternative hypothesis

(H_a) can be written as:

$$\left.\begin{aligned} H_0 &: E(X_1) = E(X_2) = ... = E(X_n) = \mu, \\ H_a &: E(X_1) = E(X_2) = ... = E(X_k) \neq E(X_{k+1}) = ... = E(X_n), \end{aligned}\right\} \tag{1.26}$$

where $1 \leq k \leq n$ is unknown. Once the observed scan statistic has been estimated, the inference on anomaly can be performed. There are three equivalent ways to perform a hypothesis test which are p-value approach, the classical approach and the confidence interval approach. For testing hypothesis, H_0 is rejected in favor of H_a if p-value $\leq \alpha$, where the size of the test, α, is the probability of rejecting H_0 when H_0 is true. Mathematically, it can be written as:

$$P_{H_0}[M_w(X) \geq C_\alpha] = \alpha, \tag{1.27}$$

where C_α is the critical value. Here the p-value has been estimated from the observed sample. In the classical approach, one can reject H_0 in favor of H_a if the observed value of $M_w(X)$ is greater than C_α. For the two-tailed test, one can use the confidence interval approach. Therefore, if the sampling distribution of the scan statistic under the null hypothesis, H_0 is known, the p-values, C_α and the test size, α can be estimated.

1.5 Time Series

A time series is a collection of random variables, $\{X_t\}$, indexed by discrete time t. Let the observed values of the random variable, X_t be $x_1, x_2, x_3, ..., x_n$. Here x_1 is the observed value of the series at the first time point, and x_2 is the observed value at the second time point and so on. The interesting feature of time series is that the current value, x_t depends on previous values $x_{t-1}, x_{t-2}, ...$ Therefore, the adjacent time points are correlated. The other features of time series are trends, seasonal variations and noise. The joint cumulative

distribution function of the stochastic process $\{X_t\}$ [SS06] is:

$$F(x_1, x_2..., x_n) = P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n). \tag{1.28}$$

If the random variables are iid standard normal, then the joint probability distribution function of the stochastic process, $\{X_t\}$, is:

$$f(x_1, x_2, ..., x_n) = \prod_{t=1}^{n} \Phi(x_t). \tag{1.29}$$

If $\{X_t\}$ are iid standard normal variables, then

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} exp(-\frac{z^2}{2})dz. \tag{1.30}$$

The one dimensional cumulative distribution function is:

$$F_t(x) = P(X_t \leq x). \tag{1.31}$$

The one dimensional density function is:

$$f_t(x) = \frac{\partial F_t(x)}{\partial x}. \tag{1.32}$$

The mean of the stochastic process, $\{X_t\}$, is defined as:

$$\mu_t = EX_t = \int_{-\infty}^{\infty} x f_t(x) dx. \tag{1.33}$$

The autocovariance of the stochastic process, $\{X_t\}$, is defined as[SS06]:

$$\gamma(s,t) = cov(X_t, X_s) = E(X_t - \mu_t)(X_s - \mu_s), \quad for \ \ all \ \ s \ \ \& \ \ t. \tag{1.34}$$

The auto correlation function (ACF) is defined as [SS06]:

$$\rho(t,s) = \frac{\gamma(t,s)}{\sqrt{\gamma(t,t)\gamma(s,s)}}, \tag{1.35}$$

where $-1 \leq \rho(t,s) \leq 1$. The assumption in behavior of time series is that the data are stationary. The time series is strictly stationary if the cumulative distribution function of the stochastic process, $\{X_t\}$, remains unchanged under a shift in time. In other words, one can write [SS06]:

$$P(X_1 \leq x_1, X_2 \leq x_2, ..., X_n \leq x_n) = P(X_1 + h \leq x_1, ..., X_n + h \leq x_n), \tag{1.36}$$

where $h = 0, \pm 1, \pm 2, ...$ is the time shift. Weak stationarity occurs if the mean does not change over time, and the autocovariance depends on separation between time, t and s. In other words, $E(X_t) = \mu_t = \mu$. Letting $s = t + h$, one obtains [SS06]:

$$\gamma(s,t) = \gamma(t+h,t) = E(X_{t+h} - \mu)(X_t - \mu) = E(X_h - \mu)(X_0 - \mu) = \gamma(h,0). \tag{1.37}$$

The stationarity is assessed visually by investigating the sample autocorrelation function and the sample partial autocorrelation function (PACF). Let X_t, $t \in Z$ be a stationary time series. The autocovariance function and the autocorrelation function (ACF) of X_t can be written as [SS06]:

$$\gamma(h) = \mathrm{Cov}(X_{t+h}, X_t) = \mathrm{E}[(X_{t+h} - \mu)(X_t - \mu)] \quad for \quad t, h \in Z, \tag{1.38}$$

$$\rho(h) = \frac{\gamma(t+h,t)}{\sqrt{(\gamma(t+h,t+h)\gamma(t,t))}} = \frac{\gamma(h)}{\gamma(0)}. \tag{1.39}$$

Therefore, the properties of a stationary time series do not depend on the time at which the series is observed. In fact, the trend and seasonality affect the value of the time series at different times, suggesting the time series with trends or, and seasonality is not stationary.

Given a probability space (Ω, F, P), where Ω is a sample space consisting of possible outcomes of an experiment, F is a σ-algebra that is a collection of subsets of Ω satisfying the following three properties [Str99], [Gra17]:

1. $\Omega \in F$ and $\Omega^c = \Phi => \Phi \in F$,

2. F is closed under complementation i.e. if $A \in F$ then $A^c \in F$,

3. F is closed under countable union, i.e. if $A_1, A_2, ... \in F$ then $\bigcup_{i=1}^{\infty} A_i \in F$.

By De Morgans law,

$$\bigcap_{i=1}^{\infty} A_i = (\bigcup_{i=1}^{\infty} A_i^c)^c, \tag{1.40}$$

which implies F is closed under countable intersection, i.e. if $A_1, A_2, ... \in F$ then $\bigcap_{i=1}^{\infty} A_i \in F$. A probability measure P on F is a function P : $F \to$ [0, 1] satisfying

1. P(0) = 0,

2. P(Ω) = 1,

3. P(A) = 1 - P(A^c),

4. For any sequence A_n of pairwise disjoint sets, P $(\bigcup_{n=1}^{\infty} A_n) = \sum_{i=1}^{\infty} P(A_n)$.

A random variable X defined on (Ω, F, P) is a function X: $\Omega \to$ R satisfying $\{\omega \in \Omega : X(\omega) \leq x\} \in F$ for all $x \in R$ [Gra17]. Let X be a random variable that is indexed to time t. The observations $\{ x_t, t \in T\}$ is defined as a time series, where T is an integer set Z. The properties of the time series $\{X_t, t \in Z\}$ include that it has a finite dimensional joint distribution and it has moments i.e. means, variances, and covariances.

The time series $\{X_t, t \in Z\}$ is said to be strictly stationary if $\mathrm{P}(X_{t1} \leq c_1, ..., X_{tk} \leq c_k) = P(X_{t1+h} \leq c_1, ..., X_{tk+h} \leq c_k)$ i.e the finite dimensional joint distributions are time invariant, and the properties of the weakly stationary are:

1. $\mathrm{E}(X_t) = \mu$, for all $t \in \mathrm{Z}$,

2. $\mathrm{Var}(X_t) = \sigma^2$ for all $t \in \mathrm{Z}$,

3. $\mathrm{Cov}(X_t, X_{(t-s)}) = \gamma_s$, for all $s, t \in \mathrm{Z}$.

1.6 Graph Features: Degree, Betweenness and Density

For a graph $G = (V, E)$ with n vertices, the degree, k_v, of a vertex, v, is the number of edges associated with it. Mathematically, if G is an undirected graph, then the degree can be written as [New10]:

$$k_v = \sum_{j=1}^{n} A_{vj}. \tag{1.41}$$

The betweenness of a vertex, v, [New10] is the number of geodesic (shortest) paths that pass through the vertex, v.

$$n_{ij}^{v} = \begin{cases} 1, & \text{if vertex } v \text{ is on the geodesic path from } i \textit{ to } j, \\ 0, & \text{otherwise.} \end{cases}$$

Then, mathematically the betweenness, b_v, is defined as:

$$b_v = \sum_{ij} n_{ij}^{v}. \tag{1.42}$$

Betweenness is then standardized as:

$$b_v = \sum_{ij} \frac{n_{ij}^v}{\sigma_{ij}}, \tag{1.43}$$

where σ_{ij} is the total number of geodesic paths from i and j. Although an individual may have little direct connections or degrees, betweenness of a node in global network measures its influential position (broker, leader or bridge) in the network. Here, betweenness centrality is applied to neighborhood network and can thus be very useful in identifying locally important individuals. The density of a network estimates community structure of a network and mathematically can be defined as [New10]:

$$\eta = \frac{m}{\binom{n}{2}}, \tag{1.44}$$

where m is the number of edges and $\binom{n}{2}$ is the number of possible edges. However, density is applied here to the neighborhood network, and therefore, can reveal how tightly-coupled the neighborhood is.

1.7 Different Surveillance Methods

The present research work deals with the retrospective surveillance using conditional scan statistics given that a fixed number of N events have occurred at random over a time period. The objective of the proposed work is to develop methodologies to investigate the excessive activities as well as identify vertices associated with these activities for inter-organization e-mail networks from 1996 to 2009. Note that scan statistics can be also used prospectively [GNW01] to monitor future data, where the number of events in the time period in not fixed. Here, the other prospective surveillance methodologies are briefly outlined.

For network data, the surveillance consists of two phases, phase-I and phase- II. In phase-I, the first step is to estimate the s-sampled statistics, S_t, which is the density of

the subgraph, G_t, $t = 1, 2, 3, ..., s$. The statistical process monitoring techniques are then introduced to identify if the future subnetworks are anomalous. The mean, μ, and the variance σ^2 of the statistics are then estimated. If S_t is approximately normally distributed, one can estimate a region of bounds, $R(\mu, \sigma^2)$. The bounds of this region are given by [JSTW18]:

$$R(\mu, \sigma^2) = \hat{\mu} \pm 3\hat{\sigma}. \tag{1.45}$$

The fluctuation within this bound is regarded as a typical event. In phase-II, the statistic, S_t, where $t > s$, for the new subgraph is calculated. If the estimated S_t for the new graph is outside the bound, the subgraph is considered anomalous.

Recently, an adaptive cumulative sum (CUSUM) based surveillance technique has been developed for detecting bioterrorism prospectively by monitoring time series of daily disease count [SKM10]. They used negative binomial regression to prospectively model the background count and then to forecast the future count. The anomaly is detected when the observed counts are greater than the threshold CUSUM score. For cyber security applications, a number of nonparametric change point detection anomaly methods have been developed [SKM10].

1.8 Chapter Summary and Layouts

In this chapter, the fundamentals of scan statistics, and the previous models for the detection of excessive activities have been presented. In addition, the research and the approach of the present research work have been discussed. Chapter 2 presents the models on network neighborhood analysis and the likelihood ratio estimation using parametric and non parametric approach to detect the excessive activities on email data. Chapter 3 demonstrates the use of univariate and multivariate time series models to detect the excessive activities and show that the important vertices/nodes/IDs associated with the excessive activities. Chapter 4 presents LDA model to obtain the proportion of topics from the content of e-mails, demonstrate the use of scan statistics and time series to obtain excessive activities

with respect to the topic. Chapter 5 provides the conclusions and future work. In addition, the network data constructed from the e-mail edgelists from June 2003 to June 2004, and a partial R code have been presented in appendix A and B, respectively.

Chapter 2: Neighborhood Analysis using Scan Statistics

2.1 Introduction

Around terabytes of unstructured electronic data are generated every day from twitter networks, scientific collaborations, organizational emails, telephone calls and websites. Fraudulent activities owing to excessive communication in communication networks continue to be a major problem in different organizations. In fact, retrieving information relating to detection of excessive activities is computationally intensive for large data sets. Therefore, one needs useful tools and techniques to analyze such a massive data set and detect anomaly. In a social network, anomalies can occur as a result of abrupt changes in the interactions among a group of individuals [SZY+14]. Analyzing the excessive activity or abnormal change [SKR99] in dynamic social networks is thus important to understand the fraudulent behavior of individuals in a subregion of a network. Recently, a network neighborhood mapping has been applied to investigate the excessive activities, particularly the changes in local leadership in the terrorist network [Kre02]. The motivation of this research work is to investigate the excessive activities and make inferences in dynamic subnetworks.

The most commonly used methods for anomaly detection in data mining are density-based techniques, such as k-nearest neighbor [Adl84] and local outlier factor [BKNS00], one class support vector machines [SPST+01], neural networks [HHWB00], cluster analysis-based outlier detection [HXD03] and ensemble techniques [LK05]. However, all these methods used to detect excessive activity, are mostly descriptive in nature, and not effective in making statistical inferences. On the other hand, scan statistics have been used to identify and test if the cluster of events is statistically significant. The one-dimensional scan statistic with a fixed scan window is first developed by Naus [Nau65] and later has been used by a number of authors to investigate the unusual clusters of events in various fields, for

example, in visual perception [Gla79], queueing theory [Gla81], telecommunication [Alm83], epidemiology [WN87], and molecular biology [KB92]. It is then extended to two or higher dimensional scan statistics with variable window size and shape by Kulldorff [Kul79].

Although considerable work has been done to detect clusters of events or anomaly using scan statistics in spatial statistics and image analysis, relatively less attention has been given to detect anomaly in social networks. Priebe et al. [PCMP05] first applied temporal scan statistics for Enron email data to detect anomaly in time series of graphs. The full network was partitioned into disjoint subregions or subnetworks over time to overcome the computational complexity associated with a global network and to uncover the interesting features of a node and neighborhood. However, this model normalized the locality statistic twice to eliminate the trend and assumed short-time, near-stationarity for the null model. They did not consider differencing, seasonal adjustment in the univariate time series of scan statistics to make the time series stationary. Furthermore, they assumed that the subgraphs are disjoint. However, the scan statistics with fixed and disjoint scan window may not be appropriate because of the occurrence of window overlaps, which may result in loss of some data on the time axis.

In this research work, scan statistics with overlapping and variable window sizes to detect anomalies of dynamic networks obtained from organizational emails has been implemented, and the log likelihood ratio (LLR) has been employed as the test statistic to rank the clusters, as the cluster size is not known. Furthermore, the structural stability has been assessed, and the differencing and seasonal adjustment have been applied to make the time series of scan statistics stationary and estimate the p-value using Monte Carlo (MC) simulation and the extreme value distribution, such as Gumbel, as the exact sampling distribution of scan statistics under the null hypothesis is not known for most of the cases. In addition, as the unstructured data set size becomes larger, the formation of dynamic network structure is computationally intensive. Instead of applying temporal scan statistics on ego sub networks directly, here the tasks have been divided into two regimes. A primary cluster using the Poisson process model has been determined first, and then the neighborhood ego

subnetworks around the primary cluster have been extracted to investigate the excessive activities, utilizing the betweenness as a locality statistic and LLR as a scan statistic. Furthermore, as an alternative approach, a univariate time series has been built using the graph edit distance (GED) between subgraphs. An autoregressive moving average (ARMA) process is then fitted to the time series, and the anomalies are assessed using residuals from the fitted model. Statistical analyses are performed using R and SaTScan software (http:www.satscan.org). The weekly network data from June 2003 to June 2004 obtained from the e-mail edgelists over 52 weeks are given as a supplement.

Here I briefly outline the process that has been used throughout this chapter. Firstly, a change point of mean using non-parametric statistical model has been detected in raw e-mail count series. In order to obtain clusters of communications in the e-mail network, a new approach, which is the two-step scan process, has been developed. In step-I of the two-step scan process, I apply scan statistics using the Poisson model in the stationary count series to get an excessive primary cluster of communication. In step-II, I extract neighborhood networks around the primary cluster obtained from the step-I, and apply scan statistics based on the binomial process to obtain more refined excessive communications and influential nodes associated with the excessive communications.

2.2 Detection of Change Point in Time Series of Raw E-Mail Count

Figs. 2.1(a,b) show the observed number of emails per month from March 1996 to November 2009, and the sample autocorrelation function (ACF), respectively. The sample ACF plot of the email count series displays a slow decay, suggesting the email count series contains a trend and the peak at 12 months, indicating the series has seasonal variation [CM09]. Fig. 2.2 shows that the count data is also highly serially correlated. All these suggest the email count series is nonstationary. To apply one dimensional scan statistics, one needs to have random or stationary time series, which is discussed below.

Here the structural stability of the univariate time series of the logarithm of email counts has been investigated using nonparametric Kolmogorov-Smirnov (KS) test statistic. Considerable work has been done to detect the structural stability using nonparametric KS test statistic [BD93]. Recently, the KS test statistics have been extended [SZ10] to estimate the size and power properties of the KS test statistics with different bandwidths. Fig. 2.3 shows that the observed number of logarithm of emails counts per month from March 1996 to November 2009 consists of two phases. The mean of the first phase between $[1996-2003)$ is much smaller than that of the second phase between $[2003-2009)$, indicating the existence of change point in mean in the email count series. Mathematically, it can be expressed for ARMA (1,1) model as:

$$X_t = \begin{cases} v_t, & 1 \leq t \leq n/2, \\ v_t + \Delta\mu, & n/2 < t \leq n, \end{cases}$$

where $v_t = \rho v_{t-1} + \theta w_{t-1} + w_t$, v_t is stationary, ρ and θ are constants ($\rho \neq 0$, $\theta \neq 0$), $\Delta\mu$ is the magnitude of change and $w_t \sim iid\ N(0,1)$. The null hypothesis is that there is no change point of the mean in the time series, and the alternative hypothesis considers that there is a change point of the mean. The goal here is first in distinguishing the change point in mean in the original email count series. The KS test statistic can be written as [BD93];

$$KS = \sup|\frac{T(k)}{\sigma}|, \quad k = 1, ..., n, \tag{2.1}$$

where $T(k) = n^{-0.5}\sum_{t=1}^{k}(X_t - \bar{X})$, $\bar{X} = n^{-1}\sum_{t=1}^{n} X_t$, $\sigma^2 = \sum_{k=-b}^{b}\gamma(k)K(k/b)$ and $\gamma(k) = n^{-1}\sum_{j=1}^{n-k}(X_j - \bar{X})(X_{j+k} - \bar{X})$, γ is the sample autocovariance estimate at lag k, $K(.)$ is the Kernel function, b is a bandwidth parameter and σ^2 is the variance. To estimate σ here, we use three types of bandwidths, fixed bandwidth (FBW), data dependent bandwidth-I (DDBW-I) and data dependent bandwidth-II (DDBW-II). The Bartlett Kernel, $K(x)$, can

be written as:

$$K(x) = \begin{cases} 1 - |x|, & |x| \leq 1, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$b = \begin{cases} \lfloor n^{\frac{1}{3}} \rfloor, & \text{for FBW}, \\ \lfloor 1.1447(\frac{4\rho_1^2 n}{(1-\rho_1^2)^2})^{\frac{1}{3}} \rfloor, & \text{for DDBW-I}, \\ \lfloor 1.1447(\frac{4\rho_2^2 n}{(1-\rho_2^2)^2})^{\frac{1}{3}} \rfloor, & \text{for DDBW-II}, \end{cases}$$

where $\rho_1 = \frac{\sum_{t=2}^{n} u_t u_{(t-1)}}{\sum_{t=2}^{n} u_{(t-1)}^2}$, $u_t = X_t - \bar{X}$, $u_{t-1} = X_{t-1} - \bar{X}$, $\rho_2 = \frac{\sum_{t=2}^{n} v_t v_{(t-1)}}{\sum_{t=2}^{n} v_{(t-1)}^2}$, $v_t = X_t - k^{-1}\sum_{t=1}^{k} X_t$; if $t = 1, ..., k$, and $v_t = X_t(n-k)^{-1}\sum_{t=k+1}^{n} X_t$; if $t = k+1, ..., n$. The observed values estimated from the data and the critical values of the KS test statistics using MC are given in Table 2.1. In addition, the asymptotic null distribution of KS is $sup_{t\in[0,1]}|B(t) - tB(1)|$, where $B(t)$ is the Brownian motion. The KS critical value, obtained from the literature, is 1.36 at 0.05 significance level (SL) which is consistent with the critical value obtained from the MC simulation. As the observed values exceed the estimated critical values (see Table 2.1), it can be concluded that the time series has a significant shift in mean implying that the statistically significant structural instability has been observed in the count series(see Fig. 2.3). It can also suggest that significantly different activity changes the mean of the series.

2.3 Discrete Scan Statistics Using Poisson Process

The discrete scan statistics have been employed here on organization emails collected between 1996 and 2009. Two kinds of data sets have been generated from the metadata of the organization emails, email count data (see Fig. 2.1) and the network data, given in the appendix.

Table 2.1: Kolmogorov Smirnov Test Statistics for different bandwidth parameters, Observed values and critical values.

Test Statistics	Observed Values	MC Critical Values (SL)
KS_{FB}	2.10971	1.28 (0.05)
KS_{DDB1}	1.359413	1.238 (0.05)
KS_{DDB2}	2.845591	1.32 (0.05)

Mathematically, the scan statistic is defined as the extremum of local statistics, which can be estimated from the scanned region of the data. If X_1, X_2, ..., X_n be i.i.d. non-negative integer valued random variables, then the one-dimensional discrete scan statistic, S_w, is written as [GB99]:

$$S_w = \max_{1 \leq t \leq n-w+1} Y_t(w), \tag{2.2}$$

where $Y_t = \sum_{i=t}^{t+w-1} X_i$ is the total number of events in the scanning window w. Below we implement scan statistics for two different models, Poisson and Binomial.

As the number of emails per month has been generated from the metadata of the organization emails, one can model it as a Poisson process. Under the null hypothesis, observations X_1, X_2, ..., X_n are from the i.i.d Poisson distribution with rate λ_0 . For the alternative hypothesis, there would be a scanning window of fixed width, w, where observations are from an i.i.d Poisson process with different rate λ_1, and the observations in the rest of the intervals, $[1, t)$ and $[t + w, n]$, are from an i.i.d Poisson process with rate λ_0 [Gen15]. For testing, the null hypothesis, $H_0 : \lambda_0 = \lambda_1$, over the alternative hypothesis, $H_1 : \lambda_1 > \lambda_0$, the likelihood under the null hypothesis $L_{H_0} = \prod_{i=1}^{n} \frac{e^{-\lambda_0} \lambda_0^{x_i}}{x_i!}$ and the likelihood under the alternate hypothesis $L_{H_1} = \left(\prod_{i=1}^{t-1} \frac{e^{-\lambda_0} \lambda_0^{x_i}}{x_i!}\right) \left(\prod_{i=t}^{t+w-1} \frac{e^{-\lambda_1} \lambda_1^{x_i}}{x_i!}\right) \left(\prod_{i=t+w}^{n} \frac{e^{-\lambda_0} \lambda_0^{x_i}}{x_i!}\right)$ [Gen15].

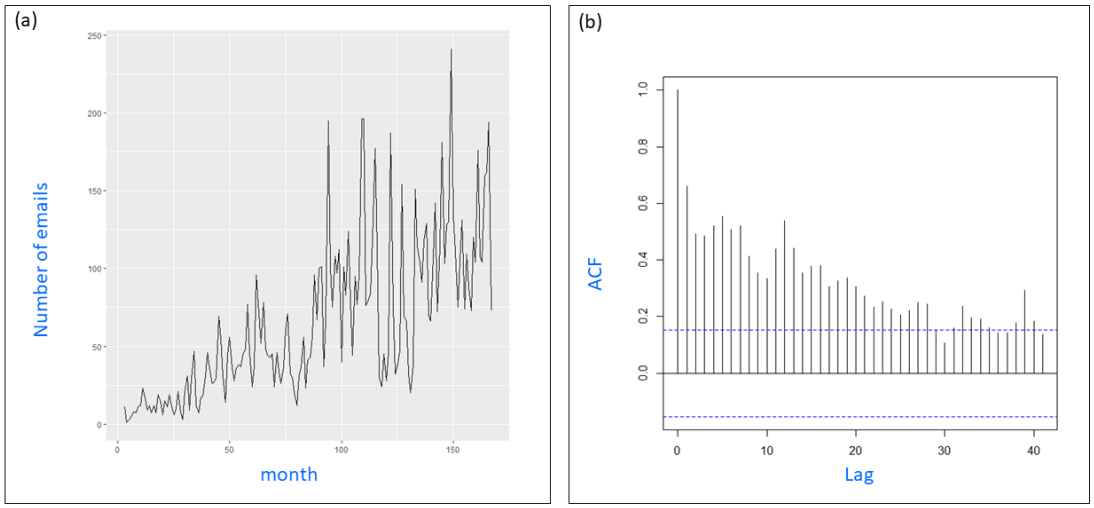


Figure 2.1: (a) Monthly number of e-mails received for the period 1996-2009. (b,c) Sample ACF and PACF of the monthly number of e-mails, respectively, showing that the time series is not stationary.

The likelihood ratio (LR), Λ, chosen as test statistic, is given by:

$$\Lambda = \frac{Likelihood\ \ under\ \ H_1}{Likelihood\ \ under\ \ H_0} = \frac{(\prod_{i=t}^{t+w-1} \frac{e^{-\lambda_1}\lambda_1^{x_i}}{x_i!})}{(\prod_{i=t}^{t+w-1} \frac{e^{-\lambda_0}\lambda_0^{x_i}}{x_i!})}. \tag{2.3}$$

Therefore, the log likelihood ratio can be written as:

$$\left.\begin{aligned} \log\Lambda &= \textstyle\sum_{i=t}^{t+w-1}[(-\lambda_1 + x_i\log(\lambda_1) - \log(x_i!)) - (-\lambda_0 + x_i\log(\lambda_0) - \log(x_i!))], \\ &= w(\lambda_0 - \lambda_1) + \log\left(\tfrac{\lambda_1}{\lambda_0}\right)\textstyle\sum_{i=t}^{t+w-1} x_i, \\ &= A_1 + A_2 N_t, \end{aligned}\right\} \tag{2.4}$$

where $A_1 = w(\lambda_0 - \lambda_1)$, $A_2 = \log(\frac{\lambda_1}{\lambda_0})$ and $N_t = \sum_{i=t}^{t+w-1} x_i =$ the number of observed emails

in the scanning window of size w for $1 \leq t \leq n - w + 1$. The one-dimensional discrete scan statistic, $S_w = \max_{1 \leq t \leq n-w+1} N_t$. Since $A_1 < 0$ and $A_2 > 0$, $\log \Lambda$ can be expressed as a monotonically increasing function of N_t for fixed w, and it can be written as for rejecting H_o over H_1:

$$\max\left(\log(\Lambda)\right) \approx S_w. \tag{2.5}$$

As the window size is not known a priori, we extend the fixed window formalism to different window length. For variable and overlapping windows, the LR and the LLR can be written as:

$$\Lambda = \frac{(\prod_{i=s}^{t} \frac{e^{-\lambda_1} \lambda_1^{x_i}}{x_i!})}{(\prod_{i=s}^{t} \frac{e^{-\lambda_0} \lambda_0^{x_i}}{x_i!})} = \prod_{i=s}^{t} e^{-(\lambda_1 - \lambda_0)} \left(\frac{\lambda_1}{\lambda_0}\right)^{x_i}, \tag{2.6}$$

$$\log(\Lambda) = \sum_{i=s}^{t} (\lambda_0 - \lambda_1) + \log(\frac{\lambda_1}{\lambda_0}) \sum_{i=s}^{t} x_i, \tag{2.7}$$

where $\lambda_0 = \frac{\sum_{i=1}^{t} x_i}{\sum_{i=1}^{t} X_i}$, $\lambda_1 = \frac{\sum_{i=s}^{t} x_i}{\sum_{i=s}^{t} X_i}$ and $\sum_{i=1}^{t} X_i =$ the total number of months in the whole study region. To perform a hypothesis test, the p-value approach is used in this work. H_0 is rejected in favor of H_1 if p-value $\leq \alpha$, where the size of the test, α, is the probability of rejecting H_0 when H_0 is true which can be written as $P_{H_0}[S_w \geq C_\alpha] = \alpha$, where C_α is the critical value. In the present work, the p-value has been estimated using the MC simulation [Kul79] or the extreme value distribution, Gumbel [PCMP05], to make an inference about the cluster of events.

2.4 Detection of Cluster of Communications Using Two-Step Scan Process

2.4.1 Removal of Trend and Seasonal Effects

As the time series data are often non-stationary/non-random, they can give rise to trends and non-constant variance over time. It is also very common for dynamic email data to

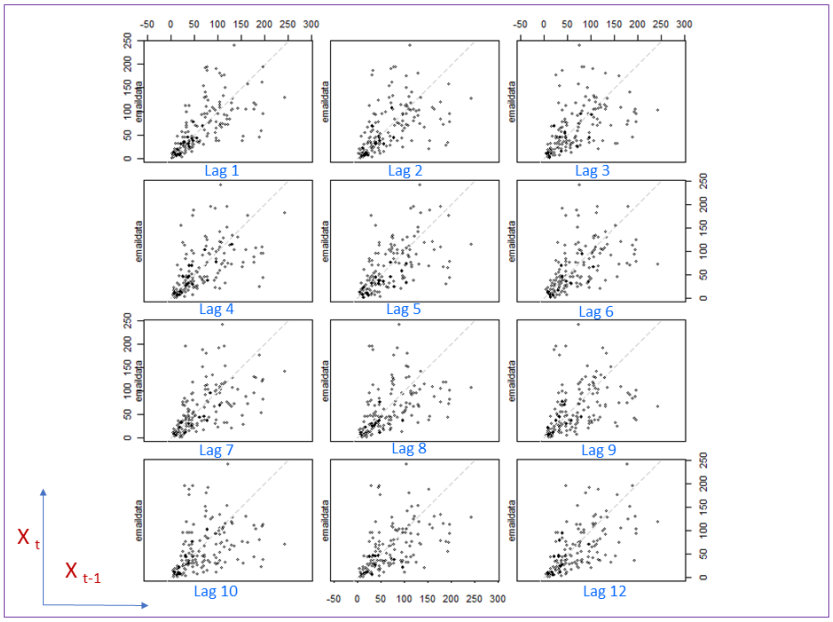


Figure 2.2: Scatter plot matrix showing the correlation of e-mail count with its own lagged values.

have seasonal effect, which can conceal the non-seasonal characteristics of the data. In order to better reveal the features of the data and apply scan statistics, removing trends and seasonal effects from time series data are necessary. To estimate the long-term trend, we fit kernel smoothing to the time series of email counts from March 1996 to November 2009, which shows a clear upward trend (See Fig. 2.4(a)).

The cycle plot which is shown in Fig. 2.4(b) is used to graph the seasonal component of the email count series [Cle93]. One could see that for months January, February, May, and July the cycle subseries appear to be increasing, whereas, for April, August, September, November and December the cycle subseries appear to be decreasing (see Fig. 2.4(b)). Therefore, the non-stationary behavior of the email count series can be due to trend and seasonality. To remove trend and seasonal effect, we will apply both the first difference and a seasonal difference to the monthly email count series where the time series is differenced

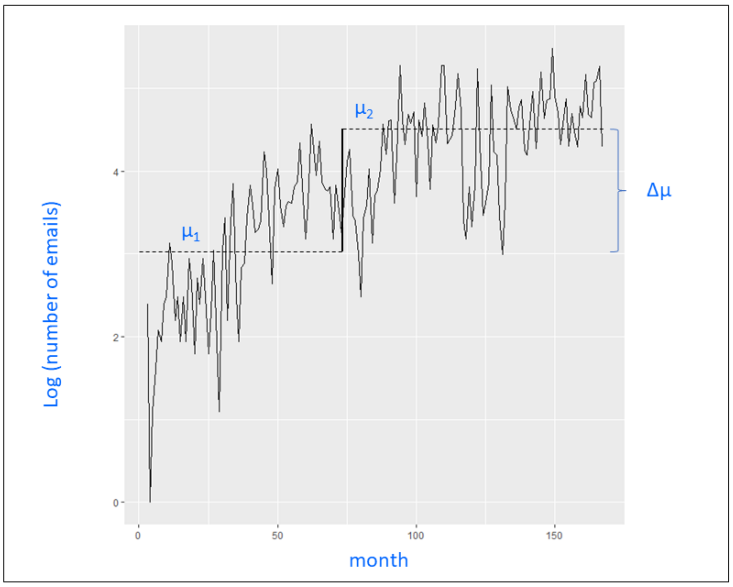


Figure 2.3: The time plot of the natural logarithm showing the change in mean.

based on its seasonal data points which is given by [SS06]:

$$\left.\begin{aligned}(1-B^{12})(1-B)x_t &= (1-B^{12}-B+B^{13})x_t \\ &= x_t - x_{t-12} - x_{t-1} + x_{t-13} \\ &= (x_t - x_{t-1}) - (x_{t-12} - x_{t-13}),\end{aligned}\right\} \tag{2.8}$$

where B is the back shift operator. Fig. 2.5(a) shows that the seasonally differenced series has neither the trend nor seasonal component that have been exhibited by the original time series of email count. The standard deviation (SD) of the original email count series turns out to be 50.31, while the SD for seasonally differenced email count series is 41.1, suggesting that the seasonal differencing is effective. The sample ACF of the seasonally differenced series has a very few smaller peak that are marginally statistically significant

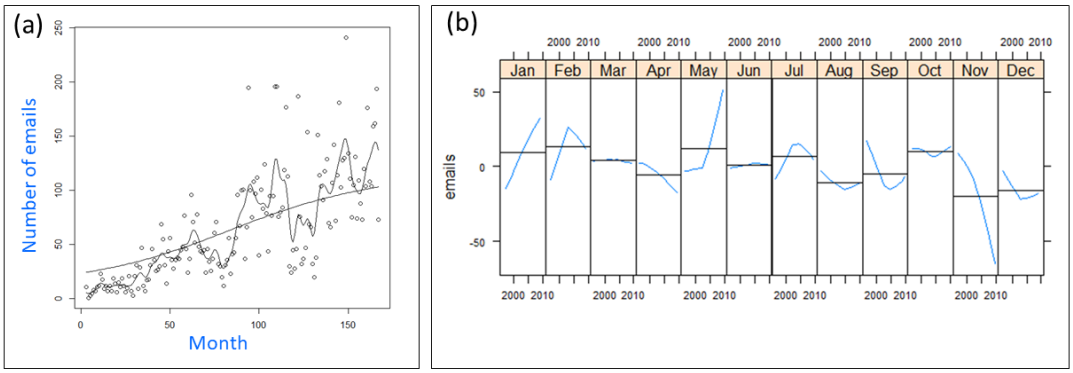


Figure 2.4: (a) Kernel smoothing of the raw e-mail count data showing an upward trend. (b) Plot showing the seasonal variations of the raw e-mail count data.

[see Fig. 2.5 (b)], so one can conclude the time series of seasonally differenced email count is nearly stationary. To confirm the stationarity, the Phillips-Perron (PP) unit root test (value of the test statistic = -8.48, lag parameter = 4, p-value = 0.01) has been performed. The Kwiatkowski-Phillips-Schmidt-Shin (KPSS) unit root test (value of the test statistic = 0.064, lag parameter = 2, p-value = 0.1) shows that the seasonally differenced series is trend-stationary. In addition, no significant mean shift of the seasonally differenced count series has been observed.

It was observed that there is a over-dispersion in the Poisson distribution of raw e-mail counts (Fig 2.1(a), and the raw count data is also highly serially correlated (see Fig. 2.2) or nonrandom, as observed by the auto correlation function (ACF) plot (see Fig. 2.1(b) in the text). The trend and the seasonal behavior of the raw e-mail count data affect the distribution, and makes it non-Poisson. The trend and the seasonal behavior in the raw data are shown in Fig. 2.4. The over-dispersion, however, has significantly reduced in the

seasonal difference count data (see Fig. 2.5(a) and the ACF plot, Fig. 2.5(b), shows the distribution becomes random, and the time series is stationary. Here the seasonal difference time series, which is stationary, has been used to estimate the maximum log likelihood ratio (LLR) using the Poisson model to obtain the primary cluster. Furthermore, it was observed that the left truncated seasonal difference data become normally distributed. It is known that that the Poisson distribution would tend to be normal, when the rate of arrival is high.

The counting processes that deal with over-dispersion have been covered in the literature by Weiss [Wei07], Weiss and Testik [WT09], Weiss [Wei09]. All those studies deal with the time series of correlated processes of the Poisson counts with over-dispersion. They used the modified ARMA models, such as integer valued ARMA (INARMA), INGARCH models to fit the over-dispersed and correlated count data to get the critical cluster. Another important part of our research, in addition to obtaining excessive activities, is to obtain the nodes/vertices associated with the excessive activities. To obtain this information, the time series of networks has been constructed using the maximum betweenness at a given time for different neighborhoods ($k = 1, 1.5, 2$ and > 2) as part of the two-step process around the primary cluster, and estimated the maximum LLR. In all cases the time series is stationary. These models (INARMA and INGARCH), referred above, identify the critical cluster only. However, these models will not provide the nodes or vertices associated with the critical cluster.

2.4.2 Step-I of Two-Step Scan Process: Estimation of LLR using Poisson Model

A purely temporal cluster analysis has now been applied to detect the temporal clusters of emails with high rates using the discrete Poisson model to the stationary time series of email count, as shown in Fig. 2.6 (upper panel), using SaTScan software [Kul79]. Here the scanning window is an interval (in time). At each location, the number of observed and expected observations inside the window has been compared. Different probability models, depending on the type of data, can be applied to evaluate the cluster using scan

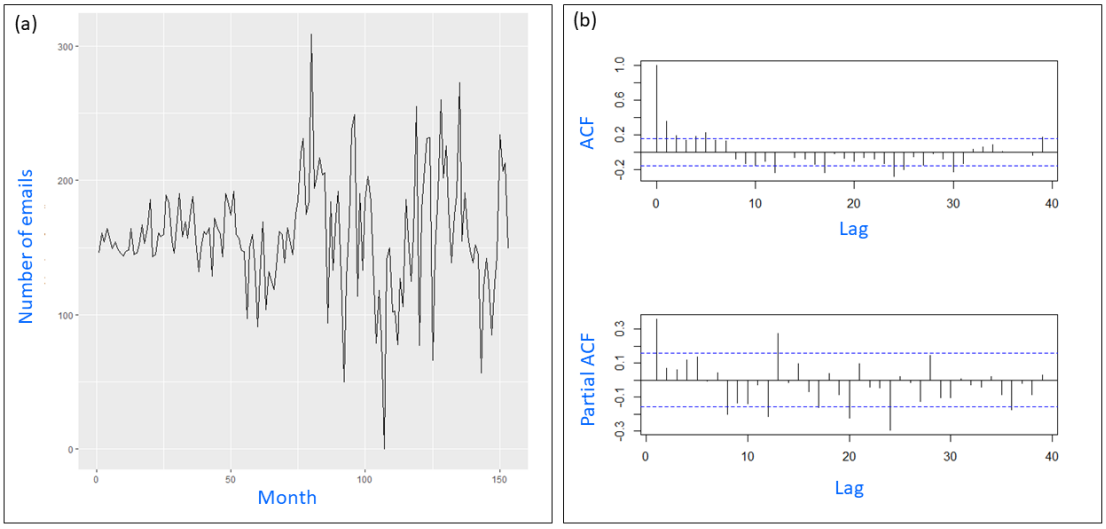


Figure 2.5: (a) Time plot of email count after removing the trend and seasonal variation by seasonal differencing. Note the change in the mean is removed. (b) The sample ACF (top right) and PACF (lower right)of seasonally adjusted and trend removed email count series.

statistics. For count data, for example, the number of phone calls or number of e-mails received over time, a Bernoulli or discrete Poisson model is used. For categorical data, a multinomial model is used in scan statistics. In addition, an exponential model can be applied for survival time data and a normal model for continuous data. In the present case, for the number of emails over time, it has been assumed that the marginal distribution of the stationary time series of integer email counts is a Poisson process. As the critical cluster size is not known, the 50% of the study period have been set as the maximum temporal cluster for the calculation of the LLR in SatScan. The p-value is obtained using MC simulation with 1000 replications. We have identified two clusters, a primary cluster and a secondary cluster using hierarchical way. The statistically significant most likely cluster is detected from June 2003 until March 2004 (LLR = 82.07 and the MC p-value = 0.001). Similarly, the statistically significant secondary cluster with log likelihood ratio of 88.94 and the MC p-value of 0.001 is observed between January 2007 and July 2008 after excluding the effect

Table 2.2: Temporal clusters of email count showing the estimated maximum loglikelihood ratio (LLR), standard Monte critical values (SMCV), Gumbel critical values (GCV) and significance level (SL) obtained using SaTSscan software.

Cluster	Time Frame	LLR	p-value	SMCV (SL)	GCV (SL)
Primary	6/1/03 - 3/31/04	82.07	0.001	12.69 (0.001)	13.99 (0.001)
Secondary	1/1/07 - 7/31/08	88.94	0.001	9.93 (0.001)	12.59 (0.001)

of the primary cluster [ZAK10], [SWY75]. The primary and the secondary clusters are shown by the rectangular boxes in Fig.2.6 (upper panel). One could see that the estimated the LLR for the primary cluster of emails is greater than the standard Monte Carlo critical value (SMCV), 12.69, and the Gumbel critical value (GCV), 13.99, at the 0.0001 level of significance (See Table 2.2).

The relative risk (RR) is an important tool that is defined as the ratio of probabilities in the two groups. In fact, the RR is the ratio of probabilities that there would be a scanning window of width, w, where observations are from the i.i.d Poisson process with different rate λ_1 and the observations in the rest of the intervals $[1, t)$ and $[t + w, N)$ are from the i.i.d Poisson process with rate λ_0. The mathematical expression of RR is given by:

$$\mathrm{RR} = \frac{\pi_1}{\pi_2}, \tag{2.9}$$

$$\log(\mathrm{RR}) = \log(\pi_1) - \log(\pi_2), \tag{2.10}$$

where π_1 is the risk inside of the scanning window of width, w, and π_2 is the risk outside of the scanning window of width, w. The null hypothesis is $H_0 : \mathrm{RR} = 1$ and the alternative hypothesis is $H_a : \mathrm{RR} > 1$.

In order to compare these results, we estimate LLR as a function of using equation 2.6

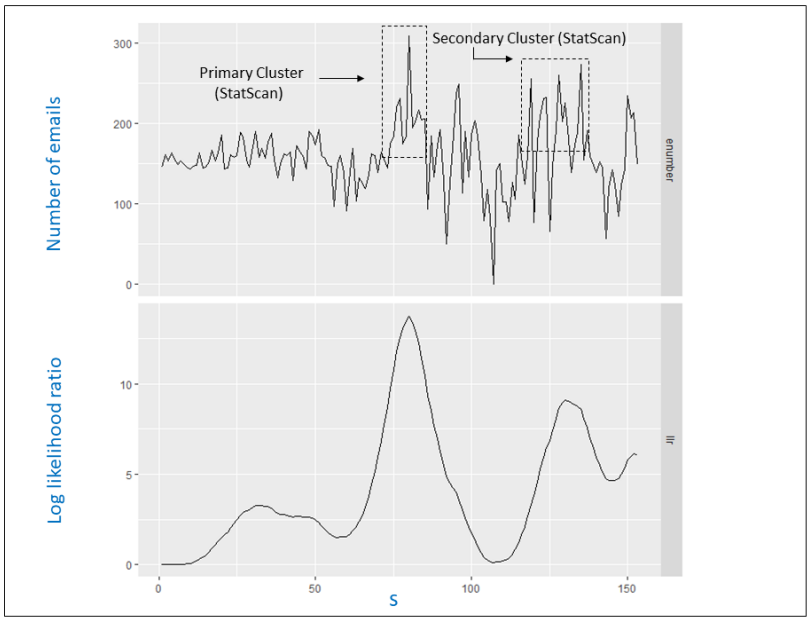


Figure 2.6: The primary and secondary clusters, obtained using SatScan software, are shown by rectangular boxes in the count series (upper panel). The LLR estimated as a function of variable and overlapping bin w, showing the primary and secondary clusters at the same time period (lower panel).

(see Fig. 2.6 (lower panel)). Note that two broad peaks appear in the time period, which are similar to the time periods estimated by SaTScan. The estimated maximum LLR is 13.75 and the estimated p-value using MC is 0.001. This analysis provides consistent result with the result obtained from the SaTScan software. Note that the SaTScan software provides the maximum likelihood for a large cluster that contains several smaller clusters.

2.5 Step-II of Two-Step Scan Process: Network Neighborhood Analysis

2.5.1 Data Processing

To form subnetworks from edgelist, the date and time were extracted from the *Date* header, and email addresses from the *From*, *Cc*, and *To* fields. In email network, an individual

may have many different email addresses known as aliases. The manual inspection was used to identify the aliases for the senders and receivers in the edge list. Usually, an email message consists of the header and the body. The header contains information on sender, receiver/receivers, subject, date, and time, and the body contains unstructured text, including the content of the email and sometimes a signature block in the end. The message may have a previous message and/or attachment. One can obtain three types of data from emails: network data, count data and text data.

An email header mainly consists of information concerning the details of the sender (who has sent the email), when the email was sent (date and time) and details of receiver (who is/are receiving the email). The most commonly used email headers are: (i) From: Senders name and email address, (ii) To: Recipients name and email address, (iii) Cc: Recipient of copies of the email, (iv) Date: Departure date and time of the email, and (v) Subject: Text entered by the sender. Furthermore, in an inter-organization, people communicate worldwide using inter-organization emails and therefore, multiple time zones may appear in *Date* header. In the current inter-organization emails, date header has Coordinated Universal Time (UTC) as well as some older standards, such as EST for Eastern Standard Time, PST for Pacific Standard Time. The UTC has a 4-digit numeric offset with a + or prefix. In this work, the local time has been converted to UTC.

In order to construct a graph from such an inter-organizational network, the participants in the emails are represented by the nodes, and the link between senders and receivers are represented as edges. Thus, in this type of network, nodes represent individuals and links represent emails, and the weight of a link between two nodes is given by the email frequencies. In addition, as the emails are time stamped, it is suitable for conducting dynamic social network analysis (DSNA). A time series of network for every week is thus constructed from emails. Furthermore, one can estimate the number of emails per week, or per month from *Date* header, thus obtaining count data.

The email data could be noisy due to the duplicate emails, multiple IDs for one person, persons having similar names, etc. The data needs to be cleaned prior to constructing the

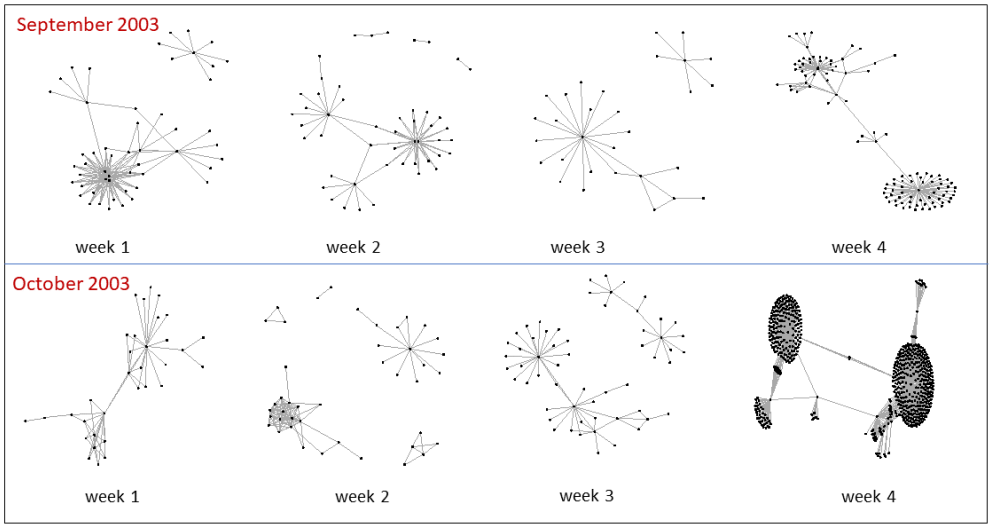


Figure 2.7: The weekly subnetworks obtained from e-mails for the period September 2003-October 2003.

network and conducting any analysis. Thus, data cleaning is a very important step, which involves cleaning the date field obtained from the *Date* header for count data, identifying the optional email addresses (aliases) for the senders and receivers in the edgelist, removing the duplicate emails and cleaning the content in text data. Fig. 2.7 shows weekly networks obtained form the email edgelist for the months of September and October, 2003.

To quantify the characteristics of a node in subgraphs, the graph invariant, such as betweenness, is used. In fact, betweenness of a node in global network measures its influential position. It is realized that more information transmits through the vertex with higher betweenness and, as a result, the vertex with higher betweenness can dominate over the entire network. A node with high betweenness can act as gatekeeper, bridge or a broker [Fre79]. Thus, the betweenness centrality applied to neighborhood network can be very useful in identifying locally important individuals and excessive activities in the network.

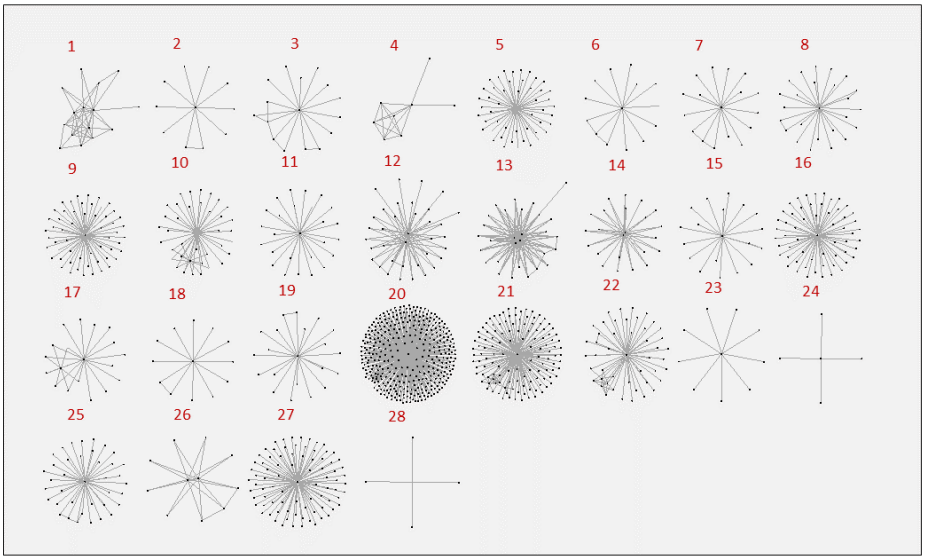


Figure 2.8: Weekly neighborhood ego subnetworks with maximum betweenness for $k = 1.5$ for the 32 week period in 2003 around the primary cluster obtained using Poisson model.

The betweenness of a vertex, v, [New10] is mathematically expressed as:

$$B(v) = \sum_{ij} \frac{g_{ij}(v)}{g_{ij}}, \tag{2.11}$$

where g_{ij} is the total number of geodesic paths from vertex i to vertex j, and $g_{ij}(v)$ is the number of geodesic paths from i to j that pass through v.

2.5.2 Formation of Ego Subnetworks Around Most Likely Cluster

The network neighborhood analysis [PCMP05, PS00, Mar12] as opposed to global network analysis is utilized here to identify interesting features at a specific point of time in which the global network is partitioned into disjoint subnetworks or subregions. The subregions are modeled subsequently by undirected binary graphs indexed by time, D_t, which are a collection of vertices that are joined by edges. Therefore, graph, D_t, can be expressed as

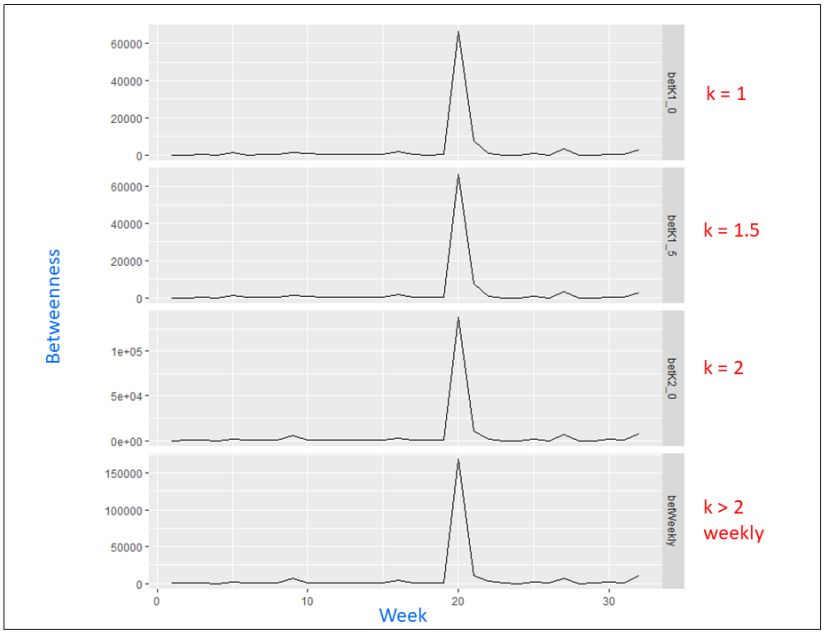


Figure 2.9: Weekly maximum betweenness series for $k = 1$ (top), 1.5 (upper middle), 2 (lower middle) and > 2 (lower).

$D_t = (V, E_t)$, where each graph has the same set of vertices, V, and different set of edges, E_t. Furthermore, a graph can be characterized by order and size, where the order of the graph, D_t, is $n = |V|$ = number of vertices, and size of graph, D_t, is $m = |E_t|$ = number of edges, respectively. From the subgraph induced at each time point, one can obtain local regions or neighborhoods of vertices. For example, mathematically the k^{th} order neighborhood of a vertex, v, of the network, D_t, is defined as $N_k[v; D_t] = \{u \in V : d_t(u, v) \leq k;\ k = 0, 1, 2, ..\}$ [PCMP05], where $d_t(u, v)$ is the geodesic distance between u and v at time, t. In fact, it is a metric space which is a set of vertices with metric defined by $d_t(u, v) \leq k$ and k is the neighborhood level. A family of neighborhood sub graphs denoted by $\Omega(N_k[v; D_t])$ with a set of vertices, $N_k[v; D_t]$, can then be generated over time.

In this work, for neighborhood network analysis, the entire data set, X, is partitioned into n disjoint subintervals $w_i;\ i = 1, 2, ..., n$ such that $X_i \in w_i$; and it can be expressed as $X = \cup_{i=1}^{n} X_i \subset \cup_{i=1}^{n} w_i$ [Mar12]. In each sub interval, the betweenness, $B_{wi}(X_i)$, was

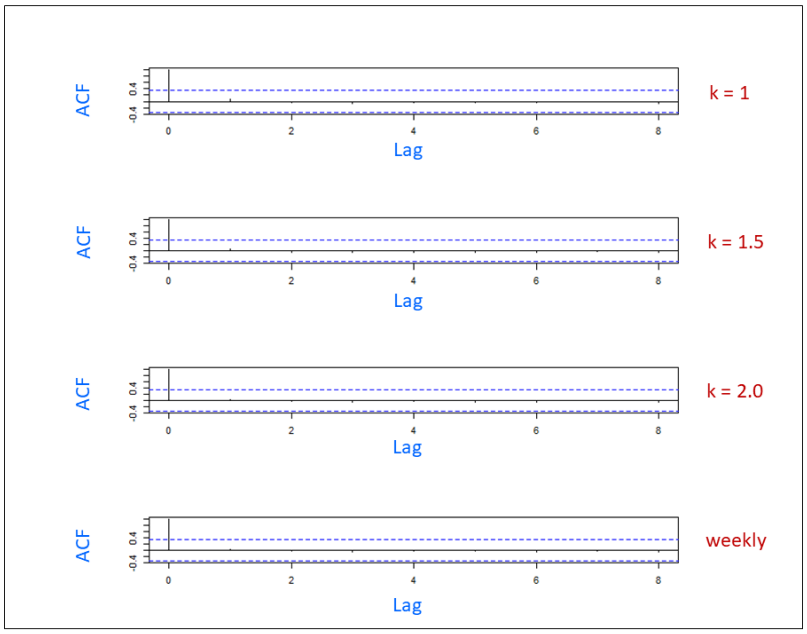


Figure 2.10: The sample ACF of the weekly maximum betweenness series for k = 1 (top), 1.5 (upper middle), 2 (lower middle) and ¿2 (lower).

estimated as the local statistics, for each vertex for $k = 1, 1.5, 2, > 2$ and then the vertex with the maximum betweenness was chosen for each neighborhood level, MB_{wi}, as the scan statistics. Fig. 2.8 is the ego subnetworks, corresponding to the highest betweenness upto 32 weeks for $k = 1.5$. One could see that the ego subnetwork [HR05] with ID = 15 with a betweenness of 66648.27 at 20^{th} week is highly dense as compared to other ego subnetworks. The corresponding time plots of the maximum betweenness for $k = 1, 1.5, 2, > 2$ are shown in Fig. 2.9. Note that a large spike for ID 15 has been observed at the 20^{th} week with a betweenness of 66795, 66648.27, 138232.77, 167972.42 for $k = 1, 1.5, 2$ and > 2, respectively. The corresponding sample ACF shows that the maximum betweenness series is stationary, as autocorrelations does not differ significantly from zero (see Fig. 2.10).

2.5.3 Estimation of LLR Using Binomial Model

As the betweenness is proportional to the number of geodesic distances around a vertex and therefore, it can be modeled as Binomial distribution. Under the null hypothesis, observations, x_1, x_2, ..., x_N are from an i.i.d Binomial distribution with success probability p_0. For alternative hypothesis, there would be a scanning window of width, w, where observations are from an i.i.d Binomial distribution with different success probability p_1, and the observations in the rest of the intervals, $[1,\ t)$ and $[t+w,\ N]$, are from an i.i.d Binomial distribution with success probability p_0. Here X_i = the number of shortest paths in the neighborhood such that:

$$X_i = \begin{cases} (n_i - 1)(n_i - 2), & \text{for directed graph,} \\ \frac{(n_i-1)(n_i-2)}{2}, & \text{for undirected graph.} \end{cases}$$

For variable and overlapping windows, p_0 and p_1 can be estimated from the data as $\hat{p_0} = \frac{\sum_{i=1}^{t} x_i}{\sum_{i=1}^{t} X_i}$ and $\hat{p_1} = \frac{\sum_{i=s}^{t} x_i}{\sum_{i=s}^{t} X_i}$ [PS00], where x_i = the number of shortest paths that is traversed by the vertex, i, in the neighborhood. The null hypothesis is $H_0 : p_0 = p_1$, and the alternative hypothesis is $H_1 : p_1 > p_0$. The likelihood under the null hypothesis is:

$$L_{H_0} = \prod_{i=1}^{N} \binom{X_i}{x_i} p_0^{x_i} (1 - p_0)^{X_i - x_i}. \tag{2.12}$$

The likelihood under the alternate hypothesis is:

$$\begin{aligned} L_{H_1} = &\left(\prod_{i=1}^{m-1} \binom{X_i}{x_i} p_0^{x_i} (1 - p_0)^{X_i - x_i} \right) \times \left(\prod_{i=m}^{t} \binom{X_i}{x_i} p_1^{x_i} (1 - p_1)^{X_i - x_i} \right) \\ &\times \left(\prod_{i=t+1}^{N} \binom{X_i}{x_i} p_0^{x_i} (1 - p_0)^{X_i - x_i} \right). \end{aligned} \tag{2.13}$$

The likelihood ratio, Λ, chosen as test statistic, is given by:

$$\Lambda = \frac{Likelihood\ \ under\ \ H_1}{Likelihood\ \ under\ \ H_0} = \frac{\left(\prod_{i=m}^{t} \binom{X_i}{x_i} p_1^{x_i}(1-p_1)^{X_i-x_i}\right)}{\left(\prod_{i=m}^{t} \binom{X_i}{x_i} p_0^{x_i}(1-p_0)^{X_i-x_i}\right)}. \tag{2.14}$$

Therefore, the LLR can be written as:

$$\left.\begin{aligned} \log(\Lambda) \quad &= \textstyle\sum_{i=m}^{t}[x_i\ \log(\frac{p_1}{p_0}) + (X_i - x_i)\ \log\frac{(1-p_1)}{(1-p_0)}] \\ &= a_1 N_t + a_2(\textstyle\sum_{i=m}^{t} X_i - N_t), \end{aligned}\right\} \tag{2.15}$$

where $a_1 = \log\left(\frac{p_1}{p_0}\right)$, $a_2 = \log\frac{(1-p_1)}{(1-p_0)}$ and $N_t = \sum_{i=s}^{t} x_i$. Using equation 2.1, one could write for rejecting H_0:

$$\max\left(\log(\Lambda)\right) \approx S_w. \tag{2.16}$$

Fig. 2.11 shows the estimated value of the LLR as a function of m. It is observed that the maximum LLR and the second maximum LLR for $k = 1.5, 2.0, \geq 2$ occur in the 20^{th} week (the 4^{th} week of October 2003) and at the 21^{st} week (the first week of November 2003), respectively. The associated ego subgraphs corresponding to the largest LLR and the second largest LLR are shown in Fig. 2.12, showing individuals with ID = 15 and 5 have the highest and the second highest betweenness, respectively. The largest LLR and the second largest LLR for $k = 1.5, 2.0, > 2.0$ are given in Table 2.3. All these estimated LLR values are greater than the critical values and therefore, statistically significant. Therefore, the clusters of shortest paths traversed by the vertices with ID = 15 (see Fig. 2.13)and 5 at the 20^{th} and the 21^{st} week, respectively, in the neighborhood are statistically significant, suggesting that excessive communications have been transmitted through these two individuals.

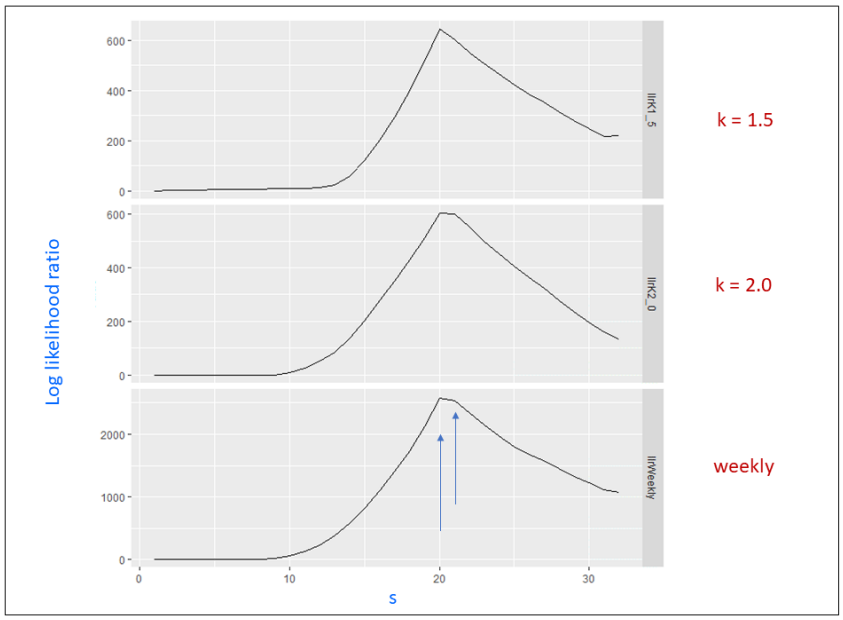


Figure 2.11: The estimated LLR as a function of variable and overlapping w for k = 1.5(top panel), 2(middle panel), > 2 (lower panel) for the 32 week period in 2003 around the primary cluster.

Table 2.3: The maximum log likelihood ratio at week 20 and week 21 respectively with Gumbel critical values (GCV), standard Monte critical values (SMCV), and significance level (SL) for k = 1.5, 2.0 and > 2.0 using the Binomial model.

K	LLR: x[20]	LLR: x[21]	GCV(SL)	SMCV(SL)]
1.5	644.11	601.069	7.73 (0.0001)	8.47 (0.001)
2.0	602.74	599.95	7.60 (0.0001)	8.51 (0.001)
> 2	2583.92	2529.41	9.00 (0.0001)	8.50 (0.001)

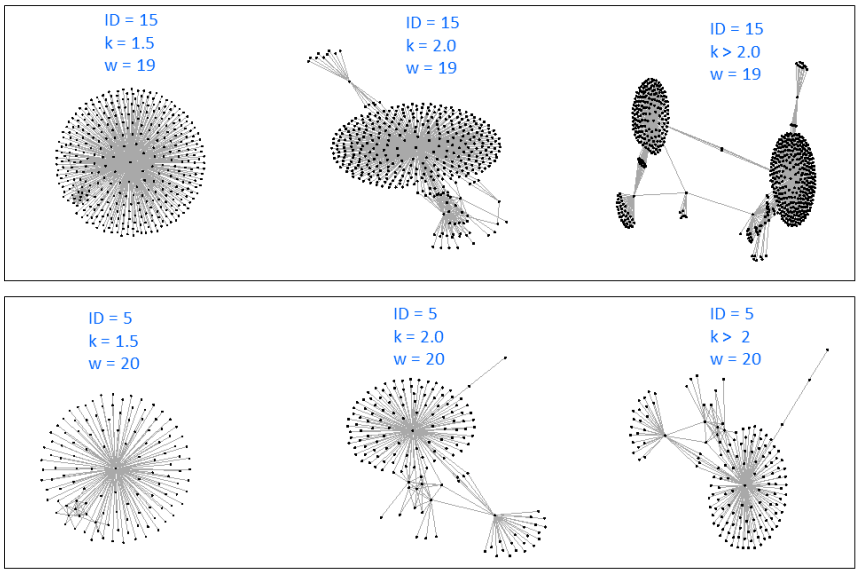


Figure 2.12: The anomalous ego sub network with ID = 15 (upper panel) detected at week $t = 20$ in 2003 for $k = 1.5$ (top left), $k = 2$(middle) and $k = > 2$ (top right). The vertex has maximum absolute betweenness score. Neighborhood ego sub networks with the second maximum absolute betweenness score (lower panel) for ID = 5.

2.5.4 Maximum Likelihood Estimation Using Non-Parametric Model

Here a nonparametric method of obtaining an estimate of the maximum likelihood of a unimodal density has been briefly discussed. This is initially reported for the continuous case by Robertson [Rob67], when mode is known. This method has later been extended by Wegman [Weg70], when the mode is unknown. He presented this estimate as conditional expectation given a Σ-lattice, $(E(f|\Sigma_L))$. Wegman [Weg70] showed that the MLE must contain the mode located at one of the observations. However, the MLE that is used as the unimodal estimator with the largest likelihood product, is not computationally very efficient as the mode has been estimated for each observation to calculate the likelihood product. In order to enhance the computational efficiency, Wegman [Weg11] later suggested placing the mode at the smallest as well as the largest order statistic. In fact, he observed that if the mode were set at the smallest order statistic, the left-hand side of the density estimate

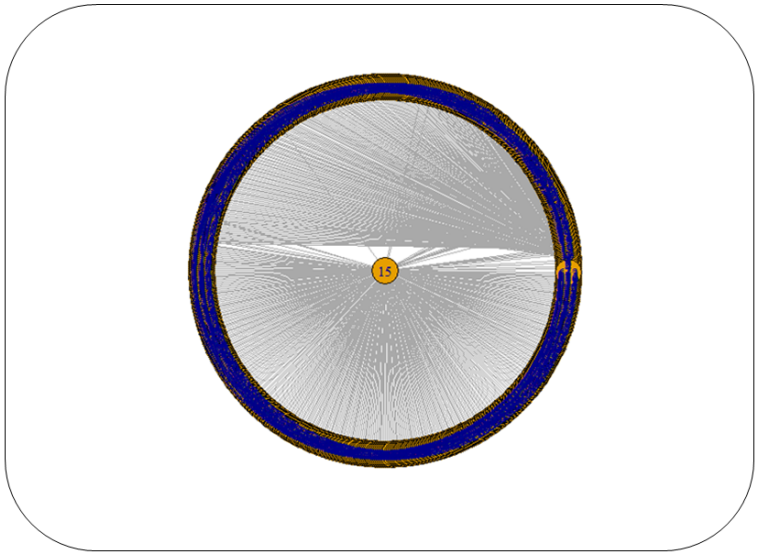


Figure 2.13: Circular plot showing ID = 15 associated with the maximum betweenness in the middle.

would be flat and the right-hand side would be non-increasing and similarly, if the mode were placed at the largest order statistic, the right-hand side would be flat and the left-hand side would be non-decreasing. Thus, the MLE would set the mode between the two flat regions/bounds. Let $x_1 \leq x_2 \leq .... \leq x_n$ is the ordered sample drawn from a unimodal density f and L and R be the left and right bounds such that $\varsigma = [L, R] = R$ - $L = \epsilon$. Let x_l and x_r be the largest observation $\leq L$ and the smallest observation $\geq R$, respectively, and h be any estimate, where $[L, R]$ is the modal interval such that:

$$E(f|\Sigma_L) = \begin{cases} h(x_l), & x_l < x < L, \\ h(x_r), & R < x < x_r, \\ h(x), & \text{otherwise.} \end{cases}$$

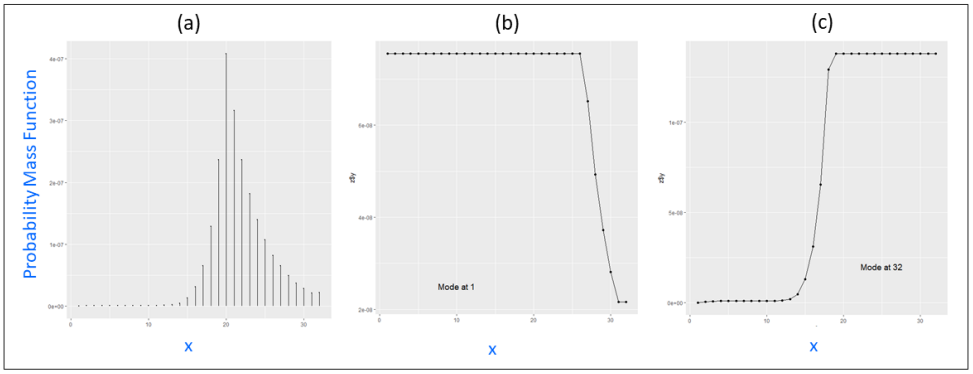


Figure 2.14: (a) The estimated pmf of the maximum betweenness as a function of ordered observations. (b) The estimated density with mode placed at the smallest order statistic. (c) The estimated density with mode placed at the largest order statistic.

Here this model was applied to estimate MLE for the unimodal discrete case. The raw probability mass function of the maximum betweenness, $f(x_i)$, such that $f(x_i) = \binom{X_i}{x_i} p^{x_i}(1-p)^{X_i - x_i}$ was initially estimated. The likelihood for each conditional mass function is written as:

$$L_j(p) = \prod_{i=1}^{32} f_j(x_{(i)}|p), \quad j = 1, 2, ...8. \tag{2.17}$$

Fig. 2.14(a) shows the raw estimate of the unimodal probability mass function (pmf) for the maximum betweenness. At first, the mode was placed at the smallest order statistic in the data sample of the maximum betweenness, which shows a flat region on the left hand side, and the non increasing region at the right (see Fig. 2.14(b)). Similarly, the mode has been placed at the largest order statistics (see Fig. 2.14 (c)) to obtain the flat region at the right, while the non-increasing region can be seen in the left. The left (L) and right

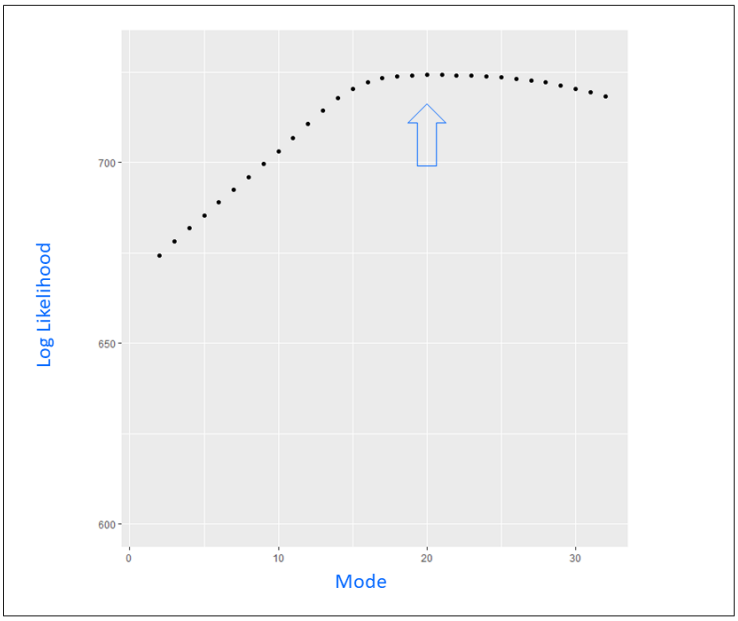


Figure 2.15: The Log likelihood estimate using a non-parametric method as a function of mode.

bounds (R) are $x[19]$ and $x[26]$, respectively. Therefore, the estimated modal interval is [19, 26]. The log likelihood estimate between two bounds as a function of mode is shown in Fig. 2.15. In fact, the MLE is associated with mode at $x[20] = 724.18$ at week 20 (the 4th week of October 2003) and the second most likely mode is $x[21] = 724.16$ (see Fig. 2.15) at week 21 (the first week of November 2003). One could see that these results are consistent with the results obtained from the parametric method.

2.6 Monte Carlo Simulation

Monte Carlo procedure is a very important method to estimate the P values. Here are the three reasons why this procedure is useful [NCS02].

1. Some test statistics do not have an exact sampling distribution.

2. Despite that a test statistic has an exact distribution, it may not be appropriate for

an inadequate sample size.

3. Estimation of exact sampling distribution may be computationally intensive.

This procedure has some limitations. To obtain p-value precision, the number of simulation needs to be increased, which is computationally intensive. On the other hand, Monte Carlo technique requires no assumption of sampling distribution of the test statistic and calculates an approximate P value. In the present work, the exact sampling distribution is not known for most cases, as the anomaly splits among multiple windows, resulting in overlapping windows and nonrandom clusters.

Let Z be the test statistic having the sampling distribution g under the null hypothesis and z be its observed value from the data then p value $= P_{H_0}(Z \geq z)$ [CH74b]. The p value can be estimated as:

$$p = \frac{r}{n}, \tag{2.18}$$

where $r = (i : z_i \geq z)$ = number of test statistic from the simulation that is greater than or equal to the observed test statistic given that n is the simulated replicates and $z_1, z_2, .., z_n$ are the test statistics from these simulation. However, this p value estimate is not strictly correct. The unbiased estimate of the true p value is given by [DH97]

$$p = \frac{(r+1)}{(n+1)}. \tag{2.19}$$

The equation 2.19 can be derived as follows. Let Y be the random variable that denotes the number that test statistic $Z \geq z$. Here the model is:

$$Y|p \approx \text{Binomial}(n, p), \tag{2.20}$$

where $p \approx \text{unif}(0, 1)$. The probability that the test statistic Z is greater than or equal to

the size z in exactly r simulations is given by [NCS02]:

$$p(Y=r) = \int_0^1 p(Y|p)f(p)dp = \binom{n}{r}\int_0^1 p^r(1-p)^{n-r}dp$$

$$= \binom{n}{r}\text{Beta}(r+1, n-r+1) = \frac{n!}{(n+1)n!} = \frac{1}{n+1}, \tag{2.21}$$

where $\text{Beta}(r+1, n-r+1) = \frac{\Gamma(r+1)\Gamma(n-r+1)}{\Gamma(n+2)} = \frac{r!(n-r)!}{(n+1)!}$. The probability that the test statistic exceeds r or greater in n simulated replicates $= p(Y \geq r) = \frac{r+1}{n+1}$.

2.6.1 Evaluation of Performance of Scan Statistic Model

The Monte Carlo simulation was used to evaluate the performance of the hypothesis test in detecting the primary cluster of events. The power of the scan statistic is the probability that the observed scan statistic exceeds the critical value k when the alternative hypothesis is true. For the discrete scan statistic case, the power is written as [GNW01]

$$P(S_w \geq k|H_A), \tag{2.22}$$

where $S_w(w) = \max\ \ Y_t(w)$, and $Y_t(w)$ is the number of events in in the scanning window w. The measured responses are the size (α) and the power of the test statistic, log likelihood ratio. The factors are test statistic, sample size and λ. Here n = 200, 500 and 1000 replications were used for the simulations. The type-I error rate (size) is measured using the proportion of the times the null hypothesis is rejected out of n replications when the null hypothesus is true and can be mathematically written as [MM16].

$$\hat{\alpha} = \frac{1}{n}\sum_{i=1}^{n}\delta_i, \tag{2.23}$$

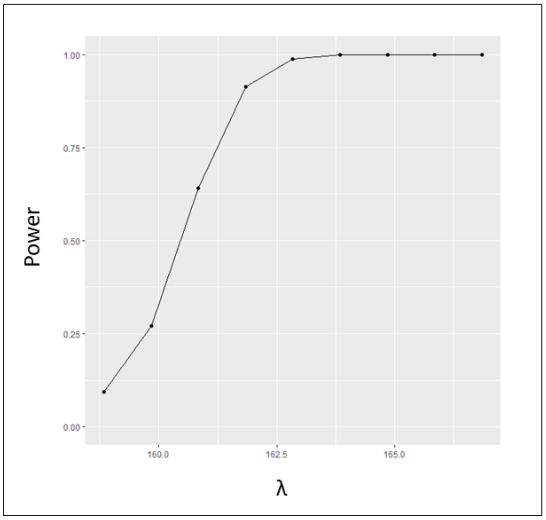


Figure 2.16: The estimated power for the Log likelihood ratio as a function of λ.

where

$$\delta_i = \begin{cases} 1, & \text{type-I error,} \\ 0, & \text{no type-I error.} \end{cases}$$

The simulated sizes and the critical values (CV) are presented in Tables 2.4 and 2.5, respectively. It was observed that the size distortion for the test statistic improves as the sample size increases. The sizes are quite closer to the nominal level when $n = 500$ and 1000.

For measuring the power, the probability of making type-II error (β), which is the proportion of the times the null hypothesis is not rejected out of n replications when the alternative hypothesis is true, has been estimated and can be mathematically expressed as [MM16]:

$$\hat{\beta} = \frac{1}{n} \sum_{i=1}^{n} \zeta_i, \tag{2.24}$$

Table 2.4: Empirical sizes (in percentage) for the LLR in testing for a cluster of events for scan statistic model.

Replications (n)	Size
200	3.5
500	4.2
1000	4.5

where

$$\zeta_i = \begin{cases} 1, & \text{type-II error,} \\ 0, & \text{no type-II error.} \end{cases}$$

Fig. 2.16 shows the estimated power as a function of λ in the range 158.8 to 170.9. The LLR based test for detection of primary cluster shows the monotonic power as λ increases. Thus, it can be concluded that the LLR based test has monotonic power, reasonable size and power performance in testing for the primary cluster of events.

2.7 Chapter Summary

Excessive activities in communication networks continue to be a major problem. In some cases, for example, Enron e-mails, frequent contact in interconnected networks could lead to fraudulent activities. Here the purely temporal clustering of emails using a two-step scan process has been investigated. In step I, the Poisson process model is initially employed to identify the most likely cluster of emails using monthly email count data for 10 year period. Initially, seasonal differencing is conducted to remove the trend and adjust the seasonal variation of monthly email count series and the most likely statistically significant

Table 2.5: The simulated critical values for the LLR for $n = 1000$.

(1- α) %	value
90	1.349809
95	1.916925
97.5	2.70383
99	3.353406
99.5	4.547692

temporal primary cluster is detected using the maximum LLR as the scan statistic (LLR = 82.07, $p = 0.001$). In step II, the binomial model is applied to weekly network data of emails for the 32 week period in 2003 in the neighborhood of the most likely cluster, where betweenness is implemented as the locality statistics and the most likely purely temporal clusters of emails are observed for $k = 1.5, 2$ and > 2 using the maximum LLR as the scan statistic. The two-step scan statistic process modeling to estimate the excessive activities as well as identifying most important nodes/vertices here for the large data set would be more scalable or computationally less intensive, as the most likely cluster in the entire data set was first identified using count data, and then the network was extracted in the vicinity of the primary cluster.

Chapter 3: Anomaly Detection Using Univariate and Multivariate Time Series Models

3.1 Introduction

In the previous chapter, the existence of significant excessive activities in the e-mail networks has been demonstrated using scan statistics. It is not apparent from the study the dynamic relationship between the influential vertices or nodes associated with excessive activities in the e-mail network. As the scan statistics measure the maximum of locality statistics, it can conceal the group of influential people in the network. In this chapter, two alternative approaches to detect clusters in a point process are implemented using the graph edit distance (GED).

Initially, the univariate time series of the GED between subsequent weekly subgraphs has been constructed and then fitted the time series to the auto regressive moving average (ARMA) model. The anomalies were then assessed using the residual thresholds obtained from the fitted time series model. Additionally, this chapter considers multiple time series of neighborhood ego subnetworks using the GED. A vector auto regressive (VAR) model was applied to fit the multiple time series. The VAR model has previously been used in economics and finance, accounting and marketing [Sim80]. In economics, this model is applied to forecast and predict macroeconomic variables, such as gross domestic product (GDP), money supply and unemployment. Here the VAR model has been implemented on time series of ego subgraphs in e-mail networks to investigate the excessive activities, as the nodes or vertices of the subgraphs are interrelated or cross correlated. Using this model, the dynamic relationship between vertices, and the excessive activities associated a vertex or node can be obtained. In addition, the group of influential persons (IDs) or vertices or nodes in the e-mail subnetworks can be identified using this process.

3.2 Univariate Time Series from e-mail Networks

3.2.1 Graph Distance Metrics

Several graph distance metrics, such as maximum common subgraph (MCS) edge distance, MCS vertex distance, graph edit distance (GED), modality distance (MD), diameter distance (DD)and spectral distance (SD) can be used to form univariate time series. Here we mostly use GED, $d(G, H)$, which is a measure of similarity (or dissimilarity) between two graphs, G and H. The GED can be defined as [Pin05]:

$$d(G, H) = |V_G|+|V_H|-2|V_G \cap V_H|+|E_G|+|E_H|-2|E_G \cap E_H|, \tag{3.1}$$

where E_G and V_G are the edge set and vertex set of a graph, G, respectively, and E_H and V_H are the edge set and vertex set of a graph, H, respectively [Pin05], [**?**]. The union of two graphs $G(V_G, E_G)$ and $H(V_H, E_H)$ are the union of their vertex sets and edge sets. It can be expressed as:

$$G \cup H = (V_G \cup V_H, E_G \cup E_H)\,. \tag{3.2}$$

Similarly, the intersection of two graphs, $G(V_G, E_G)$ and $H(V_H, E_H)$, is the union of their vertex sets and the intersection of their edge sets. It can be written as:

$$G \cap H = (V_G \cup V_H, E_G \cap E_H)\,. \tag{3.3}$$

3.2.2 Graph Edit Distance to Time Series

The GED between the weekly networks, as shown in Fig. 3.1, was obtained using Eq. 3.1. The distances between the networks for four weeks are given in Table 3.1. Fig. 3.2(a) shows the time plot of observed graph edit distance for 52 weeks around the most likely cluster obtained from scan statistics in Chapter 1. One could observe a spike at around week 20 (the 4th week of October 2003) and at week 21 (the first week of November 2003). The

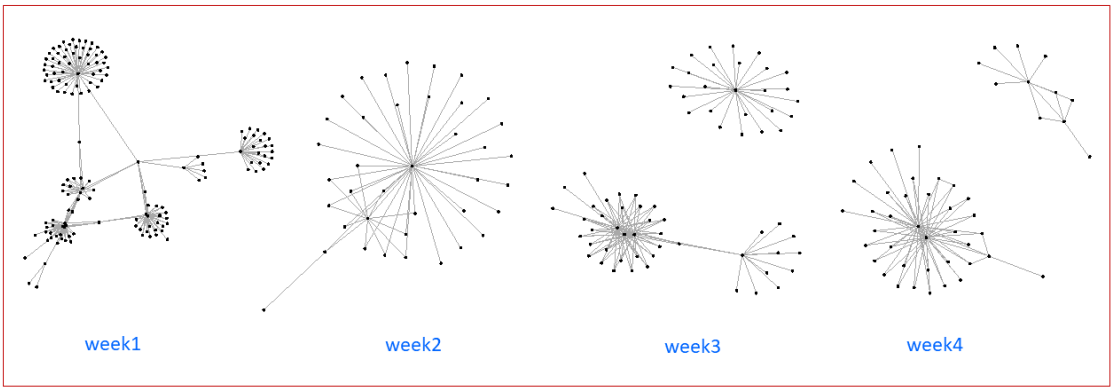


Figure 3.1: Weekly subnetworks at different time points. The GED was estimated from adjacent periods to compare subgraphs sequentially.

sample ACF and PACF show that the series is stationary (see Fig. 3.2(b)). To further validate this point, the unit root tests, such as augmented Dickey-Fuller, Phillips-Perron [CM09], [Pfa08] and KPSS [Pfa08] tests were performed to show that the time series is stationary. The values of the test statistics, lag parameters and the corresponding p-values are given in Table 3.2. A time series model was then fitted to the graph distance time series, as shown in Fig. 3.2(a) to get the residual. Here three time series models, ARMA, AR and MA are discussed.

3.2.3 AR model

The autoregressive (AR) model of order p with no constant term can be written as [SS06]:

$$x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + ... + \phi_p x_{t-p} + w_t, \tag{3.4}$$

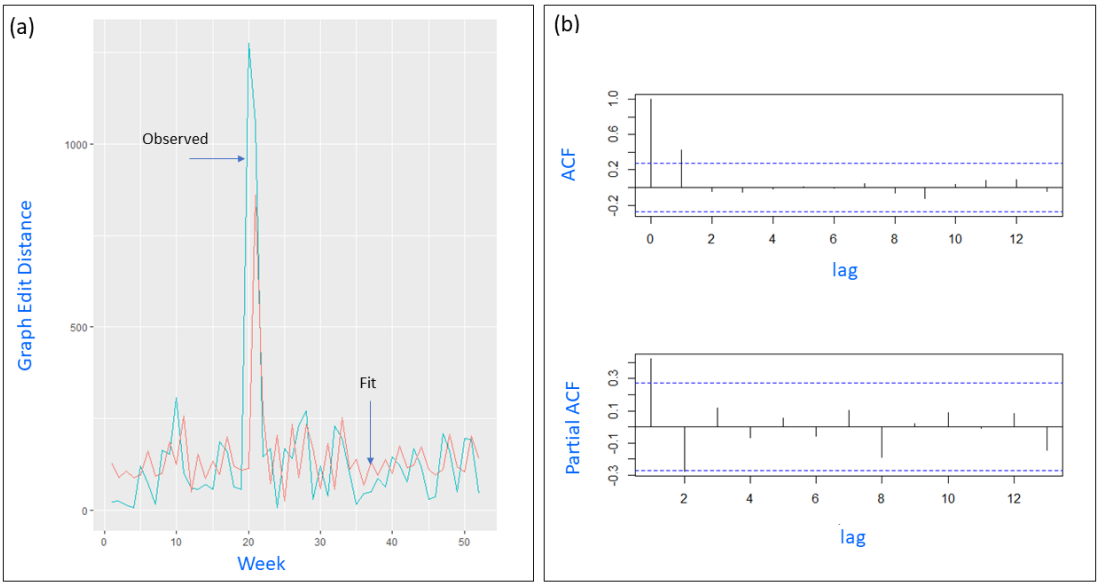


Figure 3.2: (a)Time Plots of observed and fitted GED series using ARMA Model for the 52 week period June 2003 - June 2004. (b) The sample ACF (top panel) and the sample PACF (lower panel) of weekly GED series for the 52 week period June 2003 - June 2004, respectively.

Table 3.1: The order of a graph G, the order of graph H, the size of graph G, size of a graph H, and the graph edit distance between two graphs, G and H.

	$\|V_G\|$	$\|V_H\|$	$\|E_G\|$	$\|E_H\|$	$d(G,H)$
w1-w2	73	38	93	55	73
w2-w3	38	39	55	41	15
w3-w4	39	76	41	168	164

where x_t is stationary, $\phi_1, \phi_2, ..., \phi_p$ are parameters and $\phi_p \neq 0$. Here w_t is the white noise with mean zero and variance σ_w^2. The AR (p) model is also given in useful form using

Table 3.2: Tests for unit roots showing that the time series is stationary.

Unit root test	Value of test statistics	Lag Parameter	p-value
Augmented Dickey-Fuller Test	-4.72	1	0.01
Phillips-Perron Test	-4.18	3	0.01
KPSS Test	0.145	1	0.1

backshift operator, B, as

$$\left.\begin{aligned}(1-\phi_1 B-\phi_2 B^2-...-\phi_p B^p)x_t = w_t,\\ \phi(B)x_t = w_t,\end{aligned}\right\} \tag{3.5}$$

where $Bx_t = x_{t-1}$ and $\phi(B) = 1-\phi_1 B-\phi_2 B^2+...-\phi_q B^q$, and $\phi(B)$ is called an auto regressive operator.

3.2.4 MA model

The moving average (MA) model with no constant term of order q can be written as [SS06]:

$$x_t = w_t+\theta_1 w_{t-1}+\theta_2 w_{t-2}+...+\theta_q w_{t-q}, \tag{3.6}$$

where $\theta_1, \theta_2, ..., \theta_q$ are parameters and $\theta_q \neq 0$. Using backshift operator, B, the model is written as

$$x_t = \theta(B)w_t, \tag{3.7}$$

where $\theta(B) = 1+\theta_1 B+\theta_2 B^2+...+\theta_q B^q$, and $\theta(B)$ is called a moving average operator.

3.2.5 ARMA model

The ARMA (p, q) models with no constant term are well-known statistical models for a stationary time series x_t and can be mathematically expressed as [SS06]:

$$\left.\begin{aligned} & x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t + \theta_1 w_{t-1} + ... + \theta_q w_{t-q}, \\ & (1 - \phi_1 B - \phi_2 B^2 - ... - \phi_p B^p) x_t = (1 + \theta_1 B + \theta_2 B^2 + ... + \theta_q B^q) w_t, \\ & \phi(B) x_t = \theta(B) w_t, \end{aligned}\right\} \tag{3.8}$$

where $w_t \sim iid \;\; N(0, \sigma_w^2)$ and $(\phi_1, ..., \phi_p, \theta_1, ..., \theta_q)$ is a $(p + q) \times 1$ vector of parameters, such that $\phi_p \neq 0, \theta_q \neq 0$ and $\sigma_w^2 > 0$, and p, q are the orders of the autoregressive part and the moving average part of the model, respectively. Here x_t is the current value of the time series, which is a linear function of past values of the time series and past values of the errors. In fact, x_t can be expressed as a function of autoregressive part and moving average parts. The anomalies were then assessed using the residual threshold obtained from the model fitted to the time series.

3.3 Estimation of Parameters

3.3.1 Maximum Likelihood for ARMA(p,q) Process

The likelihood function is given by [SS06]:

$$\left.\begin{aligned} & L(\phi_1, ...\phi_p, \theta_1 ....\theta_q, \sigma_w^2) = \textstyle\prod_{t=1}^{n} f(x_t | x_{t-1}, ..., x_1) \\ & = \frac{exp\left[-\frac{\sum_{t=1}^{n} \left[x_t - x_t^{t-1}(\boldsymbol{\beta}) \right]^2}{2\sigma_w^2 r_t^{t-1}(\boldsymbol{\beta})} \right]}{(2\pi\sigma_w^2)^{\frac{n}{2}} \left[(r_1^0(\boldsymbol{\beta}), r_2^1(\boldsymbol{\beta}), ..., r_n^{n-1}(\boldsymbol{\beta}) \right]^{\frac{1}{2}}}, \end{aligned}\right\} \tag{3.9}$$

where $\boldsymbol{\beta} = (\phi_1, ...\phi_p, \theta_1, ..., \theta_q)'$ and $x_t|x_{t-1}, ..., x_1 \sim N(x_t^{t-1}, P_t^{t-1})$, the conditional variance, $P_t^{t-1} = \sigma_w^2 r_t^{t-1}$ and the mean, $x_t^{t-1} = E(x_t|x_{t-1}, ..., x_1)$. The logarithm of likelihood function is given by:

$$\left.\begin{aligned} & l(\boldsymbol{\beta}, \sigma_w^2) = \log L(\boldsymbol{\beta}, \sigma_w^2) \\ & = -\tfrac{n}{2}\log(2\pi\sigma_w^2) - \tfrac{1}{2}\left[\log r_1^0(\boldsymbol{\beta}) + \log r_1^2(\boldsymbol{\beta}) ..... + \log r_n^{n-1}(\boldsymbol{\beta})\right] - \tfrac{\sum_{t=1}^n \left[x_t - x_t^{t-1}(\boldsymbol{\beta})\right]^2}{2\sigma_w^2 r_t^{t-1}(\boldsymbol{\beta})}. \end{aligned}\right\} \tag{3.10}$$

Taking the partial derivative with respect to σ_w^2 one can obtain

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma_w^2} + \frac{\sum_{t=1}^n \left[x_t - x_t^{t-1}(\boldsymbol{\beta})\right]^2}{2\sigma_w^4 r_t^{t-1}(\boldsymbol{\beta})} = 0, \tag{3.11}$$

$$\hat{\sigma}_w^2 = n^{-1}\frac{\sum_{t=1}^n \left[x_t - x_t^{t-1}(\boldsymbol{\beta})\right]^2}{r_t^{t-1}(\boldsymbol{\beta})}. \tag{3.12}$$

Newton-Raphson algorithm is a powerful numerical optimization technique for the maximum likelihood estimation of $\boldsymbol{\beta}$. Let $\boldsymbol{\beta} = (\beta_1, ..., \beta_k)$ denote k parameters. The necessary condition for maximizing $l(\boldsymbol{\beta})$ is:

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = 0, j = 1, ..., k. \tag{3.13}$$

The $k \times 1$ vector of partials is given by:

$$l^{(1)}(\boldsymbol{\beta}) = \left(\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_1}, ..., \frac{\partial l(\boldsymbol{\beta})}{\partial \beta_k}\right)'. \tag{3.14}$$

Similarly, the $k \times k$ matrix of the second order partials can be written as:

$$l^{(2)}(\boldsymbol{\beta}) = \left(\frac{\partial l^2(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \right)_{i,j=1}^{k}, \tag{3.15}$$

assuming $l^2(\boldsymbol{\beta})$ is nonsingular. If $\boldsymbol{\beta}_{(0)}$ is the initial estimate of $\boldsymbol{\beta}$, then using the Taylor expansion, one gets the following expressions:

$$0 = l^{(1)}(\hat{\boldsymbol{\beta}}) \approx l^{(1)}(\boldsymbol{\beta}_{(0)}) - l^{(2)}(\boldsymbol{\beta}_{(0)})[\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_{(0)}], \tag{3.16}$$

$$\boldsymbol{\beta}_{(1)} = \boldsymbol{\beta}_{(0)} + [l^{(2)}(\boldsymbol{\beta}_{(0)})]^{-1} l^{(1)}(\boldsymbol{\beta}_{(0)}). \tag{3.17}$$

Here $\boldsymbol{\beta}_{(2)}$ is obtained by replacing of $\boldsymbol{\beta}_{(0)}$ by $\boldsymbol{\beta}_{(1)}$, and so on. The iteration has been carried out until the sequence of estimators, $\boldsymbol{\beta}_{(1)}$, $\boldsymbol{\beta}_{(2)}$, ..., would converge to $\hat{\boldsymbol{\beta}}$, the MLE of $\boldsymbol{\beta}$.

3.3.2 Yule-Walker Estimation for an AR(p) Process

One of the commonly used methods to estimate the $AR(p)$ model parameters is Yule-Walker estimation. If $\{x_t\}$ be an autoregressive stochastic process of order p then;

$$x_t = \phi_1 x_{t-1} + ... + \phi_p x_{t-p} + w_t. \tag{3.18}$$

Multiplying Eq. 3.18 by x_{t-h} and taking expectation gives the set of equations, known as the Yule-Walker equation [SS06],

$$\left.\begin{aligned} E[x_{t-h}x_t] &= E[\phi_1 x_{t-1}x_{t-h} + ... + \phi_p x_{t-p}x_{t-h} + w_t x_{t-h}], \\ \gamma(h) &= \phi_1\gamma(h-1) + ... + \phi_p\gamma(h-p), \quad h = 1, 2, ..p, \end{aligned}\right\} \tag{3.19}$$

$$\sigma_w^2 = \gamma(0) - \phi_1\gamma(1) - ... - \phi_p\gamma(p), \;\; h = 0. \tag{3.20}$$

In matrix notation,

$$\boldsymbol{\Gamma_p \phi = \gamma_p},\tag{3.21}$$

$$\sigma_w^2 = \boldsymbol{\gamma}(0) - \boldsymbol{\phi}'\boldsymbol{\gamma}_p.\tag{3.22}$$

$$\boldsymbol{\Gamma}_p = \begin{bmatrix} \gamma(0) & \gamma(1) \ ... \ \gamma(p-1) \\ \gamma(1) & \gamma(0) \ ... \ \gamma(p-2) \\ & . \\ & . \\ & . \\ \gamma(p-1) & \gamma(p-2)...\gamma(0) \end{bmatrix}, \quad \boldsymbol{\phi} = \begin{bmatrix} \phi_1 \\ \phi_2 \\ . \\ . \\ . \\ \phi_p \end{bmatrix}, \quad \boldsymbol{\gamma}_p = \begin{bmatrix} \gamma(1) \\ . \\ . \\ . \\ . \\ \gamma(p) \end{bmatrix},\tag{3.23}$$

where $\boldsymbol{\Gamma}_p$ is $(p \times p)$ covariance matrix, and $\boldsymbol{\phi}$ and $\boldsymbol{\gamma}_p$ are $(p \times 1)$ vectors. Using the method of moments, in Eq. 3.21, replacing $\gamma(h)$ by $\hat{\gamma}(h)$, one can get

$$\hat{\boldsymbol{\phi}} = \hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p, \quad \sigma_w^2 = \hat{\gamma}(0) - \hat{\boldsymbol{\gamma}}_p'\hat{\Gamma}_p^{-1}\hat{\boldsymbol{\gamma}}_p.\tag{3.24}$$

These estimators are called the Yule-Walker estimators. In terms of sample ACF, one can write the Yule-Walker estimators as

$$\hat{\boldsymbol{\phi}} = \hat{R}_p^{-1}\hat{\boldsymbol{\rho}}_p, \quad \sigma_w^2 = \hat{\gamma}(0)\left[1 - \hat{\boldsymbol{\rho}}_p'\hat{R}_p^{-1}\hat{\boldsymbol{\rho}}_p\right],\tag{3.25}$$

where $\hat{\boldsymbol{\rho}}_p$ is a $(p \times 1)$ vector, and $\hat{\boldsymbol{R}}_p$, a $(p \times p)$ matrix, is written as:

$$\boldsymbol{R}_p = \begin{bmatrix} 1 & \rho(1) \ ... \ \rho(p-1) \\ \rho(1) & 1 \quad ... \ \rho(p-2) \\ & . \\ & . \\ & . \\ \rho(p-1) & \rho(p-2)... \quad 1 \end{bmatrix}. \tag{3.26}$$

3.3.3 Estimation Method for MA(1) Process

If $\{x_t\}$ is a moving average stochastic process of order 1, then

$$X_t = w_t + \theta w_{t-1}, \quad |\theta| < 1. \tag{3.27}$$

Multiplying both sides by X_{t-h} and taking expectation, one gets $E[X_t X_{t-h}] = E[w_t X_{t-h} + \theta w_{t-1} X_{t-h}]$. For $h = 0$, the population autocovariance is $\gamma(0) = \sigma_w^2(1 + \theta^2)$. For $h = 1$, $\gamma(1) = \sigma_w^2 \theta$, and the estimate of θ can be obtained by solving [SS06]

$$\hat{\rho}(1) = \frac{\hat{\gamma}(1)}{\hat{\gamma}(0)} = \frac{\hat{\theta}}{1 + \hat{\theta}^2}. \tag{3.28}$$

If $|\hat{\rho}(1)| \leq \frac{1}{2}$, the solutions are real. When $|\hat{\rho}(1)| < \frac{1}{2}$, the estimate is

$$\left.\begin{aligned} \hat{\theta} &= \tfrac{1-\sqrt{1-4\hat{\rho}(1)^2}}{2\hat{\rho}(1)}, \\ \hat{\sigma}_w^2 &= \tfrac{\gamma(0)}{1+\hat{\theta}^2}. \end{aligned}\right\} \tag{3.29}$$

Table 3.3: The estimated parameters and AIC of ARMA models.

Parameter	ARIMA(0,0,1)	SE	ARIMA(2,0,0)	SE	ARIMA(1,0,1)	SE
ar1			0.5470	0.1336	-0.0412	0.2417
ma1	0.611	0.1198			0.6394	0.1963
ar2			-0.2617	0.1325		
AIC	685.147		688.343		687.1195	

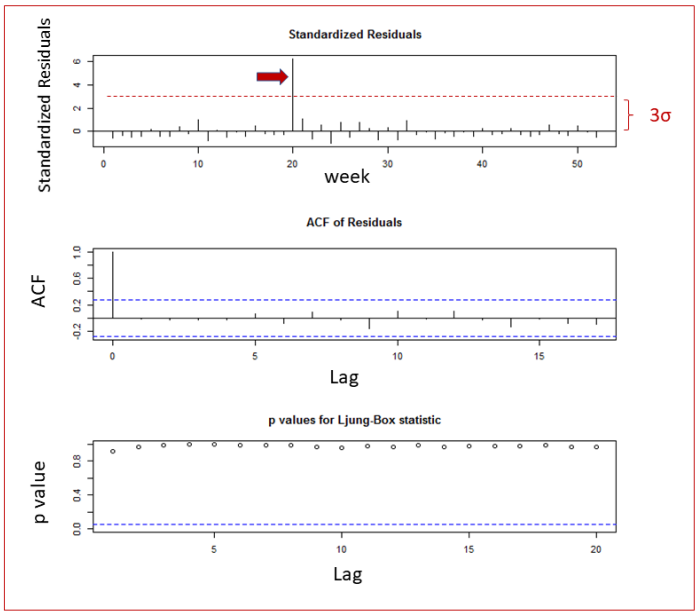


Figure 3.3: The standardized residual series (upper panel) for the MA(1)fit to the GED series and the ACF of standardized residuals for the MA(1) fit to the GED series (middle panel), and p-values for the Ljung-Box-Pierce Q-statistic (lower panel for the MA(1) fit to the GED series).

3.4 Excessive Activities and Residuals

The residual, e_t and the standardized residuals, se_t, are mathematically expressed as [SS06]

$$\left.\begin{aligned} e_t &= x_t - \hat{x_t}^{t-1}, \\ se_t &= \frac{(x_t - \hat{x_t}^{t-1})}{\sqrt{\hat{P}_t^{t-1}}}, \end{aligned}\right\} \tag{3.30}$$

where $\hat{x_t}^{t-1}$ is the one-step-ahead prediction of x_t and $\hat{P}_t^{t-1}$ is the estimated one-step-ahead error variance. The residual has some properties with mean = 0 and constant variance, i.e. homoscedasticity and normally distributed. As it is normally distributed, the 95% of all values fall between twice the standard error. Therefore, the observations lying beyond twice the standard deviations were identified as anomalies.

The augmented Dickey-Fuller test [CM09], Phillips-Perron test [CM09] and KPSS [Pfa08] test have been applied to the time series of the observed graph edit distance for 52 weeks around the most likely communication clusters obtained from the scan statistics with a spike at around week 20 (the 4^{th} week of October 2003) and week 21 (the first week of November 2003) to show that the time series is stationary. The distinct cutoff of the sample ACF at $lag1$ combined with tailing off of the sample PACF suggests an MA(1) would be an appropriate fitted model to the data [CM09]. Alternatively, the tailing off of the sample ACF and the distinct cutoff of the sample PACF suggests the AR(2) model.

The MA(1), AR(2), ARMA(1,1), ARMA(2,1) models have been fitted to the time series of GED. The optimal model is obtained based on the Akaike's information criterion (AIC) and the optimal model is an MA(1) that minimizes the AIC, followed by the ARMA(1,1),the AR(2)(see Table 3.3). The time plot of the standardized residuals (below) shows variance remains nearly constant except one observation that lies beyond the six times standard deviations (see Fig. 3.3).

The Ljung-Box-Pierce Q-statistic [SS06] is $Q = n(n+2)\sum_{h=1}^{H} \frac{\hat{\rho}_r^2(h)}{n-h}$, where $\hat{\rho}_r(h)$ is the sample autocorrelations of the residuals. The Q-test has been applied on the residuals

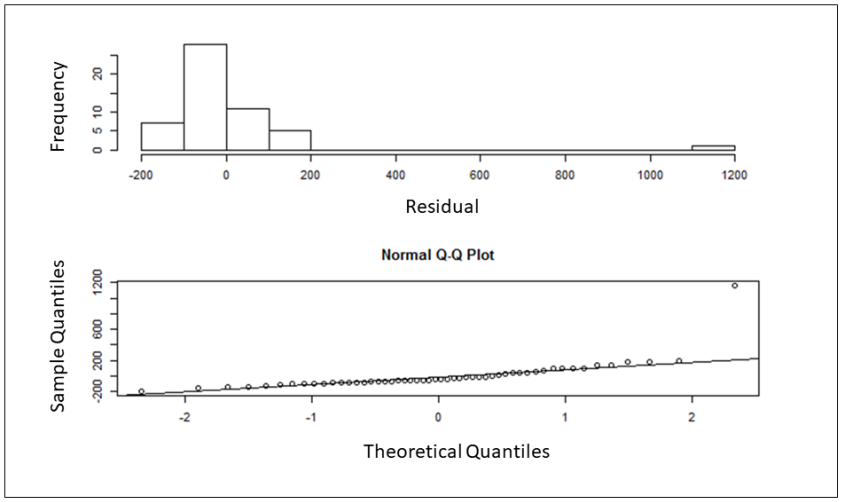


Figure 3.4: The histogram of the residuals (upper panel) and the Normal Q-Q plot of the residuals of the MA(1) fit to the GED series (lower panel), showing the residuals are close to normality except for an extreme value in the right tail.

obtained from the fitted model to check serial correlations of the residuals[Che02]. The null hypothesis for the first H autocorrelations is $H_0 : \rho_r(1) = \rho_r(2) = ... = \rho_r(H) = 0$, and the alternative hypothesis is $H_1 : \rho_r(h) \neq 0$ for some $h = 1, 2, ..., H$. Under the null hypothesis, the asymptotic distribution of Q is $Q \sim \chi^2_{H-p-q}$, where H is the number of lags. The sample auto correlation function (ACF) of standardized residual and the Q-statistic show that the residuals are random (see Fig. 3.3). One could see from the histogram and the normal Q-Q plot of residuals that the residuals are approximately normal except for an extreme value in the right tail (see Fig. 3.4). Therefore, from these analyses, it can be concluded that the spike at week 20 in the GED series is associated with excessive activities.

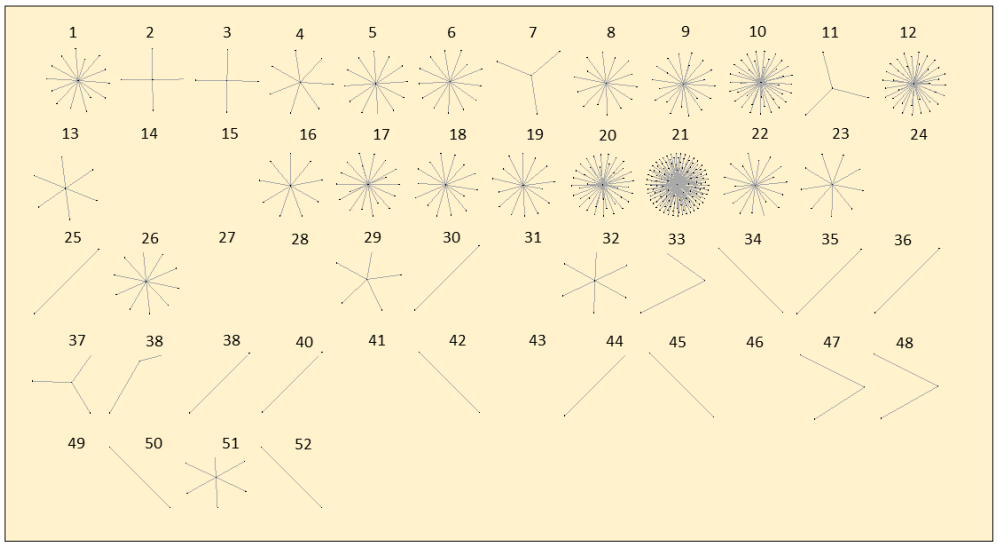


Figure 3.5: The neighborhood ego networks for $k = 1$ for ID = 5 for the 52 week period, June 2003 - June 2004.

3.5 Graph Edit Distance to Multiple Time Series

Multivariate time series is a very useful technique in time series analysis and can be used when one wants to regress a stationary time series in terms of its own past values and past values of other stationary time series. The communications in the e-mail networks are interrelated. As a result, these dynamic relationships in the email networks among IDs cannot be modeled by the univariate time series. To overcome this, the vector autoregression process has been employed to model the dynamic relationship between different IDs (nodes) that are associated with the excessive communication, and detect the chatter.

Initially, the data set, X, is partitioned around the most likely cluster obtained from the two-step scan process for 52 weeks into n disjoint weekly subintervals, w_i, $i = 1, 2, ..., n$ such that $X_i \in w_i$. A family of neighborhood ego subnetworks has been generated for different IDs in the email networks for different neighborhood levels, $k = 1, 1.5$ and 2. Figs. 3.5 and 3.6 show the ego subnetworks of ID = 5 for 52 weeks for $k = 1$ and $k = 1.5$, respectively.

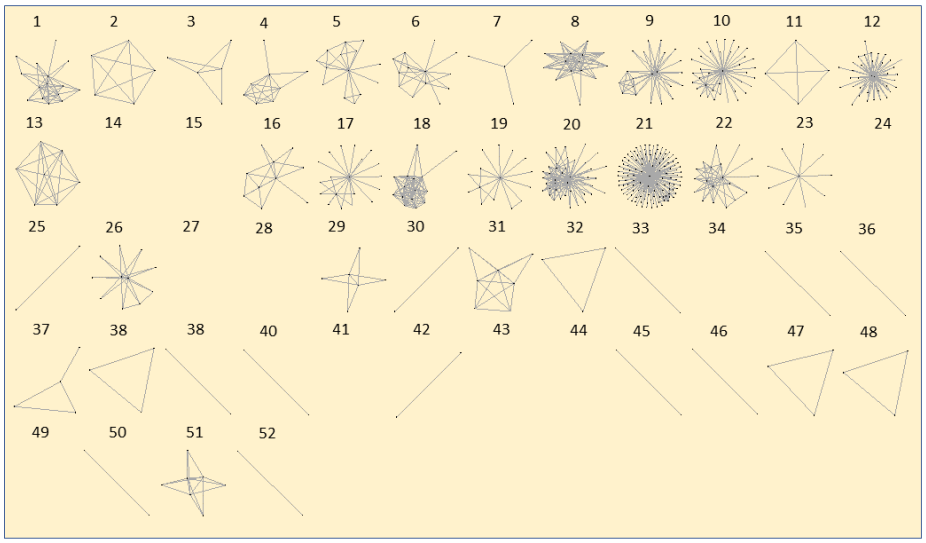


Figure 3.6: The neighborhood ego networks for $k = 1.5$ for ID = 5 for the 52 week period, June 2003 - June 2004 around the primary cluster estimated from monthly temporal scan statistic model.

The distance metric, GED, was then estimated to quantify the difference between the two consecutive ego subnetworks for $k = 1, 1.5$ and 2 to construct the multiple time series for different IDs.

In the data set, it has been observed that a number of participants have high missing values. Note that in chapter 2, the ID = 15 has been identified as the most influential person at week 20 (4th week of October 2003) using the scan statistic model. It has been observed that this individual (ID = 15) has missing values of 90% in the time series, suggesting that the individual may not be the most influential person in the network. This node (ID = 15) in the email network might be simply a connector or a hub in the email network and might not play a vital role in the network. In this research work, five individuals, IDs = 1, 5, 7, 10 and 20, with missing values between 5 to 15 % have been considered to construct the multiple time series of GED. Hence, a five-dimensional GED series has been generated.

Handling missing values is an important step in statistical analysis.In time series, the

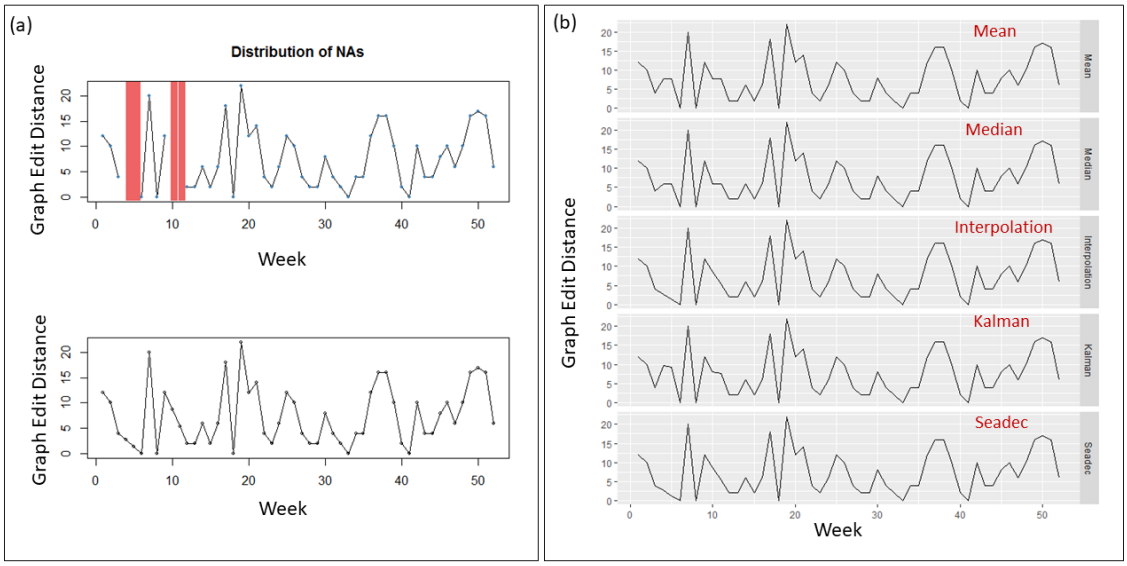


Figure 3.7: (a) Weekly GED series estimated from adjacent periods to compare subgraphs sequentially with missing values for ID = 1. Weekly GED series for ID = 1 (lower panel) after the imputation of missing values with mean. (b) The time plots of GED with different imputation methods for ID = 1.

well-known imputation algorithms are based on multiple imputation [Rub87], expectation maximization [DLR77] and nearest neighbors [VA80]. A variety of different imputation algorithms, such as imputation by mean, median and mode, imputation by linear interpolation, spline interpolation and Stineman interpolation, imputation by structural model and Kalman smoothing, and imputation by seasonally decomposed (seadec) missing values exist in the literature to handle the missing values in the time series [MBB17]. Figs. 3.7 (a) shows the time plots with missing values (upper panel) and the corresponding time plots with imputed values using mean (lower panel). Time series with the other imputation methods, such as median, interpolation, Kalman and seadec is shown for comparison (see Fig. 3.7(b)). Note that most of the algorithms give similar results. Here the mean imputation to handle time series missing values has been considered, as the time series does not show any trend and seasonality.

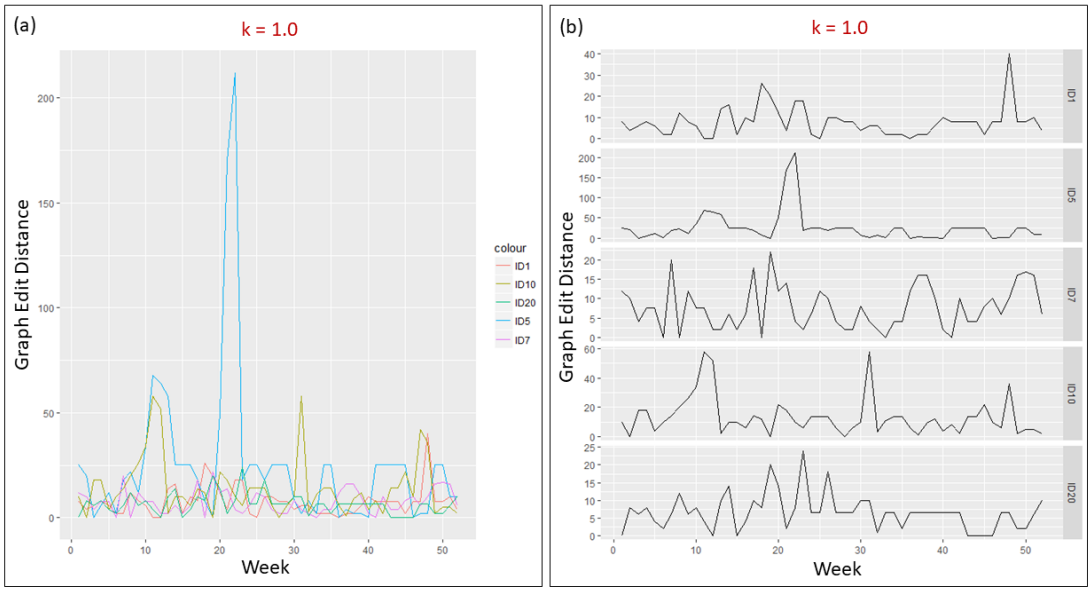


Figure 3.8: (a) A five-dimensional GED series for IDs = 1, 5, 7, 10 and 20 for $k = 1.0$. Note the spike at week = 20 for ID = 5. (b) The univariate GED series plotted separately for these IDs.

The GED between the ego subnetworks for ID = 1, 5, 7, 10 and 15 was obtained using Eq. 3.1. Figs., 3.8, 3.9 and 3.10 show the time plots of the observed graph edit distance for 52 weeks for $k = 1, 1.5$ and 2, respectively. One could see that the ego subnetworks show a large spike in the GED for ID = 5 for $k = 1.0$ and all the IDs for $k = 2$ at week 20. To implement the VAR model, the first step is to select the variables, and then investigate the stationarity of the time series. The next step is to determine the order of the VAR model, and to fit the model, and finally to perform the model checking using the residual analysis [Tsa14].

3.6 Variable Selection of the VAR model

To select the variables, the correlation between the variables need to be determined. Here the nodes with ID = 1, 5, 7, 10, and 20 have been selected for the multivariate analysis, and

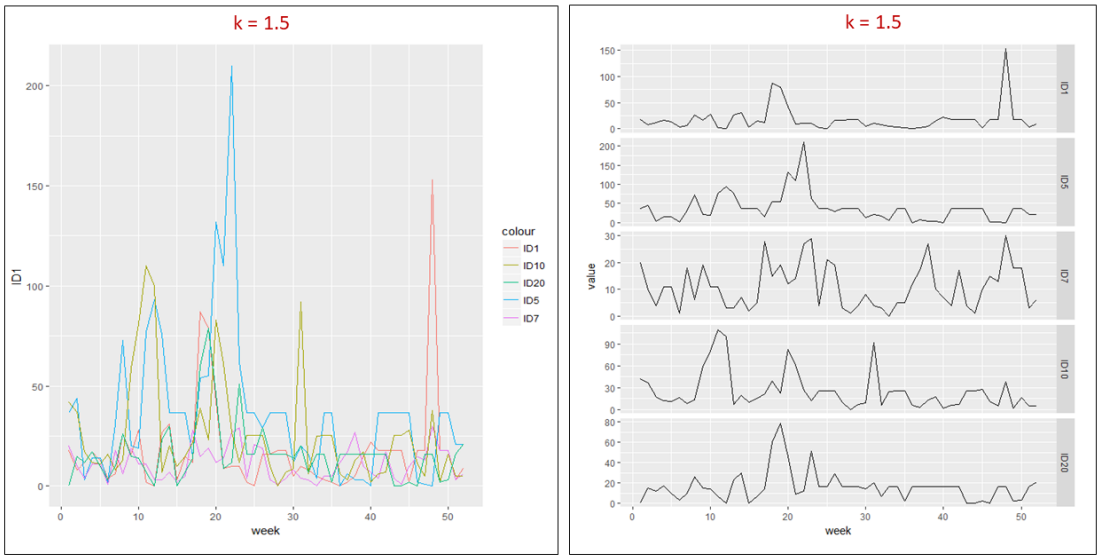


Figure 3.9: (a) A five-dimensional GED series for IDs = 1, 5, 7, 10 and 20 for $k = 1.5$. Note the spike at week = 20 for IDs = 5 and 7. (b) The univariate GED series plotted separately for these IDs.

the correlations among IDs have been checked using correlation bar plots, scatter plot matrix and parallel coordinate plots. The correlation bar plots show the correlation magnitude of the IDs increases with k (see Fig. 3.11). The different IDs are mostly positively correlated for $k = 2$. The correlation strength is more than 75 % for $k = 2$ for all variables. A similar trend for $k = 2$ can also be observed in scatter plot matrices of different variables, as shown in Fig. 3.12. For $k = 1$ and 1.5, IDs are observed to be correlated to some extent.

To extract interesting data structures, the parallel coordinate plot, which is an efficient way to represent multidimensional data, has been employed. The interesting data structures, such as the two-dimensional features (correlation and nonlinear structures) [Weg90] can be observed. Here the graph edit distances of IDs = 20, 10, 7, 5 and 1 are plotted in parallel fashion in two dimensions (see 3.13) for $k = 2$. One could see the crossing between ID 10 and ID 20 for $k = 1, 1.5, 2$, suggesting a little negative correlation whereas an approximate parallelism and relatively fewer crossings exists between the variables, suggesting a

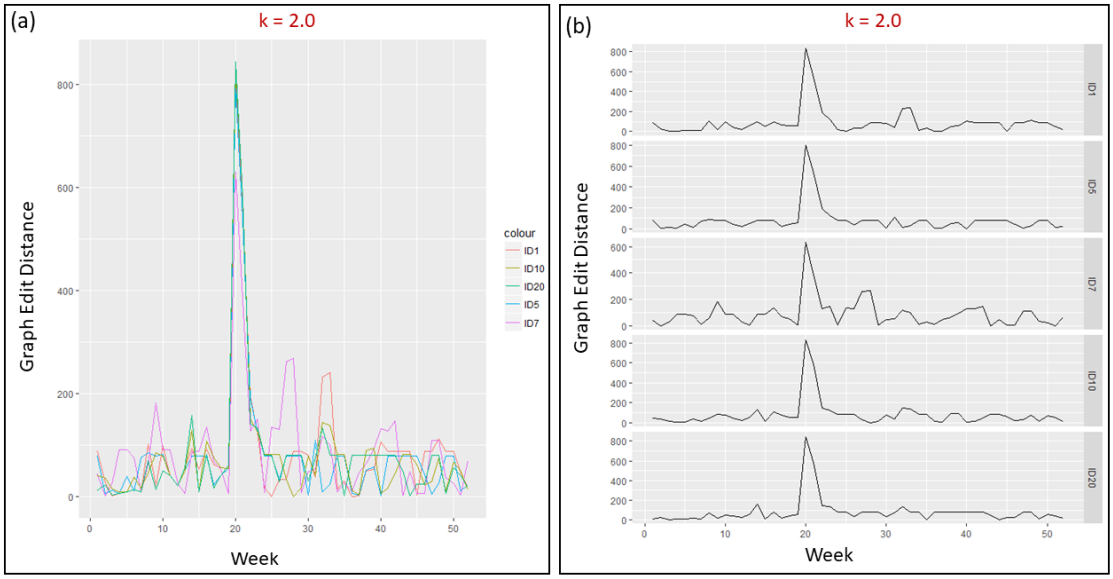


Figure 3.10: (a) A five-dimensional GED series for IDs = 1,5, 7, 10 and 20 for $k = 2.0$. Note the spike at week = 20 for IDs = 5 and 7. (b) The univariate GED series plotted separately for these IDs.

positive correlation for $k = 2$ [Weg90], [GW16]. Also, one can observe the negative slope connecting the low GED of ID 7 and ID 5 to moderate to high GED of ID 5 and ID 1, respectively, which suggests the presence of an outlier. The other interesting feature is the separation of observations for variables at high levels that is propagating across the parallel coordinates for $k = 2$, suggesting the presence of two clusters that might be associated with the excessive activities.

3.7 Vector Autoregressive Model

Let $Y_t = (y_{1,t}, y_{2,t}, ..., y_{n,t})$ be a n-dimensional multiple time series for $t = 1, 2, ..., T$, where T is the sample size of each time series with the same sample period. The first-order vector

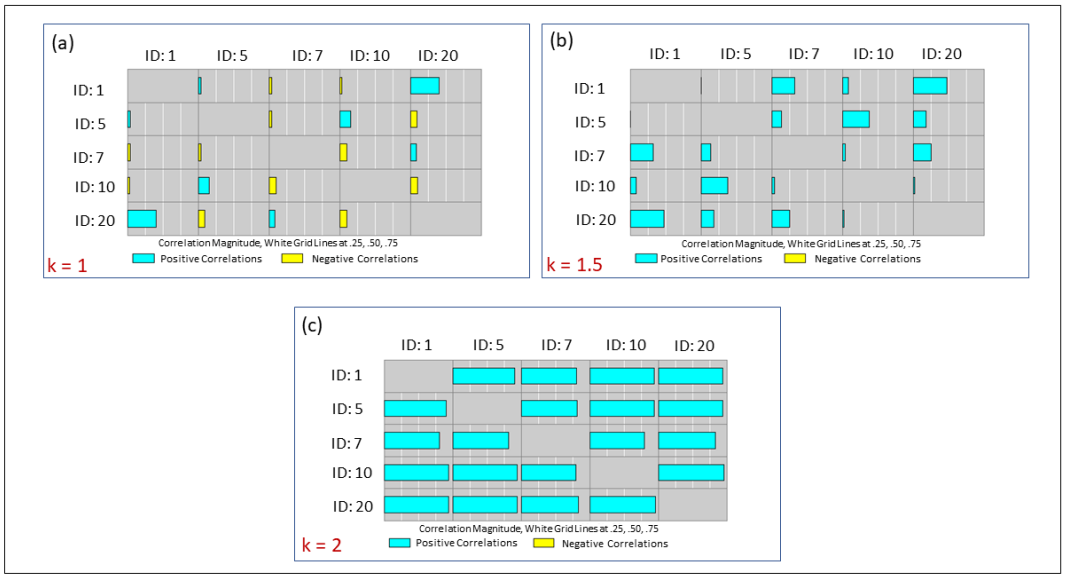


Figure 3.11: (a,b,c) Correlation bar plots of the GED for ID = 1, 5, 7, 10 and 20 with $k = 1, 1.5$ and 2, respectively.

autoregression model, VAR(1), is defined as [ZW06]:

$$\left.\begin{aligned} y_{1,t} &= \alpha_1 + \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + ... + \phi_{1n} y_{n,t-1} + w_{1,t} \\ y_{2,t} &= \alpha_2 + \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + ... + \phi_{2n} y_{n,t-1} + w_{2,t} \\ &. \\ &. \\ y_{n,t} &= \alpha_n + \phi_{n1} y_{1,t-1} + \phi_{n2} y_{2,t-1} + ... + \phi_{nn} y_{n,t-1} + w_{n,t}. \end{aligned}\right\} \tag{3.31}$$

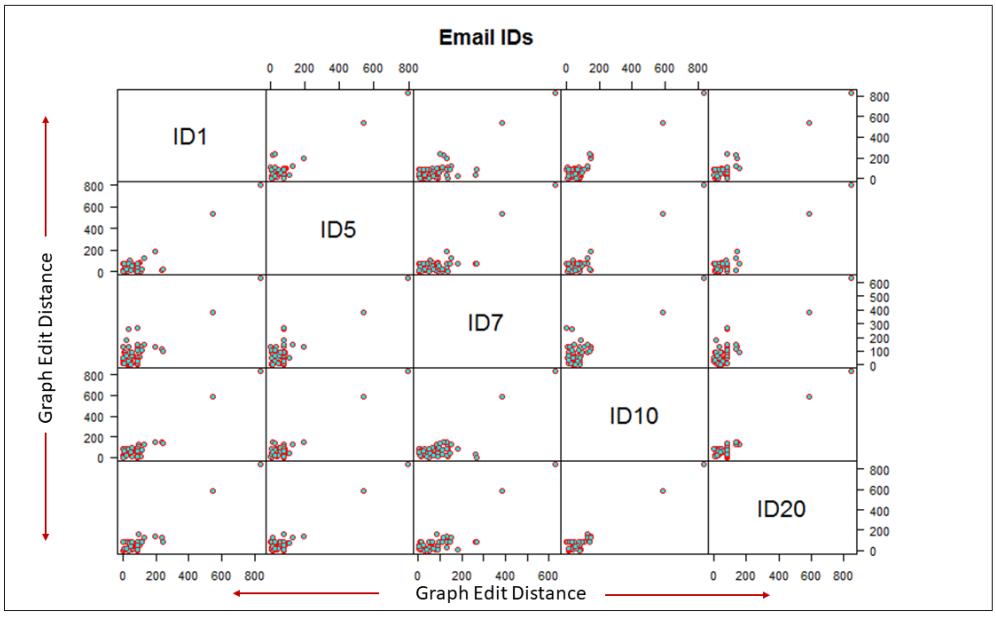


Figure 3.12: A scatterplot matrix of five-dimensional GED data illustrating correlations for $k = 2$.

The vector autoregression model of order p, VAR(p), can therefore be written as [ZW06]:

$$\left.\begin{aligned}
&y_{1,t} = \alpha_1 + \phi^1_{11} y_{1,t-1} + \phi^1_{12} y_{2,t-1} + ... + \phi^1_{1n} y_{n,t-1} + \phi^2_{11} y_{1,t-2} + \phi^2_{12} y_{2,t-2} + \\
&... + \phi^2_{1n} y_{n,t-2} + ... + \phi^p_{11} y_{1,t-p} + \phi^p_{12} Y_{2,t-p} + ... + \phi^p_{1n} y_{n,t-p} + w_{1,t} \\
&y_{2,t} = \alpha_2 + \phi^1_{21} y_{1,t-1} + \phi^1_{22} y_{2,t-1} + ... + \phi^1_{2n} y_{n,t-1} + \phi^2_{21} y_{1,t-2} + \phi^2_{22} y_{2,t-2} + \\
&... + \phi^2_{2n} y_{n,t-2} + ... + \phi^p_{21} y_{1,t-p} + \phi^p_{22} y_{2,t-p} + \ + \phi^p_{2n} y_{n,t-p} + w_{2,t} \\
&. \\
&. \\
&y_{n,t} = \alpha_n + \phi^1_{n1} y_{1,t-1} + \phi^1_{n2} y_{2,t-1} + ... + \phi^1_{nn} y_{n,t-1} + \phi^2_{n1} y_{1,t-2} + \phi^2_{n2} y_{2,t-2} \\
&+ ... + \phi^2_{nn} y_{n,t-2} + ... + \phi^p_{n1} y_{1,t-p} + \phi^p_{n2} y_{2,t-p} + \ + \phi^p_{nn} y_{n,t-p} + w_{n,t}.
\end{aligned}\right\} \quad (3.32)$$

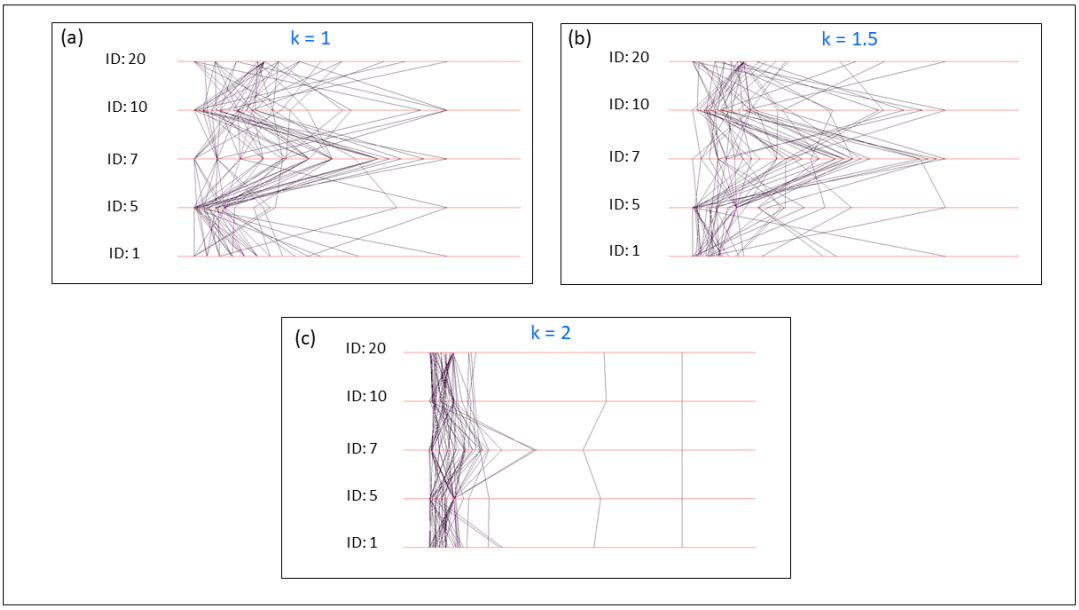


Figure 3.13: (a,b,c) Parallel coordinate plot of five-dimensional GED data showing correlations for $k = 1$, $k = 1.5$, and $k = 2$ respectively.

In matrix notations, VAR(p) can be written as:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ . \\ . \\ . \\ y_{n,t} \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ . \\ \alpha_n \end{bmatrix} + \begin{bmatrix} \phi_{11}^1 & . & \phi_{1n}^1 \\ \phi_{21}^1 & . & \phi_{2n}^1 \\ . & . & . \\ . & . & . \\ . & . & . \\ \phi_{n1}^1 & . & \phi_{nn}^1 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ . \\ . \\ . \\ y_{n,t-1} \end{bmatrix} . + \begin{bmatrix} \phi_{11}^p & \phi_{12}^p & . & \phi_{1n}^p \\ \phi_{21}^p & \phi_{22}^p & . & \phi_{2n}^p \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ \phi_{n1}^p & \phi_{n2}^p & . & \phi_{nn}^p \end{bmatrix} \begin{bmatrix} y_{1,t-p} \\ y_{2,t-p} \\ . \\ . \\ . \\ y_{n,t-p} \end{bmatrix} + \begin{bmatrix} w_{1,t} \\ w_{2,t} \\ . \\ . \\ . \\ w_{t,n} \end{bmatrix} . \tag{3.33}$$

Similarly, one can write [Lut06]:

$$\boldsymbol{Y_t = \alpha + \Phi_1 Y_{t-1} + \Phi_2 Y_{t-2} + ... + \Phi_p Y_{t-p} + W_t.} \tag{3.34}$$

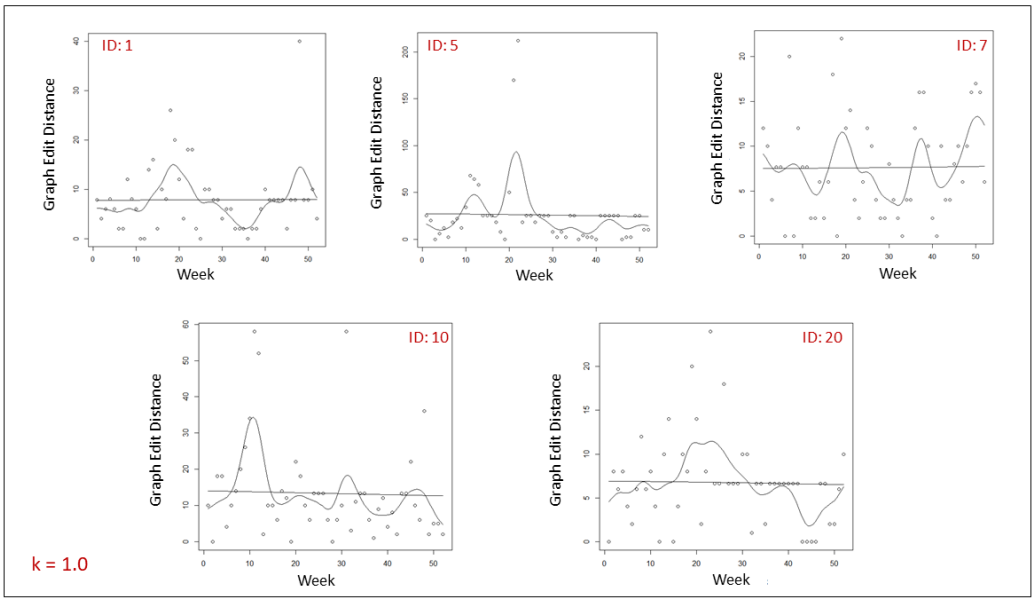


Figure 3.14: (a) Weekly GED series for IDs = 1,5, 7, 10 and 20 for $k = 1.0$ with kernel smoothing, showing no trend in the kernel fit to the series.

where

$$\boldsymbol{Y_t} = \begin{bmatrix} y_{1,t} \\ y_{2,t} \\ . \\ . \\ . \\ y_{n,t} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ . \\ . \\ . \\ \alpha_n \end{bmatrix}, \quad \boldsymbol{W_t} = \begin{bmatrix} w_{1,t} \\ w_{2,t} \\ . \\ . \\ . \\ w_{n,t} \end{bmatrix} \tag{3.35}$$

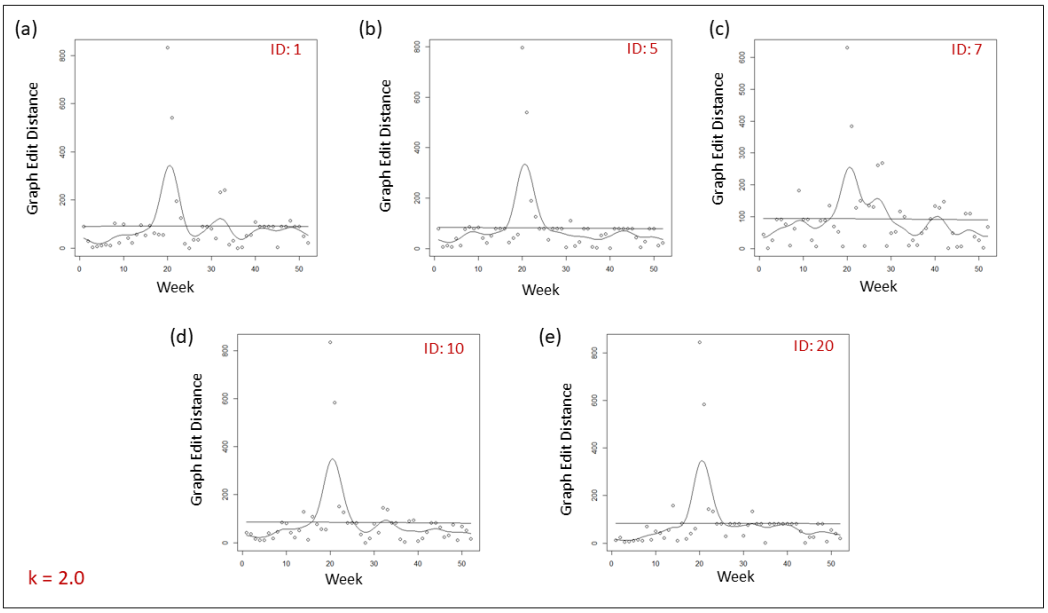


Figure 3.15: (a) Weekly GED series for IDs = 1, 5, 7, 10 and 20 for $k = 2.0$ with kernel smoothing, showing no trend in the kernel fit to the series.

are $(n \times 1)$ vectors and $\boldsymbol{\Phi_1}, \boldsymbol{\Phi_2} ... \boldsymbol{\Phi_p}$ are $(n \times n)$ coefficient matrices. $\boldsymbol{W_t}$ is a multivariate white noise, such that[Pfa08]:

$$\left.\begin{aligned} E(\boldsymbol{W_t}) &= \mathbf{0}, \\ E(\boldsymbol{W_t}\boldsymbol{W_\tau'}) &= \boldsymbol{\Sigma_W}, \quad if \;\; t = \tau, \\ E(\boldsymbol{W_t}\boldsymbol{W_\tau'}) &= \mathbf{0}, \quad if \;\; t \neq \tau, \end{aligned}\right\} \tag{3.36}$$

where $\boldsymbol{\Sigma}_W$ is time invariant positive definite covariance matrix.

3.7.1 Bivariate VAR(1) Model

The bivariate VAR(1) model is given by [Tsa14]:

$$\left.\begin{aligned} y_{1,t} &= \alpha_1 + \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + w_{1,t}, \\ y_{2,t} &= \alpha_2 + \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + w_{2,t}. \end{aligned}\right\} \tag{3.37}$$

In matrix notation,

$$\boldsymbol{Y_t = \alpha + \Phi Y_{t-1} + W_t.} \tag{3.38}$$

$$\mathbf{Y_t} = \begin{bmatrix} y_{1,t} \\ y_{2,t} \end{bmatrix}, \quad \mathbf{Y_{t-1}} = \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix}, \quad \boldsymbol{\alpha} = \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad \boldsymbol{\Phi} = \begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix}, \quad \boldsymbol{W_t} = \begin{bmatrix} w_{1,t} \\ w_{2,t} \end{bmatrix}, \tag{3.39}$$

such that $\mathbf{E(W_t W_t^{'})} = \begin{bmatrix} \sigma^2_{w_1} & \sigma^2_{w_1 w_2} \\ \sigma^2_{w_1 w_2} & \sigma^2_{w_2} \end{bmatrix} = \mathbf{\Sigma_W}$ and $\mathbf{E(W_t W_s^{'})} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix} = \mathbf{0}$ for $t \neq s$. In Eq.(3.41), $\mathbf{\Phi}$ estimates the dynamic dependence in $\mathbf{Y}$. Here the term, ϕ_{12} describes the linear dependence of $y_{1,t}$ on $y_{2,t-1}$ in presence of $y_{1,t-1}$. Similarly, the term, ϕ_{21} describes the linear dependence of $y_{2,t}$ on $y_{1,t-1}$ in presence of $y_{2,t-1}$. Here one can consider the following three cases [Tsa14].

(i) In $\mathbf{\Phi}$, if the off diagonal elements, $\phi_{12} = \phi_{21} = 0$, then Eq. 3.40 reduces to the following univariate AR(1) model

$$\left.\begin{aligned} y_{1,t} &= \alpha_1 + \phi_{11} y_{1,t-1} + w_{1,t} \\ y_{2,t} &= \alpha_2 + \phi_{22} y_{2,t-1} + w_{2,t}. \end{aligned}\right\} \tag{3.40}$$

Note that two series, $y_{1,t}$ and $y_{2,t}$, are uncoupled and dynamically uncorrelated.

(ii) If $\phi_{12} = 0$, but $\phi_{21} \neq 0$, then we get

$$\left.\begin{aligned} y_{1,t} &= \alpha_1 + \phi_{11} y_{1,t-1} + w_{1,t} \\ y_{2,t} &= \alpha_2 + \phi_{21} y_{1,t-1} + \phi_{22} y_{2,t-1} + w_{2,t}. \end{aligned}\right\} \tag{3.41}$$

In this model, note that $y_{1,t}$ does not depend on the past value of $y_{2,t}$. However $y_{2,t}$ depends on the past value of $y_{1,t}$.

(iii) If $\phi_{12} \neq 0$, but $\phi_{21} = 0$, then we have

$$\left.\begin{aligned} y_{1,t} &= \alpha_1 + \phi_{11} y_{1,t-1} + \phi_{12} y_{2,t-1} + w_{1,t} \\ y_{2,t} &= \alpha_2 + \phi_{22} y_{2,t-1} + w_{2,t}. \end{aligned}\right\} \tag{3.42}$$

In this case, note that $y_{1,t}$ depends on the past value of $y_{2,t}$. However, $y_{2,t}$ does not depend on the past value of $y_{1,t}$.

3.8 The Stationarity of Time Series

3.8.1 Stationarity Condition

One can write the bivariate VAR(1) model in companion form without the constant term as [Pfa08]:

$$\boldsymbol{Y_t = \Phi Y_{t-1} + W_t.} \tag{3.43}$$

The characteristic polynomial of $\boldsymbol{\Phi}$ is

$$\det\left(\begin{bmatrix} \phi_{11} & \phi_{12} \\ \phi_{21} & \phi_{22} \end{bmatrix} - \lambda \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right) = \det\begin{bmatrix} \phi_{11} - \lambda & \phi_{12} \\ \phi_{21} & \phi_{22} - \lambda \end{bmatrix} = \lambda^2 - (\phi_{11} + \phi_{22})\lambda + (\phi_{11}\phi_{22} - \phi_{12}\phi_{21}). \tag{3.44}$$

Table 3.4: Roots of the characteristic polynomial for $k = 1, 1.5$ and 2.

k	1	2	3	4	5
1	0.5664	0.3701	0.2957	0.2957	0.004345
1.5	0.6794	0.392	0.392	0.297	0.2318
2.0	0.5026	0.346	0.346	0.05812	0.05812

Thus, the eigenvalues of $\mathbf{\Phi}$ are the values of λ such that

$$\det(\Phi - \lambda I) = 0. \tag{3.45}$$

Y_t is a stationary process if and only if the eigenvalues of $\mathbf{\Phi}$ have moduli less than 1 [Pfa08], i.e. the values are within the unit circle. One could observe that the roots of the characteristic polynomial are less than 1, suggesting that all the variables in the VAR process are integrated of order 0 (see Table 3.4).

3.8.2 Stationarity Condition: ADF Tests

The observed GED series of ID1, ID5, ID7, ID10 and ID20 exhibits no increasing or decreasing trend with a nonzero mean (see Figs. 3.14 and 3.15). To check the stability of the GED series for all IDs, the ADF test, which is one of the unit root tests [Bro08], has been applied. The ADF test regression with constant only (no trend and nonzero mean) can be written as [NP95]:

$$\Delta Y_t = \alpha + \psi Y_{t-1} + \sum_{j=1}^{p} \lambda_j \Delta Y_{t-j} + \epsilon_t, \tag{3.46}$$

where α is a constant, $\Delta Y_t = Y_t - Y_{t-1}$ and p is the number of lagged differenced terms. The test statistic, τ, for the ADF test is defined as [Bro08]:

$$\tau 2 = \frac{\hat{\psi}}{SE(\hat{\psi})}. \tag{3.47}$$

The ADF test hypothesis is as follows,

$$\left.\begin{aligned} H_0 : \psi = 0 \Rightarrow Y_t \sim I(1), \quad \text{without drift,} \\ H_a : \psi < 0 \Rightarrow Y_t \sim I(0), \quad \text{with nonzero mean,} \end{aligned}\right\} \tag{3.48}$$

where the $I(1)$ series has one unit root, and the $I(0)$ series is a stationary process. If a nonstationary series, Y_t, becomes stationary after differencing d times, then the series Y_t is said to be integrated of order d, which can be written as $Y_t \sim I(d)$ and $\Delta^d Y_t \sim I(0)$ [Bro08]. If $\{X_t\}$ and $\{Y_t\}$ are integrated of order 1, then $X_t \sim I(1)$ and $Y_t \sim I(1)$, suggesting $\{X_t\}$ and $\{Y_t\}$ are stationary after differencing once. Then these two series, $\{X_t\}$ and $\{Y_t\}$, are cointegrated if there exists a such that $\{X_t + aY_t\}$ is stationary [EG87], [EG87], [BDGH93].

For ADF test, one has to specify the order of the serial correlation (p) or the lag length of the error term, ϵ_t [Sta10]. Ng and Perron [NP95] suggested the data dependent lag length selection method for ADF test. In fact, it is observed that with this lag length selection the size of the ADF test is reliable, and the loss of power is minimum [NP01]. The steps for this selection procedure are [NP95]; (i) Select an upper bound p_{max} for p. (ii) Then estimate the ADF test regression with $p = p_{max}$ (iii) If the absolute value of the t statistic for testing the significance of the last lagged difference is > 1.6, then the ADF test is performed. Otherwise, the lag length is reduced by one, and the process is repeated.

For an upper bound p_{max} for p, Schwert [Sch89] suggested that $P_{max} = [x]$, where $x = 12 \times (\frac{T}{100})^{\frac{1}{4}}$ and $[x]$ is the integer part of x. The $p_{max} = 10$ has been estimated for

the 5-dimensional GED series. Since the observed processes contain no trend and have nonzero mean, an ADF regression [Pfa06] with a constant (no trend and nonzero mean) has been estimated using $p_{max} = 10$. Here $lag = 1$ is selected using the Ng-Perron [NP95] algorithm. In addition, $lag = 1$ is obtained by minimizing the AIC. The critical values of the test statistic, $tau2$, for the significance levels of 1%, 5%, and 10% [Pfa06] are -3.51, -2.89 and -2.58, respectively. The test statistics and critical values are given in Table 3.5 for the observed series, ID1, ID5, ID7, ID10, and ID20 for $k = 1, 1.5$ and 2. It can be concluded that the observed series, ID1, ID5, ID7, ID10, and ID20 are integrated of order zero or stationary with a nonzero mean at 5% significance level. In addition, the autocorrelation function of each of the series decays, suggesting the observed processes are stationary (see Figs. 3.16 and 3.17).

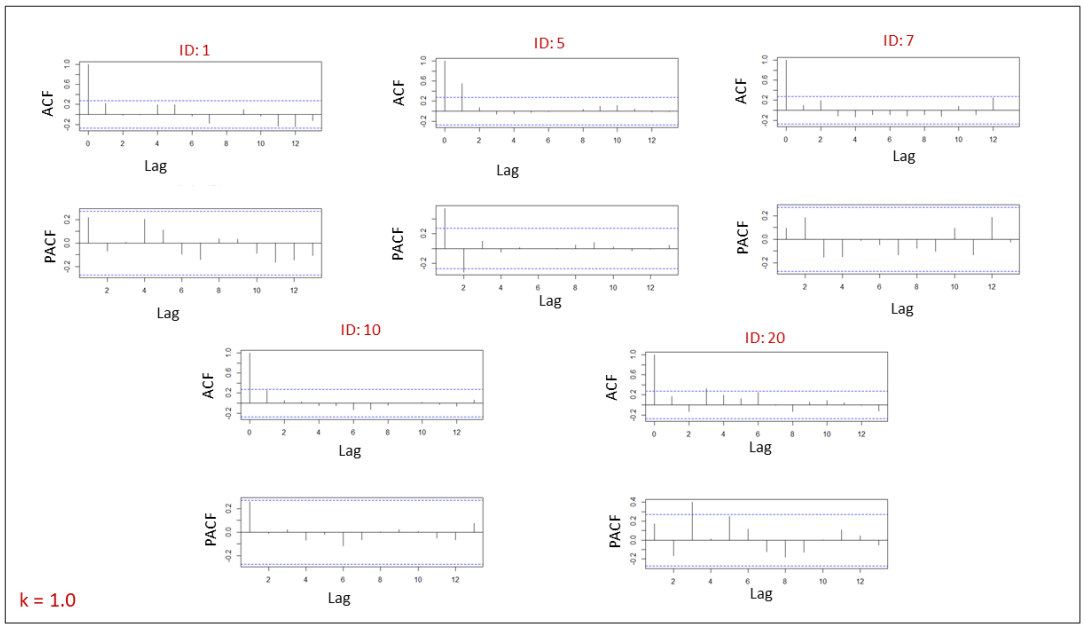


Figure 3.16: The ACF and PACF plots of the weekly GED series for IDs = 1,5, 7, 10 and 20 for $k = 1.0$, showing the series is stationary.

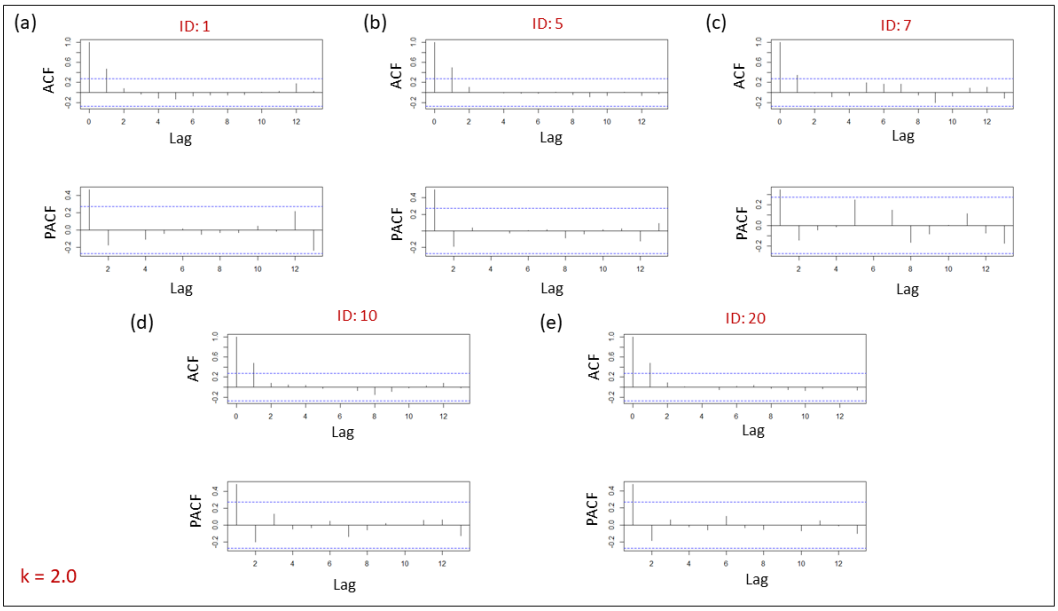


Figure 3.17: The ACF and PACF plots of the weekly GED series for IDs = 1, 5, 7, 10 and 20 for $k = 2.0$, showing the series is stationary.

3.8.3 Estimation of Parameters: Multivariate

In a VAR (p) model, the parameters of interest are $(\boldsymbol{\alpha}, \boldsymbol{\phi}_0, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_p)$ and $\boldsymbol{\Sigma}_w$. Here the multivariate least squares (MLS) and the ordinary least squares (OLS) estimation methods have been discussed for estimating the parameters. Using the matrix notation, the VAR(p) model can be written as [Lut06]:

$$\left.\begin{aligned} \boldsymbol{Y} &= \boldsymbol{BZ} + \boldsymbol{W}, \\ \text{vec}(\boldsymbol{Y}) &= \text{vec}(\boldsymbol{BZ}) + \text{vec}(\boldsymbol{W}), \\ \text{vec}(\boldsymbol{Y}) &= (\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\text{vec}(\boldsymbol{B}) + \text{vec}(\boldsymbol{W}), \\ \boldsymbol{y} &= (\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\boldsymbol{\beta} + \boldsymbol{w}, \end{aligned}\right\} \tag{3.49}$$

Table 3.5: Critical values for the ADF tests for the GED series of ID = 1, 5, 7, 10 and 20 for k = 1, 1.5 and 2.

variable	$\tau 2$ (k = 1)	$\tau 2$(k = 1.5)	$\tau 2$ (k = 2)	CV(1%)	CV(5%)	CV(10%)
ID1	-4.60	-4.25	-4.21	-3.51	-2.89	-2.58
ID5	-4.54	-2.92	-4.15	-3.51	-2.89	-2.58
ID7	-3.82	-3.87	- 4.58	-3.51	-2.89	-2.58
ID10	-4.22	-3.60	-4.26	-3.51	-2.89	-2.58
ID20	-5.20	-4.74	-4.23	-3.51	-2.89	-2.58

where $\boldsymbol{Y} = (y_1, ...y_T)$ be a $n \times T$ matrix, $\boldsymbol{Z_t} = [1, y_{t-1}, ..., y_{t-p}]'$ be a $(np+1) \times 1$ vector, $\boldsymbol{B} = (\alpha, \Phi_1, ..., \Phi_p)$ be a $n \times (np+1)$ matrix, $\boldsymbol{Z} = (Z_1, ...Z_T)$ be a $(np+1) \times T$ matrix, $\boldsymbol{W_t} = (w_1, ..., w_T)$ be a $(n \times T)$ matrix, $\boldsymbol{y} = \text{vec}(\boldsymbol{Y})$ is $(nT \times 1)$ vector by stacking the column, $\boldsymbol{\beta} = \text{vec}(\boldsymbol{B})$ is a $(n^2p + n) \times 1$ vector and $\boldsymbol{w} = \text{vec}(\boldsymbol{W})$ be a $(nT \times 1)$ vector.

The covariance matrix, $\boldsymbol{w}$, is

$$\boldsymbol{\Sigma_w} = \boldsymbol{I_T} \bigotimes \boldsymbol{\Sigma_w}. \tag{3.50}$$

The MLS estimation of $\boldsymbol{\beta}$ is obtained by minimizing

$$\left.\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{w'}(\boldsymbol{I_T} \bigotimes \boldsymbol{\Sigma_w})^{-1}\boldsymbol{w} \\ &= [\boldsymbol{y} - (\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\boldsymbol{\beta}]'(\boldsymbol{I_T} \bigotimes \boldsymbol{\Sigma_w}^{-1})[\boldsymbol{y} - (\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\boldsymbol{\beta}] \\ &= \boldsymbol{y}'(\boldsymbol{I_T} \bigotimes \boldsymbol{\Sigma_w}^{-1})\boldsymbol{y} + \boldsymbol{\beta}'(\boldsymbol{ZZ'} \bigotimes \boldsymbol{\Sigma_w}^{-1})\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\boldsymbol{Z} \bigotimes \boldsymbol{\Sigma_w}^{-1})\boldsymbol{y}. \end{aligned}\right\} \tag{3.51}$$

Taking partial derivative with respect to $\boldsymbol{\beta}$, one can get $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2(\boldsymbol{ZZ'} \bigotimes \boldsymbol{\Sigma_w}^{-1})\boldsymbol{\beta} -$

$2(\boldsymbol{Z} \bigotimes {\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1})\boldsymbol{y}$, and setting it to zero, one can obtain:

$$(\boldsymbol{Z}\boldsymbol{Z}' \bigotimes {\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1})\hat{\boldsymbol{\beta}} = (\boldsymbol{Z} \bigotimes {\boldsymbol{\Sigma}_{\boldsymbol{w}}}^{-1})\boldsymbol{y}. \tag{3.52}$$

The least squares (LS) estimator is:

$$\hat{\boldsymbol{\beta}} = ((\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z} \bigotimes \boldsymbol{I_n})\boldsymbol{y}. \tag{3.53}$$

The OLS estimation of $\boldsymbol{\beta}$ is obtained using:

$$\left.\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{w'w} = [\boldsymbol{y} - (\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\boldsymbol{\beta}]'[\boldsymbol{y} - (\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\boldsymbol{\beta}] \\ &= \boldsymbol{y'y} + \boldsymbol{\beta'}(\boldsymbol{Z}\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\boldsymbol{\beta} - 2\boldsymbol{\beta'}(\boldsymbol{Z} \bigotimes \boldsymbol{I_n})\boldsymbol{y}. \end{aligned}\right\} \tag{3.54}$$

Hence, $\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 2(\boldsymbol{Z}\boldsymbol{Z'} \bigotimes \boldsymbol{I_n})\boldsymbol{\beta} - 2(\boldsymbol{Z} \bigotimes \boldsymbol{I_n})\boldsymbol{y}$. Equating to zero, one obtains

$$\left.\begin{aligned} \hat{\boldsymbol{\beta}} &= ((\boldsymbol{Z}\boldsymbol{Z'})^{-1}\boldsymbol{Z} \bigotimes \boldsymbol{I_n})\boldsymbol{y} \\ &\Rightarrow \text{vec}(\hat{\boldsymbol{B}}) = ((\boldsymbol{Z}\boldsymbol{Z'})^{-1}\boldsymbol{Z} \bigotimes \boldsymbol{I_n})\text{vec}(\boldsymbol{Y}) \\ &= \text{vec}(\boldsymbol{Y}\boldsymbol{Z'}(\boldsymbol{Z}\boldsymbol{Z'})^{-1}. \end{aligned}\right\} \tag{3.55}$$

Thus

$$\hat{\boldsymbol{B}} = (\boldsymbol{Y}\boldsymbol{Z'}(\boldsymbol{Z}\boldsymbol{Z'}))^{-1}. \tag{3.56}$$

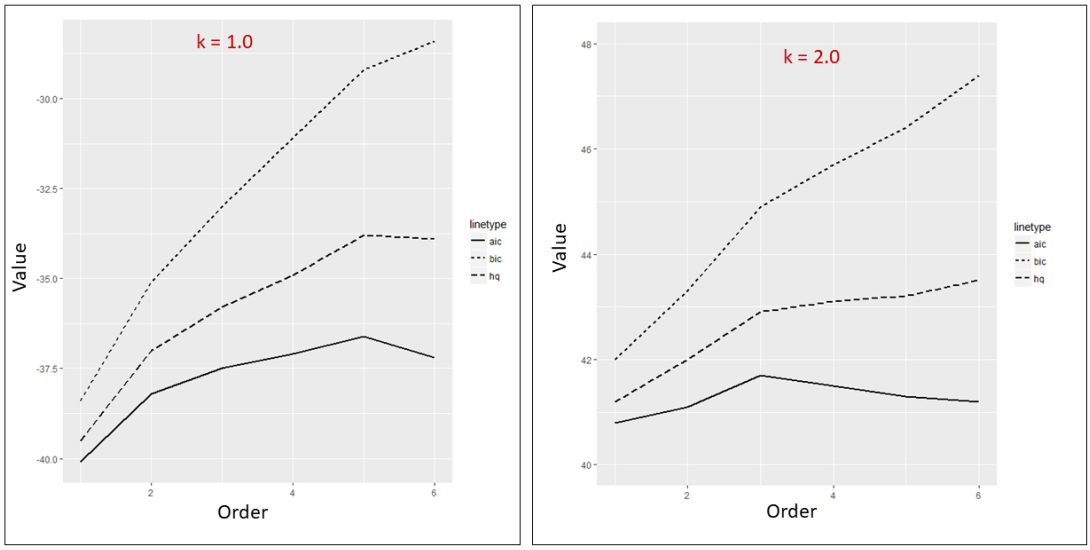


Figure 3.18: (a,b) The information criteria for the VAR models fitted to 5-dimensional series showing that the AIC, BIC and HQ are minimized when the order is 1 for $k = 1$ and 2, respectively.

The LS estimate of the covariance matrix, $\boldsymbol{\Sigma_w} = E(\boldsymbol{w_t w_t'})$, can be obtained as follows.

$$\left.\begin{aligned} \tilde{\boldsymbol{\Sigma}}_w &= \tfrac{1}{T}\textstyle\sum_{t=1}^{T} \hat{\boldsymbol{w}}_{\boldsymbol{t}}\hat{\boldsymbol{w}}_{\boldsymbol{t}}' \\ &= \tfrac{1}{T}\hat{\boldsymbol{W}}\hat{\boldsymbol{W}}' = \tfrac{1}{T}(\boldsymbol{Y}-\hat{\boldsymbol{B}}\boldsymbol{Z})(\boldsymbol{Y}-\hat{\boldsymbol{B}}\boldsymbol{Z})' \\ &= \tfrac{1}{T}\boldsymbol{Y}\left(\boldsymbol{I_T} - \boldsymbol{Z}'(\boldsymbol{Z}\boldsymbol{Z}')^{-1}\boldsymbol{Z}\right)\boldsymbol{Y}'. \end{aligned}\right\} \tag{3.57}$$

Therefore, the LS estimate of $\boldsymbol{\Sigma_w}$ is

$$\hat{\boldsymbol{\Sigma}}_{\boldsymbol{w}} = \frac{T}{(T-np-1)}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{w}}. \tag{3.58}$$

3.8.4 Information Criteria for Order Selection of VAR Model

The information criteria can be used to determine the order or lag length for the VAR(p) model. These methods have been demonstrated to be the most effective techniques for selecting a statistical model. The approach is to select the value of p that minimizes the information criteria, which is given by [ZW06]:

$$\mathrm{IC}(p) = \ln|\hat{\mathbf{\Sigma}}(p)| + c_T \times \phi(n,p), \tag{3.59}$$

where $\hat{\mathbf{\Sigma}}(p)$ is the estimated residual covariance matrix from a VAR (p) model, c_T is a sequence indexed by the sample size, T, and $\phi(n,p)$ is the penalty function. The commonly used criterion functions to determine the VAR (p) order are [Tsa14], [ZW06]:

$$\left.\begin{aligned} \mathrm{AIC}(p) &= \ln|\hat{\mathbf{\Sigma}}(p)| + \tfrac{2pn^2}{T}, \\ \mathrm{BIC}(p) &= \ln|\hat{\mathbf{\Sigma}}(p)| + \tfrac{\ln(T)pn^2}{T}, \\ \mathrm{HQ}(p) &= \ln|\hat{\mathbf{\Sigma}}(p)| + \tfrac{2ln[\ln(T)]pn^2}{T}, \end{aligned}\right\} \tag{3.60}$$

where n is the number of variables, T is the sample size and $\hat{\mathbf{\Sigma}}$ is an estimate of the covariance matrix, $\mathbf{\Sigma}$. Here AIC is the Akaike information criterion [Aka73], BIC is the Bayesian information criterion [Sch78], and HQ is proposed by Hannan and Quinn [HQ79]. Figs. 3.18(a,b) show the plots of the three information criteria, AIC, BIC and HQ as a function of order p for $k = 1$ and 2, respectively. All three criteria show that the lag length for the VAR (p) model would be 1 (see Table 3.6). A VAR(1) model, therefore, would be appropriate for the five-dimensional GED series for $k = 1, 1.5$ and 2.0. The fitted VAR(1)

model for the graph edit distance series of ID1, ID5, ID7, ID10 and ID20 for $k = 1$ is:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ y_{4,t} \\ y_{5,t} \end{bmatrix} = \begin{bmatrix} 4.804861 \\ -11.45478 \\ 7.708346 \\ 8.11249 \\ 5.272634 \end{bmatrix} + \begin{bmatrix} 0.2325 & 0.04513 & 0.1559 & -0.0516 & -0.0697 \\ -0.2704 & 0.53579 & 1.2216 & 0.6363 & 1.0922 \\ 0.2684 & -0.03356 & 0.1099 & -0.0387 & -0.2585 \\ -0.5392 & -0.00279 & 0.1053 & 0.2818 & 0.7557 \\ 0.0757 & 0.05101 & 0.0128 & -0.1042 & 0.13630 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \\ y_{4,t-1} \\ y_{5,t-1} \end{bmatrix} . \tag{3.61}$$

Similarly, the VAR(1) model for $k = 2.0$ for the graph edit distance series of ID1, ID5, ID7, ID10, and ID20 can be written as:

$$\begin{bmatrix} y_{1,t} \\ y_{2,t} \\ y_{3,t} \\ y_{4,t} \\ y_{5,t} \end{bmatrix} = \begin{bmatrix} 55.36 \\ 42.78 \\ 67.00 \\ 52.78 \\ 48.94 \end{bmatrix} + \begin{bmatrix} 0.33 & 0.27 & -0.24 & -0.49 & 0.56 \\ 0.089 & 0.29 & -0.14 & -0.021 & 0.24 \\ -0.25 & 0.076 & 0.12 & -0.14 & 0.52 \\ 0.22 & 0.38 & -0.41 & -0.16 & 0.37 \\ 0.16 & 0.31 & -0.25 & -0.30 & 0.51 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \\ y_{3,t-1} \\ y_{4,t-1} \\ y_{5,t-1} \end{bmatrix} . \tag{3.62}$$

All estimates are significant at 5% level of significance. The fitted five-dimensional models show that the graph edit distance of ID1 is dynamically related to the graph edit distance of ID5, ID7, ID10, ID20. Similarly, the graph edit distance of ID5 depends on the lagged graph edit distance of ID1, ID7, ID10, ID20, and the graph edit distance of ID7 depends on the lagged graph edit distance of ID1, ID5, ID10, ID20 and so on. One can conclude that these five individuals are interrelated. Now the residual analysis to investigate the excessive activities using VAR(1) model can be performed.

Table 3.6: Information Criteria for the VAR(p) model selection for $k = 2$.

p	AIC	BIC	HQ
1	40.78	41.98	41.24
2	41.15	43.33	41.97
3	41.68	44.86	42.87
4	41.49	45.67	43.06
5	41.26	46.43	43.20
6	41.23	47.39	43.54

3.9 Excessive Activities Using Residual Analysis of VAR(1) Model

The residual analysis for the VAR(1) model fitted to the 5-dimensional GED series has been performed to investigate the execssive activity and detect chatter. Assumptions of the model include that the residuals have no significant serial or cross-sectional correlations and heteroscedasticity. In addition, the residuals are assumed to be multivariate normally distributed. The residual matrix of a fitted VAR model is given by [LK04], [Lut06]

$$\hat{\boldsymbol{W}} = \boldsymbol{Y} - \hat{\boldsymbol{B}}\boldsymbol{Z}. \tag{3.63}$$

To check the overall significance of the residual autocorrelations [LM81] up to lag h, the Portmanteau test [Arr05], [CD04] has been performed. The null hypothesis is $H_0 : \mathbf{R_h} =$

$(R_1 = ... = R_h) = \mathbf{0}$ versus $H_1 : R_h \neq \mathbf{0}$. The Portmanteau statistic is defined as [Lut06]:

$$\left.\begin{aligned} \boldsymbol{Q_h} &= T\sum_{l=1}^{h} tr\left(\hat{\boldsymbol{R}}_{\boldsymbol{i}}{}'\hat{\boldsymbol{R}}_{\boldsymbol{w}}{}^{-\mathbf{1}}\hat{\boldsymbol{R}}_{\boldsymbol{i}}\hat{\boldsymbol{R}}_{\boldsymbol{w}}{}^{-\mathbf{1}}\right) \\ &= T\sum_{l=1}^{h} tr\left(\hat{\boldsymbol{R}}_{\boldsymbol{i}}{}'\hat{\boldsymbol{R}}_{\boldsymbol{w}}{}^{-\mathbf{1}}\hat{\boldsymbol{R}}_{\boldsymbol{i}}\hat{\boldsymbol{R}}_{\boldsymbol{w}}{}^{-\mathbf{1}}\hat{\boldsymbol{D}}^{-\mathbf{1}}\hat{\boldsymbol{D}}\right) \\ &= T\sum_{l=1}^{h} tr\left(\hat{\boldsymbol{D}}\hat{\boldsymbol{R}}_{\boldsymbol{i}}{}'\hat{\boldsymbol{D}}\hat{\boldsymbol{D}}^{-\mathbf{1}}\hat{\boldsymbol{R}}_{\boldsymbol{w}}{}^{-\mathbf{1}}\hat{\boldsymbol{D}}^{-\mathbf{1}}\hat{\boldsymbol{D}}\hat{\boldsymbol{R}}_{\boldsymbol{i}}\hat{\boldsymbol{D}}\hat{\boldsymbol{D}}^{-\mathbf{1}}\hat{\boldsymbol{R}}_{\boldsymbol{w}}{}^{-\mathbf{1}}\hat{\boldsymbol{D}}^{-\mathbf{1}}\right) \\ &= T\sum_{l=1}^{h} tr\left(\hat{\boldsymbol{C}}_{\boldsymbol{i}}{}'\hat{\boldsymbol{C}}_{\mathbf{0}}{}^{-\mathbf{1}}\hat{\boldsymbol{C}}_{\boldsymbol{i}}\hat{\boldsymbol{C}}_{\mathbf{0}}{}^{-\mathbf{1}}\right). \end{aligned}\right\} \tag{3.64}$$

The lag i residual cross-correlation matrix is written as[Lut06]:

$$\hat{\boldsymbol{C}}_i = \hat{\boldsymbol{D}}\hat{\boldsymbol{R}}_i\hat{\boldsymbol{D}}, \tag{3.65}$$

where $\hat{\boldsymbol{D}} = \sqrt{\text{diag}(\hat{C}_0)}$ is the diagonal matrix of the standard errors of the residual series and $\boldsymbol{R}_i$ is the residual correlation matrix. The test statistic has an approximate asymptotic χ^2 distribution [Lut06].

The fitted values and residuals for the VAR(1) fit to the 5-dimensional graph edit distance series of ID1, ID5, ID7, ID10, and ID20 are shown in the upper panel of the Figs. 3.19 and 3.20 for $k = 1$ and 2. For $k = 1$, the residual of ID5 at lag 20 is 113.96, which is 3.06 standard deviations above the mean indicating that it is an anomaly (see Table 3.7). On the other hand, for $k = 1$, the residual of ID1 at lag 47 is 33.12, which is only 1.32 standard deviations above the mean, indicating that it is not an anomaly. Similarly, the residuals of ID7 at lag 6, and of ID10 at lag 30, and of ID20 at lag 18 are 12.73 and 40.85 and 12.51, respectively. As these residuals are well below the 2.5 standard deviations above the mean, they are not considered anomalies.

On the other hand, the scenarios are entirely different with respect to excessive activities, when k increases to 2. For $k = 2$, the time plots of the residuals for the ID1, ID5, ID7, ID10, and ID20 series show that the variance remains nearly constant except one observation at

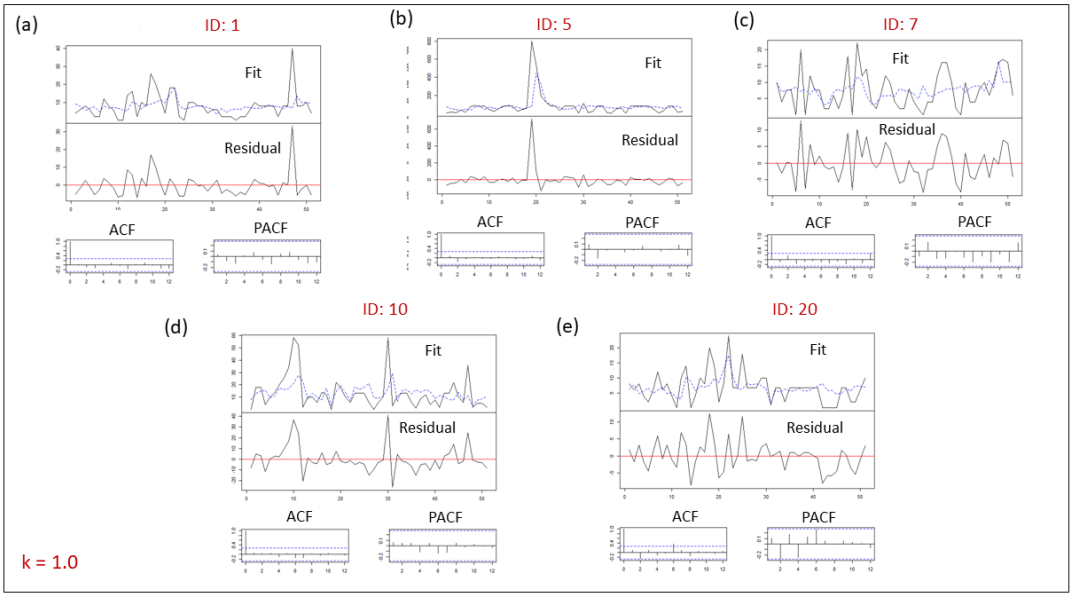


Figure 3.19: (a) Time plots of observed and fitted GED series (upper panel) and residual series (middle panel) of the VAR(1) model fit to the 5-dimensional GED series of ID = 1, 5, 7, 10 and 20 for $k = 1.0$. The ACF and PACF of the residuals are shown in the lower panel.

week 20 for each of the ID GED series that lies beyond the five times standard deviation, indicating an outlier or anomaly. The value of the residual is 739.39, 720.046, 548.087, 737.47, 756.58, respectively, for the ID1, ID5, ID7, ID10, and ID20 series. The standard deviations of the residuals obtained from the VAR(1) model fitted to the GED series of ID1, ID5, ID7, ID10, and ID20 are 134.99, 127.16, 106.79, 134.81 and 135.99, respectively. One could observe that residuals are greater than 5 times standard deviations at week 20 for all IDs, suggesting the occurrence of excessive activities at this time point (see Table 3.7).

Furthermore, the residual correlograms, and the residual partial correlograms, (see Figs. 3.19 and 3.20, bottom panels), and the residual cross-correlation matrices of the VAR(1) model (see Fig. 3.21 for $k = 2$), indicate that the residuals do not have significant serial or cross correlations as the residuals of ACF, PACF and CCF are within the bounds at the 5% level of significance. The multivariate Portmanteau test statistics have also been applied to

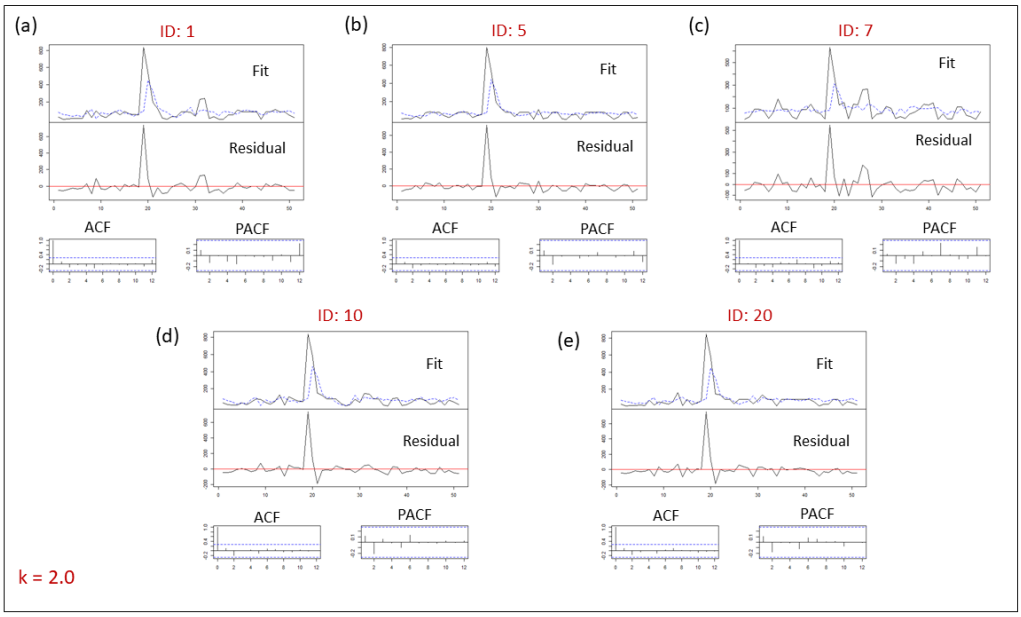


Figure 3.20: (a) Plots showing the fit (upper panel) and residual (middle panel) for the VAR(1) fit to the 5-dimensional GED series of ID = 1, 5, 7, 10 and 20 for $k = 2.0$. The ACF and PACF of the residuals are shown in the lower panel.

the residuals of the fitted VAR(1) model. From the Table 3.8 and Fig. 3.22(a, b), it can be concluded that the Q-statistic is never significant, and therefore, the residuals are not serially correlated.

3.10 Detecting Chatter

Priebe et al. [PCMP05] have used scan statistic model for detecting chatter. They have defined the order 2 ($k = 2$) statistic as the locality statistic, which is given by:

$$\tilde{\psi}_t(v) = \frac{\left(\tilde{\psi}_{2t}(v) I_{t,\tau}(v)\right)}{\max(\gamma_t(v), 1)}. \tag{3.66}$$

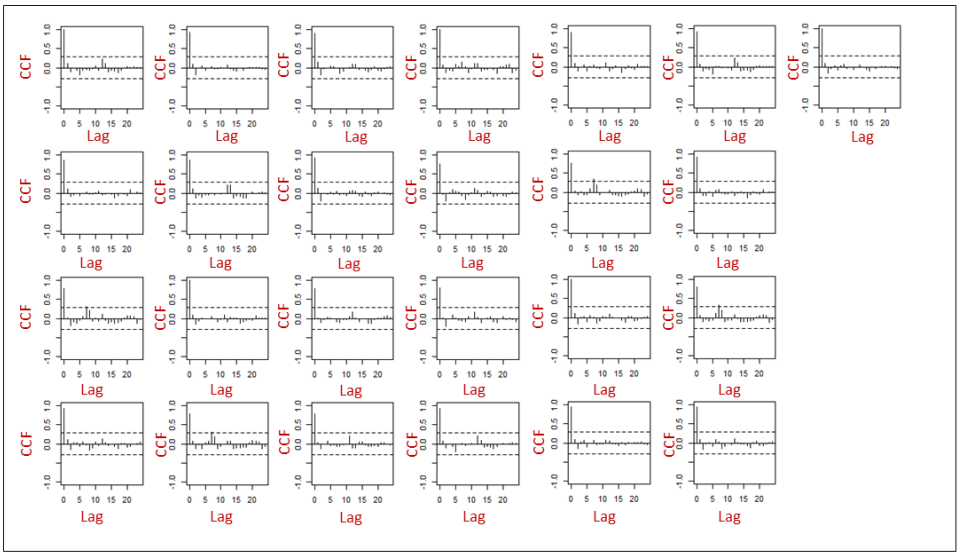


Figure 3.21: (a) The residual cross-correlation matrices for the VAR(1) model fit to the 5-dimensional GED series of ID = 1, 5, 7, 10 and 20 for k = 2.0.

Table 3.7: Excessive activity for $k = 1, 1.5$ and 2.

k	ID	Maximum Residual	SD	Times SD	Lag	Anomaly
1	ID1	33.12	25.11	1.32	47	Not an anomaly
1	ID5	113.96	37.21	3.06	20	Anomaly
1	ID7	12.73	8.22	1.55	6	Not an anomaly
1	ID10	40.85	26.18	1.56	30	Not an anomaly
1	ID20	12.51	14.99	0.83	18	Not an anomaly
2.0	ID1	739.39	134.99	5.477	20	Anomaly
2.0	ID5	720.046	127.16	5.662	20	Anomaly
2.0	ID7	548.087	106.79	5.13	20	Anomaly
2.0	ID10	737.47	134.81	5.47	20	Anomaly
2.0	ID20	756.58	135.99	5.56	20	Anomaly

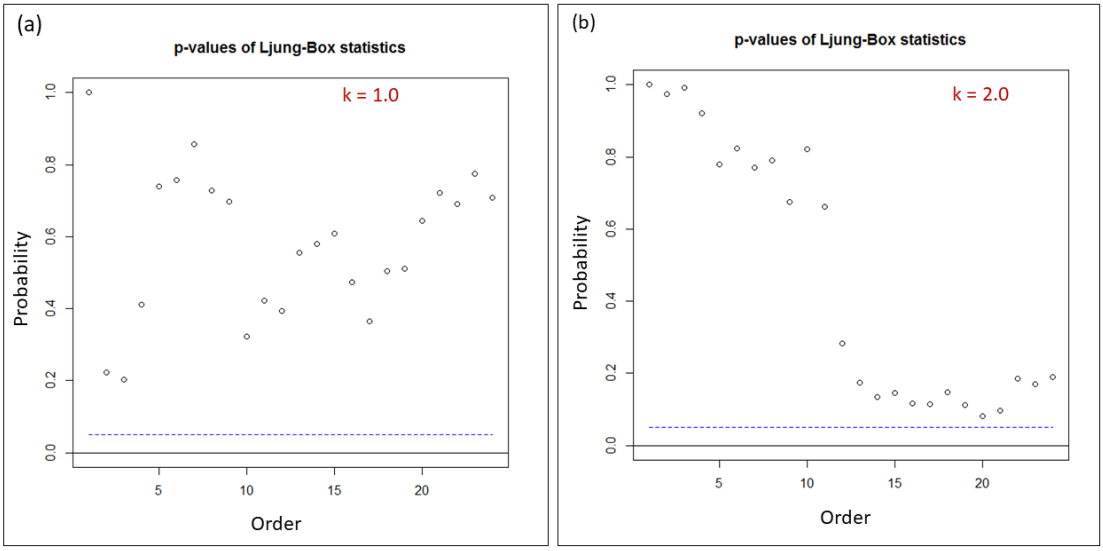


Figure 3.22: (a,b) The p-values of the multivariate Ljung-Box statistics ($Q_k(m)$) applied to the residuals of the VAR(1) model fit to the 5-dimensional GED series of ID = 1, 5, 7, 10 and 20 for $k = 1$ and 2, respectively.

where the indicator function, $I_{t,\tau}(v)$, is written as the product of three indicator functions,

$$\left.\begin{aligned} & I\left(\hat{\mu}_{0,t,\tau} > c_1\right), \\ & I\left(\psi_0(v) < \hat{\sigma}_{0,t,\tau}(v)C_2 + \hat{\mu}_{0,t,\tau}(v)\right), \\ & I\left(\psi_1(v) < \hat{\sigma}_{1,t,\tau}(v)C_3 + \hat{\mu}_{1,t,\tau}(v)\right). \end{aligned}\right\} \tag{3.67}$$

However, the scan statistic measures the maximum of the locality statistic, which can conceal the group of individuals that are associated with the excessive activity. In the present work, the chatter has been identified using a different approach, which is based on the residual obtained from the fitted VAR(1) model. The mean, μ, and the variance σ^2 of the statistic, residual, are then estimated. As the residual for each ID is approximately

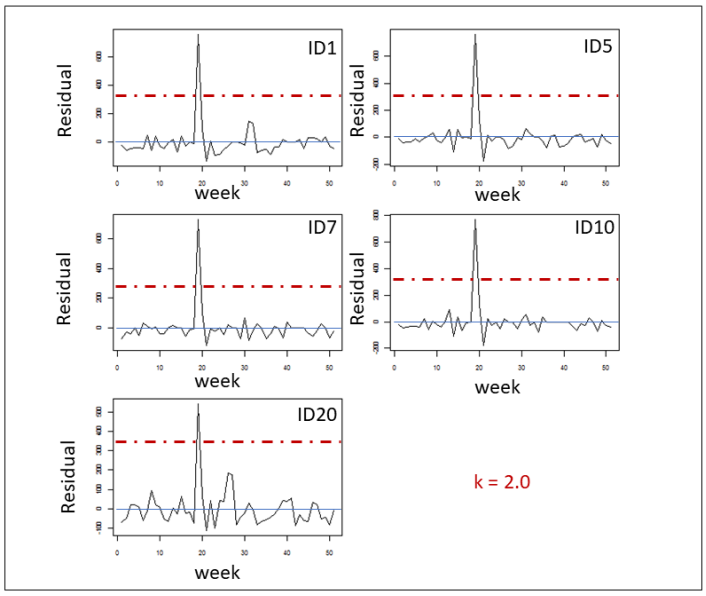


Figure 3.23: Time plots of the residuals for the VAR(1) model fit to the 5-dimensional GED series of ID = 1, 5, 7, 10 and 20 for $k = 2.0$. The chatter is detected at week = 20. The dotted red line corresponds to the residual exceeding 2.5σ standard deviations above the mean.

normally distributed with mean = 0, one can estimate a region of bounds, $R(\mu, \sigma^2)$. The bounds of this region are given by:

$$R(\mu, \sigma^2) = \hat{\mu} \pm 2.5\hat{\sigma}. \tag{3.68}$$

The fluctuation within this bound is regarded as a typical event. If the estimated residual is outside the bound, the residual is considered anomalous. It has been observed that the chatter is initiated by ID5 for $k = 1$, and as k increases, the information diffuses among the other IDs, 1, 7, 10 and 20 at week 20 (see Fig. 3.23). Therefore, ID5 is the most influential person in the email network.

3.11 Chapter Summary

In this chapter, the influential vertices or nodes associated with excessive activities in the e-mail network have been investigated. As the scan statistics measure the maximum of locality statistics, it can conceal the group of influential people in the network. Two alternative approaches have been implemented based on time series model to detect the group of influential nodes and their dynamic relationships in a point process. Initially, a univariate time series has been built using the graph edit distance, which has been estimated by the sequential comparison of the ego subnetworks for each week with those of the previous week. An autoregressive moving average process is then fitted to the GED time series, and the anomalies were assessed using residuals from the fitted model exceeding a threshold.

In addition, a vector autoregressive model fitted to the 5-dimensional GED series of the email neighborhood ego subnetworks. This represents the first known application of the VAR model to detect the chatter. This model considers the time series simultaneously of the dynamic email ego networks for the k^{th} order neighborhood, where the nodes or vertices of the subgraphs are interrelated. Anomalies in the networks are investigated using the residuals from the fitted model exceeding a threshold. A VAR(1) model using lag selection methods based on the minimization of AIC for the 5-dimensional GED series has been obtained. As the residuals for each ID is multivariate normally distributed with mean = 0, one can estimate a region of bounds, $R(\mu, \sigma^2)$. The bounds of this region are given by $R(\mu, \sigma^2) = \hat{\mu} \pm 2.5\hat{\sigma}$. The fluctuation within this bound is regarded as a typical event. If the estimated residual is outside the bound, the residual is considered anomalous. One could observe the residual greater than 3 times the standard deviations at week 20 for $k = 1$ only for ID = 5. However, the pattern changes dramatically as k increases. For $k = 2$, the time plots of the residuals show that the variance remains nearly constant except one observation at week 20 that lies beyond the five times the standard deviations, indicating an outlier or anomaly. From this multivariate time series analysis, it is concluded that the chatter has been initiated by ID = 5, and as k increases, the information spreads among other IDs. In

Table 3.8: Multivariate Portmanteau statistics of GED for $k = 2$.

m	Q (m)	df	p-value
1.00	3.43	25.00	1.00
2.00	32.40	50.00	0.97
3.00	48.97	75.00	0.99
4.00	80.81	100.00	0.92
5.00	112.61	125.00	0.78
6.00	133.91	150.00	0.82
7.00	160.92	175.00	0.77
8.00	183.67	200.00	0.79
9.00	214.83	225.00	0.68
10.00	229.31	250.00	0.82
11.00	264.76	275.00	0.66
12.00	313.64	300.00	0.28
13.00	348.79	325.00	0.17
14.00	379.44	350.00	0.13
15.00	404.02	375.00	0.15
16.00	434.05	400.00	0.12
17.00	460.29	425.00	0.11
18.00	481.49	450.00	0.15
19.00	512.90	475.00	0.11
20.00	544.87	500.00	0.08
21.00	567.45	525.00	0.10
22.00	579.69	550.00	0.18
23.00	607.46	575.00	0.17
24.00	630.23	600.00	0.19

addition, this analysis clearly demonstrates the dynamic social relationship between ID = 5 and other IDs = 1, 7, 10 and 20. It can be concluded that the ID = 5 is the most influential person of this email network.

Chapter 4: Pattern Retrieval and Anomaly Detection from E-Mail Content

4.1 Introduction

In the previous two chapters, purely temporal clustering of emails using scan statistics and time series models was investigated. The Poisson process model is initially employed to identify the most likely cluster of emails using email count data for 10 year period. The most likely statistically significant temporal primary cluster is detected using the maximum LLR as the scan statistic (LLR = 82.07, p = 0.001). Then the binomial model is applied to network data of emails for 52 weeks in the neighborhood of the most likely cluster, where betweennes is implemented as the locality statistic and the most likely purely temporal clusters of emails are observed for $k = 1.5$, 2 and > 2 using the maximum LLR as the scan statistic. I also perform the residual analysis of the MA(1) fitted to the GED series, and observe the statistically significant excessive activity. Both approaches, scan statistics applied to count and network data, and residual analysis of graph edit distance, provide consistent results, exhibiting excessive activity in email data. Here I analyze an unstructured textual data obtained from email contents around the primary cluster from June 2003 to June 2004 (52 weeks), and investigate the major topic discussed in this period using text mining algorithms and probabilistic modeling, such as latent Dirichlet allocation (LDA) modeling. I then use scan statistics to get the excessive topic activities.

With the increasing amount of text documents, the need for extracting information quickly from the massive unstructured textual data, such as, emails, tweets and other social media, websites, research reports, survey and blogs using statistical natural language processing (SNLP) has grown. Applications and techniques of SNLP include text clustering, information retrieval, text categorization or text classification, and text summarization

[MS02]. In text clustering, a corpus, which is a collection of documents, is partitioned into groups [Ren94]. On the other hand, documents are classified through text categorization into two or more specified classes, and text summarization automatically extracts a summary of a document [SSBM96], [SM83]. The most widely used models for information retrieval is a vector space model (VSM)[SM83], where documents are modeled as vectors in multi-dimensional term space. The natural language processing (NLP) and VSM have been used to represent the text document. In VSM, each dimension corresponds to a word in the document set. Thus, one needs higher dimensions to represent a document. A number of classification techniques, such as Bayesian methods (BM) [BNJ03], decision trees (DT) [ADW94], k-nearest neighbor (KNN) [MGW92], and support vector machines (SVM) [Joa98] have been used to the vector space representation of text documents. Although considerable efforts have been made to extract information from unstructured textual data using text clustering, classification, and summarization, relatively few attempts have been carried out to investigate the excessive topic activities using e-mail content. Pattern recognition and probabilistic modeling on unstructured records, therefore, would be very useful for further research to study fraudulent activities owing to excessive activities in some communication networks.

4.2 Content Analysis and Anomaly

Recently, Priebe et al [PPM$^+$10] reported a model by combining the graph features and e-mail content to investigate the anomaly. They estimated a local topic parameter, $\hat{\theta}_t(v)$ for each vertex, v, and time, t. This parameter represents the proportion of messages in the local region $G_t(v)$. For each vertex and time, the locality region, $G_t(v)$, is defined as:

$$G_t(v) = \Omega\left(N_1[v; G_t]; G_t\right). \tag{4.1}$$

They defined the test statistic as:

$$T_t^c = \Sigma_v I\left(\arg \max_k \hat{\theta}_t(v) \neq arg \max_k \hat{\theta}_{t-1}(v)\right), \tag{4.2}$$

where T_t^c is the number of vertices or nodes that experience a change in the main topic between the two consecutive time points, $t-1$ and t. The large values of T signifies anomaly, suggesting an excessive number of vertices seeing a change in the main topic in the neighborhood at time t.

Here I devise a different approach by combining the probabilistic topic model [SG07], latent Dirichtlet allocation (LDA) algorithm [BNJ03], and the scan statistics. In this case, the locality statistic is the topic proportion, $\hat{\theta}_t$, obtained from LDA, and the maximum of the topic proportion is the scan statistic, which can be written as:

$$S_t = \max_k \hat{\theta}_t(k), \tag{4.3}$$

where k is the topic obtained from LDA. The text processing, LDA and other dimension reduction methods have been discussed below. I then apply temporal scan statistic to obtain excessive topic activities.

4.3 Documents Preprocessing

The content of an email may contain a current message, previous message and a signature block. The signature block in the current inter-organization emails consists of name, title, company, website, telephone number, email address and famous quotes. The previous message is called parent email usually attached with Original Message or Forwarded by. Signature block and previous message have been removed from the current message. Duplicate emails from the dataset have been removed.

The unstructured textual data has been preprocessed to reduce the size of the lexicon

and to preserve the semantic content. It generally consists of two steps, denoising and stemming. The denoising process removes stop words, such as the, for, and, of, are, but, an, by, etc. One can also use own stop words, which are corpus dependent. A commonly used measure for identifying the stop words is to use the term frequency-inverse document frequency (TFIDF) [SM83], [Hoh08], which is defined as:

$$\left.\begin{aligned} w_{ij} &= TF_{ij}\ \log(IDF_j), \\ IDF_j &= \tfrac{D}{b_j}, \end{aligned}\right\} \tag{4.4}$$

where TF_{ij} is the term frequency, which is the number of times word j appears in a document i, IDF_j is the inverse document frequency of word j in the corpus, D is the total number of documents in the corpus, and b_j is the number of documents that have word j. A stemming is a process that removes the suffix from words [Lov68]. Thus, by denoising and stemming, the length of the lexicon is reduced and at the same time, the semantic content of the lexicon is preserved. After denoising and stemming, and removing punctuation, numbers, common stop words and own stop words, we obtain a histogram of words presented in Fig. 4.1.

4.4 The Term Document Matrix

The VSM is used to obtain a term document matrix (TDM), which is a $t \times d$ matrix, where t is the number of terms and d is the number of documents in the pre-processed lexicon, and rows and columns in the TDM represent terms and documents, respectively. The TDM is usually a sparse matrix and the matrix entries is the term frequency count. In the present case, the initial corpus consists of a list of 52 documents and 14883 words around the primary cluster obtained from the scan statistics using Poisson count from June 2003 to June 2004 (see Chapter 1). After denoising, stemming, removing the header, footer text, white space, stop words and own stop words, the preprocessed lexicon consists of 52

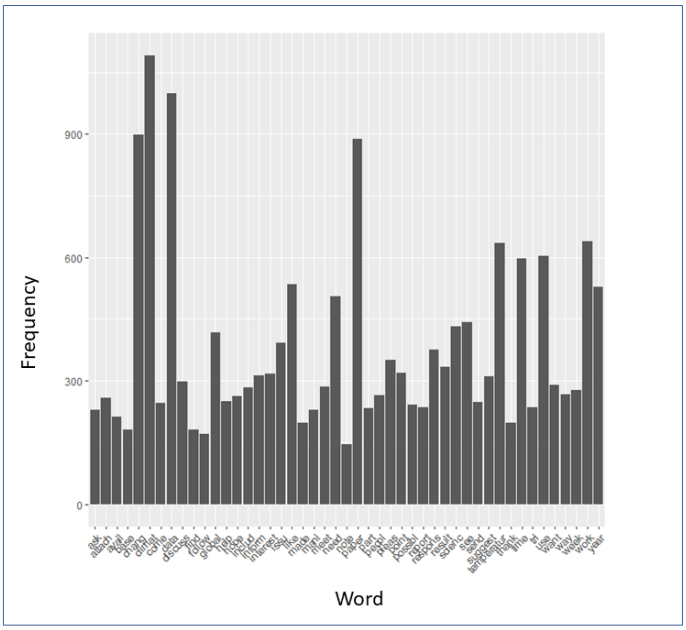


Figure 4.1: Histogram showing frequency of words after denoising and stemming.

documents, $d_1, d_2, d_3, ...d_{52}$, and 10315 words around the primary cluster obtained from the two-step scan process in chapter 1. Thus, the TDM turns out to be a 10315×52 matrix. A partial document term matrix (DTM) (see Table 4.1) shows the frequency of different terms in documents d_1 to d_9.

4.5 Document Similarity

The similarity between documents, $\text{sim}(d_j, d_k)$, is measured by distance based on the angular separation between documents, which is estimated by the cosine of the angle between documents [SM83]. Let d_j and d_k be the two document vectors. The cosine of the angle between d_j and d_k is given by:

$$\text{sim}(d_j, d_k) = \cos(d_j, d_k) = \frac{d_j.d_k}{|d_j||d_k|}. \tag{4.5}$$

Table 4.1: A partial document term matrix for e-mail content from June 2003 to June 2004 around the primary cluster obtained from scan statistics showing the frequency of words in documents.

Docs	abl	academ	accommod	accompani	accord	account	accur	achiev
Jun.W1	5	1	1	1	1	1	1	1
Jun.W2	2	0	0	0	0	2	1	0
Jun.W3	1	1	0	0	3	1	0	0
Jun.W4	3	0	0	0	4	4	0	1
July.W1	3	0	0	0	3	2	0	0
July.W2	0	1	0	1	0	1	0	0
July.W3	1	0	1	0	3	1	0	0
July.W4	5	0	0	1	1	2	1	0
Aug.W1	2	3	0	0	0	1	0	1

where $|d_j|$ and $|d_k|$ are the L_2 norm of the document vectors d_j and d_k. One can see that large values of this measure imply small angular separation between vector documents d_j and d_k, which indicate that the documents d_j and d_k are close to each other. On the other hand, smaller values represent large angular separation between the documents d_j and d_k. Documents d_5 and d_7 with cosine similarity (0.5), documents d_1 and d_4 with cosine similarity (0.48), and documents d_1 and d_6 with the cosine similarity (0.39) are somewhat similar. This measure is used in the multidimensional scaling and the hierarchical agglomerative clustering processes.

4.5.1 Multidimensional Scaling (MDS)

A dimensionality reduction technique, known as multidimensional scaling which uses a spectral decomposition of the dissimilarity matrix [Hoh08], has been used. The objective

of multidimensional scaling is to find a pattern of proximities (i.e. similarities or distances) among a collection of objects. The other widely used dimension reduction technique proposed by Hofmann [Hof99] is probabilistic latent semantic indexing (LSI).

The steps to form the 52×52 dissimilarity matrix from the 10315×52 TDM, $\boldsymbol{A}$, are given below [Hoh08]:

(i) Construct a diagonal matrix, $\boldsymbol{L}$, with diagonal entries L_{ii}, such that $L_{ii} = \log(IDF_i)$, where IDF_i is the inverse document frequency for word, i, and $IDF_i = \frac{D}{b_i}$. Here D is the number of documents and b_i is the number of documents that contain word i.

(ii) Obtain a matrix $\boldsymbol{V}$ such that $\boldsymbol{V} = \boldsymbol{A}^T\boldsymbol{L}$.

(iii) Create the 52×52 similarity matrix, $\boldsymbol{S}$, which is defined as:

$$S_{ij} = \frac{V_i V_j}{|V_i||V_j|}, \tag{4.6}$$

where V_i is the i^{th} row of $\boldsymbol{V}$. The S_{ij} is the cosine of angel between the vector documents i and j. The matrix, $\boldsymbol{S}$, is transformed to a dissimilarity matrix, $\boldsymbol{D}$ by:

$$D_{ij} = \begin{cases} 0, & if \;\; i = j, \\ \max(\boldsymbol{S}) - S_{ij}, & if \;\; i \neq j. \end{cases}$$

Therefore, the dissimilarity matrix, $\boldsymbol{D}$, is a square matrix, where the diagonal elements are

zero, and the off diagonal elements are distances, and is defined as [Gen02]:

$$\boldsymbol{D} = \begin{bmatrix} 0 & d_{12} & d_{13} & d_{14} & . & . & . & d_{1n} \\ d_{21} & 0 & d_{23} & d_{24} & . & . & . & d_{2n} \\ d_{31} & d_{32} & 0 & d_{34} & . & . & . & d_{3n} \\ . & . & . & . & . & . & . & . \\ . & . & . & . & . & . & . & . \\ d_{n1} & d_{n2} & d_{n3} & d_{n4} & . & . & . & 0 \end{bmatrix}.$$

The multi-dimensional scaling is applied on $\boldsymbol{D}$.

4.5.2 Singular Value Decomposition (SVD)

The singular value decomposition of the TDM, $\boldsymbol{A}_{t\times d}$, is written as [MS02]

$$\boldsymbol{A}_{t\times d} = \boldsymbol{T}_{t\times n}\boldsymbol{S}_{n\times n}(\boldsymbol{D}_{d\times n})^T, \tag{4.7}$$

where $\boldsymbol{T}$ is a $(t \times n)$ matrix whose columns is left singular vectors, $\boldsymbol{S}$ is a diagonal $n \times n$ matrix of singular values of $\boldsymbol{A}$, and $\boldsymbol{D}^T$ is the transpose of matrix $\boldsymbol{D}$ whose columns are right singular vectors. The $\boldsymbol{T}$ and $\boldsymbol{D}$ matrices are orthonormal such that $\boldsymbol{T}^T\boldsymbol{T} = \boldsymbol{I}$ and $\boldsymbol{D}^T\boldsymbol{D} = \boldsymbol{I}$, where $\boldsymbol{I}$ is the identity matrix. The diagonal elements of $\boldsymbol{S}$ are all positive and are in the decreasing order. The column vectors of $\boldsymbol{T}$ span the document space, while the column vectors of $\boldsymbol{D}$ span the term space [Sol08]. In a lower dimensional space, the least square approximation of $\boldsymbol{A}$ is estimated as:

$$\hat{\boldsymbol{A}} = \boldsymbol{T}_{t\times k}\boldsymbol{S}_{k\times k}(\boldsymbol{D}_{d\times k})^T, \tag{4.8}$$

such that $||A - \hat{A}||_2$ is minimized. Here $\boldsymbol{T}$, $\boldsymbol{S}$ and $\boldsymbol{D}$ are estimated based on k smaller than the full rank (n columns) [MS02].

4.6 Latent Dirichtlet Allocation Method (LDA)

This method is an unsupervised learning technique, which is used to obtain probabilistic topic models across a large collection of documents. Such method has been applied in various fields, such as Artificial Intelligence, Bioinformatics, Geography and Political Science. It has been considered as the most effective generative probabilistic methods for topic modeling. Here the latent Dirichtlet allocation (LDA) method [BNJ03] has been employed to obtain topics from the content of e-mails received around the most likely (primary) cluster of e-mail data, discussed in chapters 2 and 3. The LDA method has been briefly described below.

Let M be the number of documents in the corpus, D, where $D = \{\mathbf{w_1}, \mathbf{w_2}, ..., \mathbf{w_M}\}$. Here each document is a vector of N words $\mathbf{w} = (w_1, w_2, ..., w_N)$, and is represented as a vector space. In the vector space model, a word from a vocabulary indexed by {1,2,...,V} is represented as unit basis vector. The v^{th} word in the vocabulary is, in fact, a V-vector w defined by [BNJ03]:

$$\left.\begin{array}{ll} w^v & = 1, \\ w^u & = 0, \quad for \;\; u \neq v. \end{array}\right\} \tag{4.9}$$

For each document w:

a. Draw $N \sim$ Poisson(ξ),

b. Draw topic proportion $\boldsymbol{\theta} \sim$ Dir(α),

c. For each of the N words w_n: i) Draw a topic assignment $z_n \sim$ Multinomial($\boldsymbol{\theta}$), ii) Draw w_n, the n^{th} word, from $p(w_n|z_n, \beta)$, a multinomial probability conditioned on the topic assignment, z_n.

In Dirichlet distribution, a k-dimensional $\boldsymbol{\theta}$ is enclosed by (k-1) simplex as $\theta_i \geq 0$ and

$\sum_{i=1}^{K} \theta_i = 1$. The probability density of $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$ is

$$f(\boldsymbol{\theta}|\alpha) = \frac{\Gamma(\sum_{i-1}^{K} \alpha_i)}{\prod_{i=1}^{K} \Gamma(\alpha_i)} \prod_{i=1}^{K} \theta_i^{\alpha_i - 1}, \tag{4.10}$$

where $\boldsymbol{\theta} \in (K-1)$ simplex. The Dirichlet distribution is the conjugate prior to multinomial distribution. The posterior distribution of the latent variables, $\boldsymbol{\theta} = (\theta_1, ..., \theta_k)$, given the document is written as:

$$p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{w}, \alpha, \beta) = \frac{p(\theta, \mathbf{z}, \mathbf{w}|\alpha, \beta)}{p(\mathbf{w}|\alpha, \beta)}. \tag{4.11}$$

The joint distribution of a topic mixture $\boldsymbol{\theta}$, a set of N topics $\boldsymbol{z}$, and a set of N words $\boldsymbol{w}$, given that the parameters α and β are given by:

$$p(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{w}|\alpha, \beta) = p(\boldsymbol{\theta}|\alpha) \prod_{n=1}^{N} p(z_n|\boldsymbol{\theta}) p(w_n|z_n, \beta). \tag{4.12}$$

The marginal distribution of a document:

$$p(\boldsymbol{w}|\alpha, \beta) = \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N} \Sigma_{Z_n} [p(z_n|\theta) p(w_n|z_n, \beta] \right) d\theta. \tag{4.13}$$

The probability of a corpus is written as:

$$p(D|\alpha, \beta) = \prod_{d=1}^{M} \int p(\theta_d|\alpha) \left(\prod_{n=1}^{N} \Sigma_{z_n} [p(z_n|\theta_d) p(w_{dn}|Z_{dn}, \beta] \right) d\theta_d. \tag{4.14}$$

4.6.1 Parameters estimation: Gibbs sampling

The Gibbs sampling, which is the most commonly used Monte Carlo Markov Chain (MCMC) algorithm, is applied when the posterior distribution cannot be directly estimated. In order

to estimate the posterior distribution, $g(\boldsymbol{\theta}|\mathbf{x}) = g(\theta_1, \theta_2, ..., \theta_p|\mathbf{x})$, the Gibbs sampling is used. The algorithm of the Gibbs sampling is given below [HTF11], [CG92].

a. Initialize $\left(\theta_1^{(0)}, \theta_2^{(0)}, ..., \theta_p^{(0)}\right)$,

b. Compute $\left(\theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_p^{(j)}\right)$. For j = 1 to M:

(1) Sample, $\theta_1^{(j)} \sim g\left(\theta_1|\theta_2^{(j-1)}, ..., \theta_p^{(j-1)}, \mathbf{x}\right)$,

(2) Sample, $\theta_2^{(j)} \sim g\left(\theta_2|\theta_1^{(j)}, \theta_3^{(j-1)}..., \theta_p^{(j)}, \mathbf{x}\right)$,

...

(i) Sample, $\theta_i^{(j)} \sim g\left(\theta_i|\theta_1^{(j)}, ..., \theta_{(i-1)}^{(j)}, \theta_{(i+1)}^{(j-1)}, ..., \theta_p^{(j-1)}, \mathbf{x}\right)$,

...

(p) Sample, $\theta_p^{(j)} \sim g\left(\theta_p|\theta_1^{(j)}, ..., \theta_{(p-1)}^{(j)}, \mathbf{x}\right)$.

where $\boldsymbol{\theta}^{(j)} = \left(\theta_1^{(j)}, \theta_2^{(j)}, ..., \theta_p^{(j)}\right)^T$ converges in distribution to $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)^T \sim g(\theta|x) = g(\theta_1, \theta_2, ..., \theta_p|\mathbf{x})$.

4.7 Evolution of Topics Across Time Using LDA

The LDA model has been applied to obtain topics from the e-mail content. This model transforms the document-term matrix of dimension $(n \times m)$ into two matrices of the lower dimensions, $\boldsymbol{D}_1$ and $\boldsymbol{D}_2$. Here the matrix, $\boldsymbol{D}_1$, is termed as the document-topic matrix with dimension $(n \times k)$, and the matrix, $\boldsymbol{D}_2$, is called the topic-term matrix with dimension $(k \times m)$, where n is the number of documents, k is the number of topics and m is the number of terms. In the present case the DTM is (52×10315) matrix. Here $k = 19$ is initially chosen, and the partial D_1 matrix is given in Table 4.2. However, choosing the optimal number of topics in LDA can be computationally complex and tedious. Recently, the log

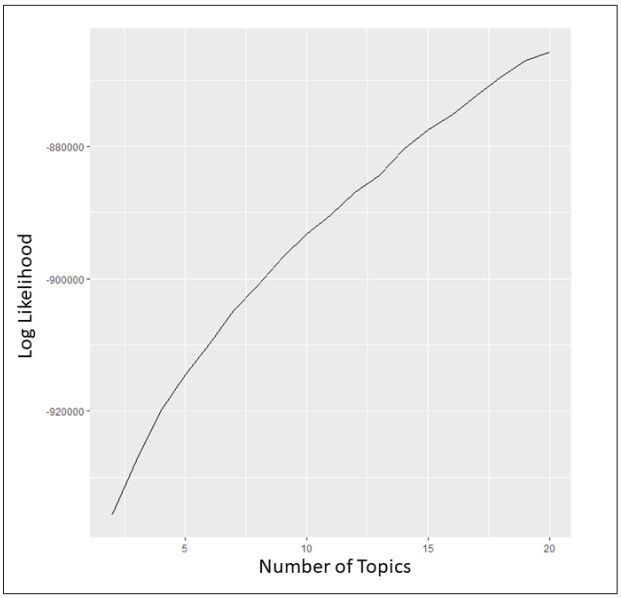


Figure 4.2: Model selection using the log likelihood for the number of topics showing that it does not converge to a global maximum with the increase of number of topics.

likelihood has usually been used to select the number of topics [GS04]. However, in the present case, the log likelihood increases with topics, as shown in Fig.4.2, suggesting that the log likelihood does not converge to a global maximum, as k increases. To overcome this condition, the information criterion techniques, AIC and BIC, have been used to select the optimal number of topics. Fig. 4.3 shows the AIC becomes the minimum, when the number of topics, $k = 3$, and the BIC is minimum when the number of topics equals 2. The number of optimum topics turned out to be three, based on the minimum AIC that gives the best fit to these data.

Recently, a heauristic approach based on analysis of variation of statistical perplexity has been applied to determine the number of topics [ZCP$^+$15]. In fact, this method measures the effectiveness of the statistical model fitted to the data set. To verify the number of topics obtained from the information criterion technique, the dimensionality reduction techniques, such as MDS and SVD have been applied to the document-topic matrix, $\boldsymbol{D}_1$.

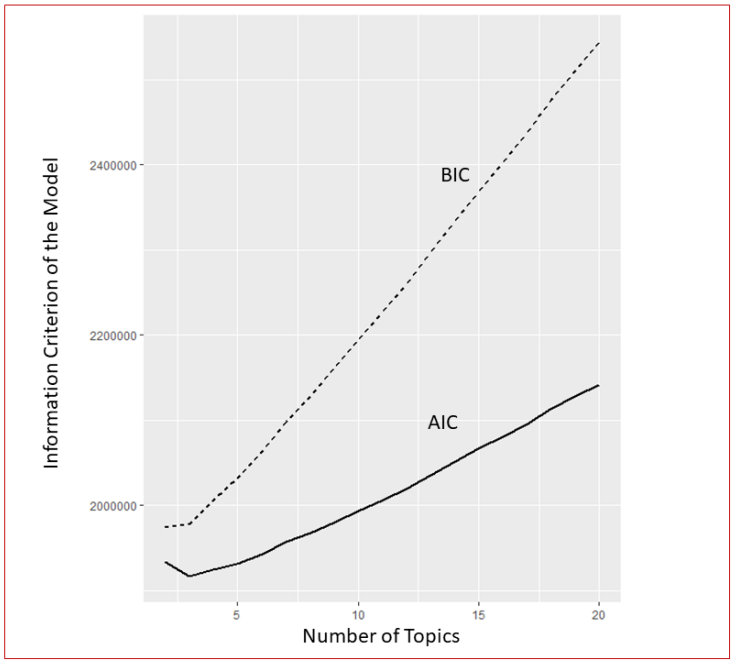


Figure 4.3: Model selection using the information criteria for the number of topics.

A reasonable drop can be observed after three dimensions (see 4.4 (b)), suggesting that the three dimensions are sufficient. The inter-topic distance plot has been presented using MDS in Fig. 4.4(a). Here topics are projected onto the first two dimensions, and each number represents a topic. Note that a number of topics on the right are overlapped, and from this analysis, three non-overlapping topics are obtained.

Similarly, SVD has been applied to the document-topic matrix, $\boldsymbol{D}_1$ and the analysis are presented in Fig. 4.5. Note that the topics are projected onto the first two singular vectors (see Fig. 4.5(a)). The majority of topics, which are overlapped, are at the lower right part of the figure. The largest dimension is 19. The variance vs. the singular vector (see Fig. 4.5(b)) shows that the three dimensions are sufficient to adequately fit the documents into the Euclidean space. The first two singular vectors explain $\approx 82\%$ of the variance. The partial estimated $\boldsymbol{T}$, $\boldsymbol{S}$ and $\boldsymbol{D}$ matrices are shown in Tables 4.3, 4.4 and 4.5, respectively. Note that the $\boldsymbol{S}$ matrix is diagonal with non-negative values in the decreasing order. One

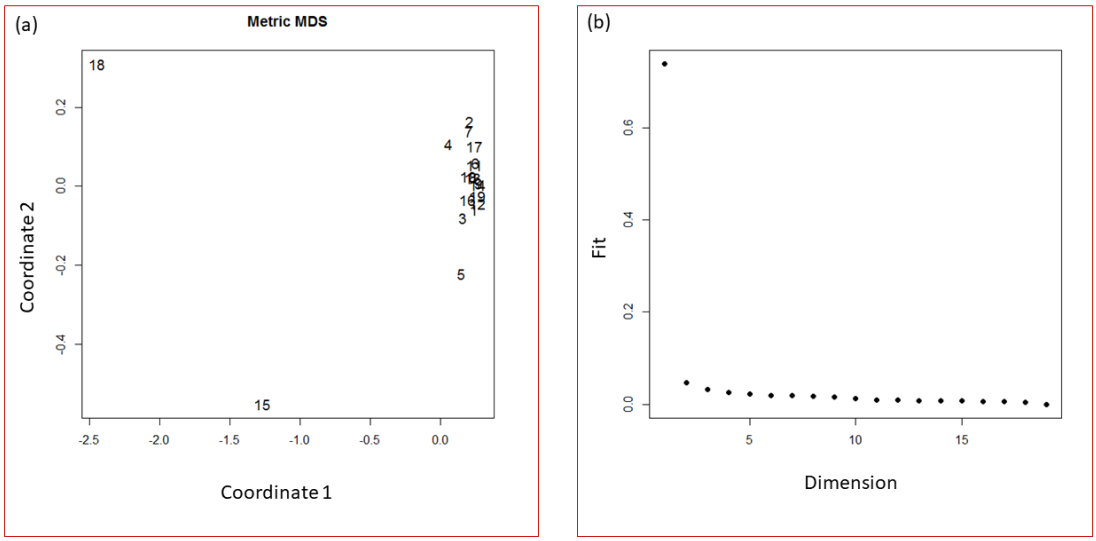


Figure 4.4: (a,b) Multidimensional scaling for the original 52 X 10315 DTM, showing 3 major dimensions. The largest dimension is 19.

could observe that both dimensionality reduction techniques, SVD and MDS, reveal three-major similar components, which is consistent with the results obtained from LDA using the AIC criterion.

Three topics are listed in Table 4.6, showing the top six terms in topics 1 through 3. The topic 1 is associated with a paper, data used, review and response, and ID5, the topic 2 is related to the research work on global weather change, and the topic 3 is about the data reflecting a temperature increase and weather change with time. Topic probabilities of three topics for 52 documents are given in Tables 4.7 and 4.8. One could observe that the largest probabilities of the topic 1 are 0.8333, 0.6331 and 0.6488, respectively, for documents 20, 21 and 22. The e-mail content of week 1 is represented in the LDA analysis as document 1, and the content of week 2 is document 2, and so on. Further, one can plot the evolution of topics across time, as the documents are time stamped. Fig. 4.6 (a) shows the proportion of three topics over 52 weeks, exhibiting the major topic is the topic 1 at week 20. The

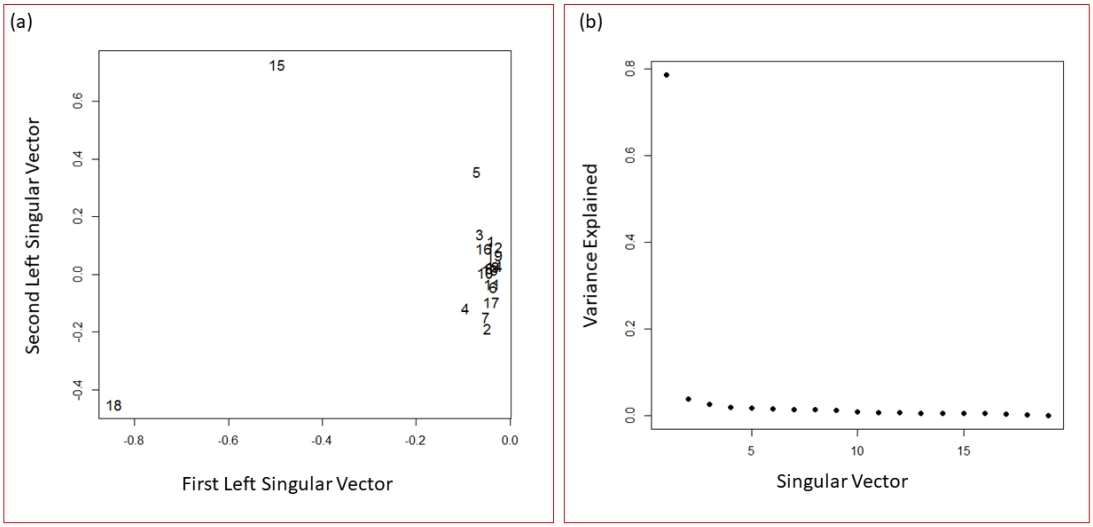


Figure 4.5: (a,b) Singular value decomposition for the original 52 x 10315 DTM, showing 3 dimensions.

topics are shown separately in Fig. 4.6(b).

4.8 Clustering

Two clustering methods, such as hierarchical clustering and K-means clustering, have been applied [MC85] to the document-topic matrix to group 52 documents. For hierarchical clustering, the agglomerative (bottom-up) hierarchical clustering method [MS02] has been used. In this method, initially each object is assigned as a separate cluster, and a new cluster is formed in each step by merging two clusters based on the similarity between the clusters. Three types of similarity functions, such as single-link, complete-link, and group-average are used in selecting clusters to merge. Here, a single - link function is used to the document-topic matrix to cluster 52 documents. The dendrogram is given in Fig. 4.7, showing three clusters that consist of a large cluster having 42 documents and two small clusters with two and eight documents.

Table 4.2: The partial document-topic matrix.

Document	Topic-1	Topic-2	Topic-3	Topic-4	Topic-5	Topic-6
1	0.0394170	0.0015138	0.0162988	0.1235568	0.0009762	0.0031267
2	0.0113448	0.0092984	0.0024772	0.3960651	0.0038414	0.0086163
3	0.0015130	0.0063491	0.0114539	0.3360106	0.0044684	0.0052744
4	0.0033900	0.0008414	0.0059387	0.0552899	0.0050119	0.0013048
5	0.0106559	0.0067150	0.0200155	0.1402126	0.0145968	0.0037593
6	0.0126032	0.0052728	0.0119924	0.0877407	0.0095489	0.0077162
7	0.0150977	0.0093834	0.0093834	0.0120501	0.0318596	0.0029072
8	0.2243859	0.0020768	0.0087188	0.0040303	0.0323562	0.0024675
9	0.1659811	0.0030672	0.0083652	0.0030672	0.0348553	0.0050540
10	0.0298719	0.0388809	0.0253674	0.0208629	0.0253674	0.0208629

Table 4.3: The partial $\boldsymbol{T}_{52\times19}$ matrix of e-mail content.

	v1	v2	v3	v4	v5
1	-0.16184732	0.077103136	-0.18295134	-0.0074151010	-0.019091389
2	-0.13737595	-0.270722214	-0.43221408	-0.3703561140	0.052869896
3	-0.13777215	-0.061069179	-0.42583607	-0.2518886292	0.015319542
4	-0.13279061	0.021356232	-0.08141087	0.0097054464	-0.021760463
5	-0.13523075	0.137339256	-0.20861933	-0.0157525753	-0.071209642
6	-0.14248863	0.145744288	-0.15188913	0.0355148942	-0.069943338
7	-0.11550435	0.092063276	0.01841244	0.1088373227	-0.231220370
8	-0.13698574	0.242338464	-0.04709021	0.0943636136	-0.081727403
9	-0.14366957	0.049399435	0.03671274	0.0333295796	-0.044148842
10	-0.12355908	0.148173944	-0.02858152	0.0677259471	-0.034001568

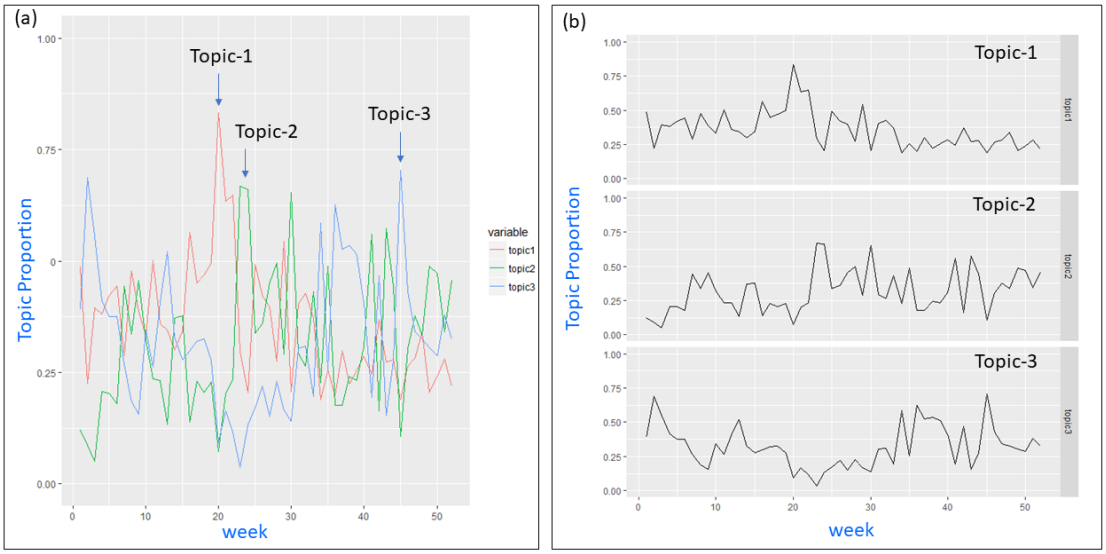


Figure 4.6: (a) Topic proportion of all three topics for the 52 week period around the primary cluster using scan statistic model. (b) Topic proportion plotted separately with time.

Table 4.4: The partial $\boldsymbol{S}_{19\times 19}$ matrix of e-mail content.

	v1	v2	v3	v4	v5	v6	v7
1	3.338276	0.00000	0.00000	0.00000	0.00000	0.00000	0.00000
2	0.0000	0.7431123	0.00000	0.00000	0.00000	0.00000	0.00000
3	0.0000	0.00000	0.6079248	0.00000	0.00000	0.00000	0.00000
4	0.0000	0.00000	0.00000	0.5386343	0.00000	0.00000	0.00000
5	0.0000	0.00000	0.00000	0.00000	0.4988096	0.00000	0.00000
6	0.0000	0.00000	0.00000	0.00000	0.00000	0.4680062	0.00000
7	0.0000	0.00000	0.00000	0.00000	0.00000	0.00000	0.4589441

Table 4.5: The partial $(\boldsymbol{D}_{19\times 19})^T$ matrix of e-mail content.

	v1	v2	v3	v4	v5
1	-0.04045566	0.11565104	-0.015159817	0.047869952	-0.08160274
2	-0.04792393	-0.18657792	0.106860269	-0.027032479	0.01441465
3	-0.06423974	0.13939360	-0.012146631	0.297756953	0.78853319
4	-0.09435856	-0.11601124	-0.798778568	-0.493581174	0.02233819
5	-0.06998048	0.35691651	0.455700782	-0.739323117	0.14316247

The K-means clustering is a non-hierarchical clustering technique that divides the data into K clusters such that the within-cluster variation is minimized[Gen02]. The objective is to determine the optimal number of K using Calinski-Harabasz (CH) index. The null hypothesis vs. the alternative hypothesis [CH74a], [DPL$^+$15] is $H_0 : K = 1$ vs. $H_1 : K > 1$. A pseudo-F statistic known as the CH index is defined as [CH74a], [Gen02] :

$$\text{CH index} = \frac{\frac{b}{(k-1)}}{\frac{w}{(n-k)}}, \tag{4.15}$$

where b is the between cluster sum-of-square, w is within cluster sum-of-square, k is the number of clusters and n is the number of observations. Here, k is chosen such that the CH index is a global or a local maximum, or grows rapidly. The number of the clusters is turned out to be 3 based on a local maximum or a rapid increase that corresponds to the CH index of 141.58 (p-value < 0.0001). Fig. 4.8) reveals the similar structure in the reduced dimensions, suggesting that there are three clusters with black dots indicating the centroids of clusters. From both agglomerative and K-means clustering, three major clusters have been obtained.

Table 4.6: Three major topics with top six terms obtained using the LDA from e-mail content around the most likely cluster.

Topic 1	Topic 2	Topic 3
data	work	temperatur
paper	global	weather
use	weather	change
review	new	warm
ID-5	chang	data
respons	research	time

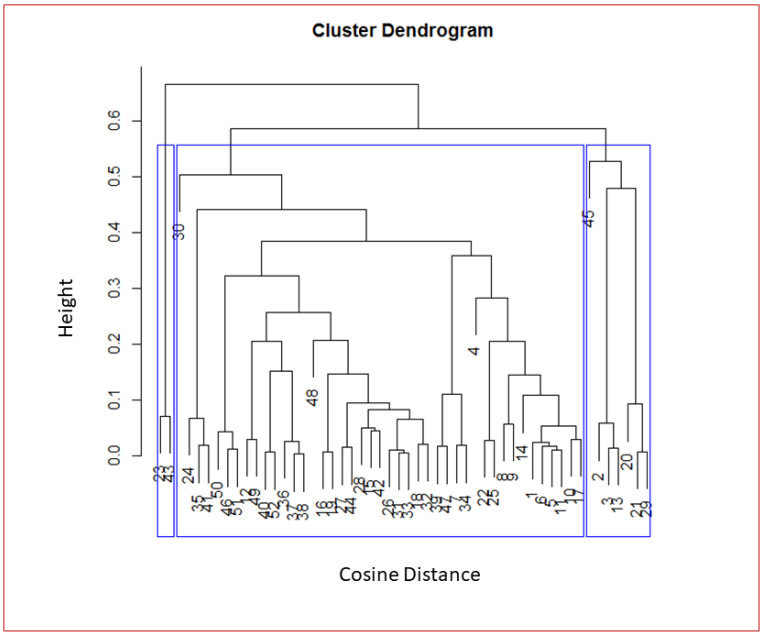


Figure 4.7: Hierarchical agglomerative clustering on the document term matrix, showing the dendrogram with three major clusters using the single link method.

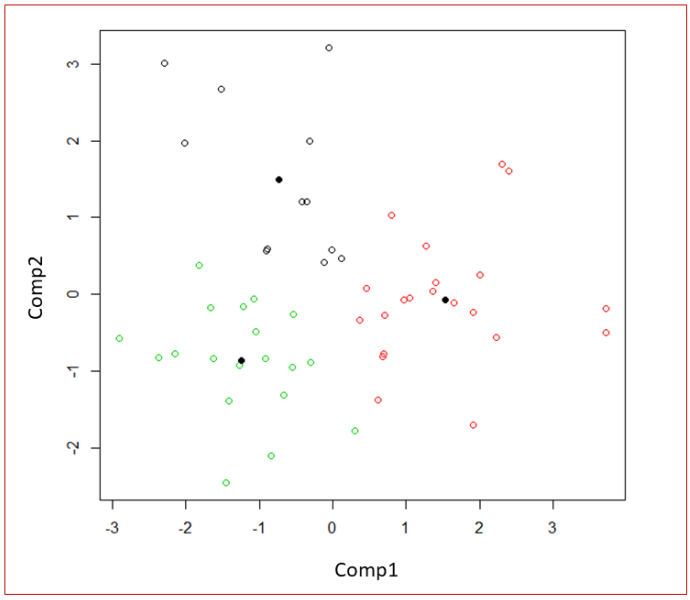


Figure 4.8: K-means clustering showing three clusters.

4.9 Scan Statistics on Topic Proportions Using Normal Distribution

To carry out scan statistics for the time series of the maximum proportions, $p_t \in (0, 1)$, the logistic transformation [Wal87] has been applied to the data. This is given by:

$$y_t = \text{logit}\, p_t = \log\left(\frac{p_t}{1 - p_t}\right). \tag{4.16}$$

This transformation stabilizes the variance, and the transformed data become normally distributed [BVAMF07]. Figs. 4.9(a) shows the maximum topic proportion over 52 weeks around the most likely cluster obtained from the two-step scan process. The logistic transformation of the maximum proportion is shown in Fig. 4.9(b). The normal Q-Q plot shows that the transformed maximum topic proportion is approximately normally distributed (see

Fig. 4.10). The scan statistics have been applied here to investigate if there exists a critical topic cluster. As discussed in chapter 1, the one-dimensional continuous scan statistic, S_w, is the largest number of points that are observed in any subinterval of [0, T) of length w. Under the null hypothesis, observations $x_1, x_2, .., x_n$ are from the i.i.d normal distribution with mean, μ, and variance, σ^2 . For the alternative hypothesis, there would be a scanning window of width, w, where observations are from an i.i.d normal distribution with mean, μ, and common variance, σ^2, and the observations in the rest of the intervals, [1, t) and [$t+w$, n], are from an i.i.d normal distribution with mean, η, and variance, σ^2 [Kul79], [Jun16]. For testing, the null hypothesis, $H_0 : \mu = \eta$, over the alternative hypothesis, $H_1 : \mu > \eta$, the likelihood under the null hypothesis:

$$L_{H_0} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = (2\pi)^{-\frac{n}{2}} \sigma^{-n} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}. \tag{4.17}$$

The log likelihood function under the null hypothesis is given by:

$$\log L_{H_0}(\mu, \sigma) = -\frac{n}{2} \log(2\pi) - n \log(\sigma) - \sum_{i=1}^{n} \frac{(x_i - \mu)^2}{2\sigma^2}. \tag{4.18}$$

Taking the derivative of the log likelihood with respect to μ and setting to zero, one can get:

$$\frac{\partial \log L_{H_0}}{\partial \mu} = 2 \sum_{i=1}^{n} \frac{(x_i - \mu)}{2\hat{\sigma}^2} = 0. \tag{4.19}$$

Therefore,

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}. \tag{4.20}$$

Similarly, taking the derivative of the log likelihood with respect to σ and setting to zero, one obtains:

$$\frac{\partial \log L_{H_0}}{\partial \sigma} = -\frac{n}{\sigma} + \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{\sigma^3} = 0. \tag{4.21}$$

Therefore,

$$\hat{\sigma^2} = \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{n}. \tag{4.22}$$

The likelihood under the alternate hypothesis:

$$\left.\begin{aligned} L_{H_1} &= \left(\prod_{i\in w} \tfrac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{(x_i-\mu_w)^2}{2\sigma^2}}\right)\left(\prod_{i\neq w} \tfrac{1}{\sqrt{2\pi}\sigma_w} e^{-\frac{(x_i-\eta_w)^2}{2\sigma_w^2}}\right) \\ &= (2\pi)^{-\frac{n}{2}} (\sigma_w)^{-n} e^{-\frac{1}{2\sigma_w^2}\left[\sum_{i\in w}(x_i-\mu_w)^2 + \sum_{i\neq w}(x_i-\eta_w)^2\right]}. \end{aligned}\right\} \tag{4.23}$$

The log likelihood function under the alternative hypothesis is given by:

$$\log L_{H_1} = -\frac{n}{2}\log(2\pi) - n\log(\sigma_w) - \frac{1}{2\sigma_w^2}\left[\sum_{i\in w}(x_i-\mu_w)^2 + \sum_{i\neq w}(x_i-\eta_w)^2\right]. \tag{4.24}$$

By taking the derivative of the log likelihood with respect to σ and setting to zero, one obtains:

$$\frac{\partial \log L}{\partial \sigma_w} = -\frac{n}{\sigma_w} + \frac{1}{\sigma_w^3}\left[\sum_{i\in w}(x_i-\mu_w)^2 + \sum_{i\neq w}(x_i-\eta_w)^2\right] = 0. \tag{4.25}$$

Therefore, the estimated $\hat{\sigma}_w^2$, $\hat{\mu}_w$ and $\hat{\eta}_w$ are given by the following equations.

$$\hat{\sigma}_w^2 = \frac{1}{n}\left[\sum_{i\in w}(x_i-\mu_w)^2 + \sum_{i\neq w}(x_i-\eta_w)^2\right], \tag{4.26}$$

$$\hat{\mu}_w = \frac{1}{n_w} \sum_{i \in w} x_i, \tag{4.27}$$

$$\hat{\eta}_w = \frac{\sum_{i \in w} x_i}{n - n_w}, \tag{4.28}$$

where n_w is the number of observations inside the window, w. The log-likelihood ratio, $\log \Lambda$, is given by:

$$\log \Lambda = n \log(\hat{\sigma}) + \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{2\hat{\sigma}^2} - \frac{n}{2} - n \log(\sqrt{\hat{\sigma}_w^2}). \tag{4.29}$$

The maximum log-likelihood ratio is given by:

$$\max_w(\log \Lambda) = \max_w \left(n \log(\hat{\sigma}) + \sum_{i=1}^{n} \frac{(x_i - \hat{\mu})^2}{2\hat{\sigma}^2} - \frac{n}{2} - n \log(\sqrt{\hat{\sigma}_w^2}) \right). \tag{4.30}$$

Therefore, the most likely cluster can be identified when variance will be minimized.

I now apply a purely temporal scan statistic to detect the primary clusters using the normal model to the stationary time series (See Fig. 4.11) of the transformed maximum topic proportion, as shown in Fig. 4.9. The stationarity of the transformed series has been confirmed by the unit root tests (see Table 4.9). The p-value is obtained using the MC simulation with 1000 replications. A primary cluster from October 22 to November 30 2003 has been identified, which is statistically significant with the maximum LLR = 10.24 and the MC p-value = 0.01. One could see that the estimated the maximum LLR for the primary cluster of topics is greater than the standard Monte Carlo critical value, 7.81, at the 0.0001 level of significance (See Table 4.10). In addition, it is observed that the primary cluster consists of mostly the topic 1. It can, therefore, be concluded that the scan statistic identifies statistically significant excessive topic activities, where mostly the topic 1 is discussed. To investigate the topic proportion, I consider here compositional approach

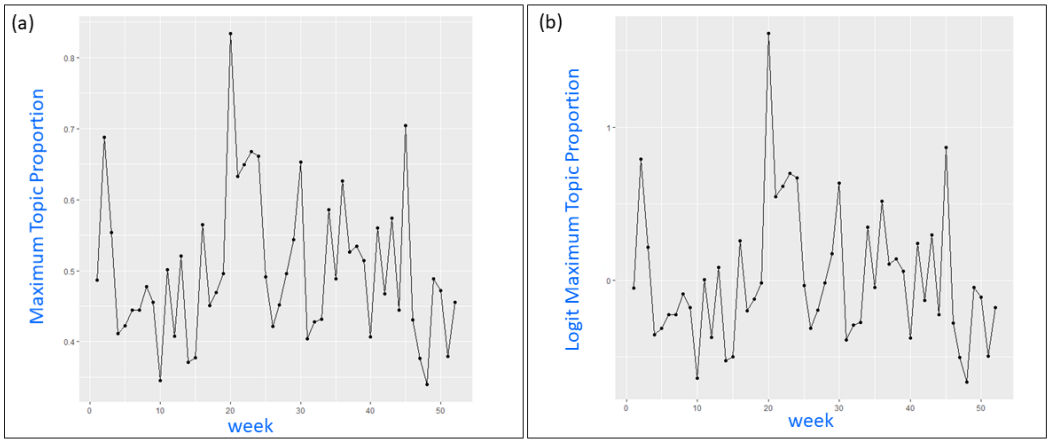


Figure 4.9: (a) The maximum proportion of topics for the 52 week period around the primary cluster using scan statistic model . (b) The logistic transformed maximum proportion of topics for the 52 week period around the primary cluster using scan statistic model.

to the time series of topic 1 proportion [BVAMF07].

4.10 Time Series Models on Topic 1: Compositional ARIMA (C-ARIMA) Model

A process of continuous proportions, $p_t, t = 0 \pm 1, \pm 2, ...,$ is a C-ARMA (p,q) process if for every t [BVAMF07],

$$\text{logit}(p_t) = \phi_1 \text{logit}(p_{t-1}) + ... + \phi_p \text{logit}(p_{t-p}) + \text{logit}(w_t) - \theta_1 \text{logit}(w_{t-1}) - .... - \theta_q \text{logit}(w_{t-q}), \tag{4.31}$$

where $\text{logit}(w_t) \sim N(0, 2\tilde{\sigma}_w^2)$. One can write in terms of odds [BVAMF07]:

$$\text{odds } p_t = (\text{odds } p_{t-1})^{\phi_1} \times ... \times (\text{odds } p_{t-p})^{\phi_p} \times w_t \times (w_{t-1})^{-\theta_1} \times ... \times (w_{t-q})^{-\theta_q}, \tag{4.32}$$

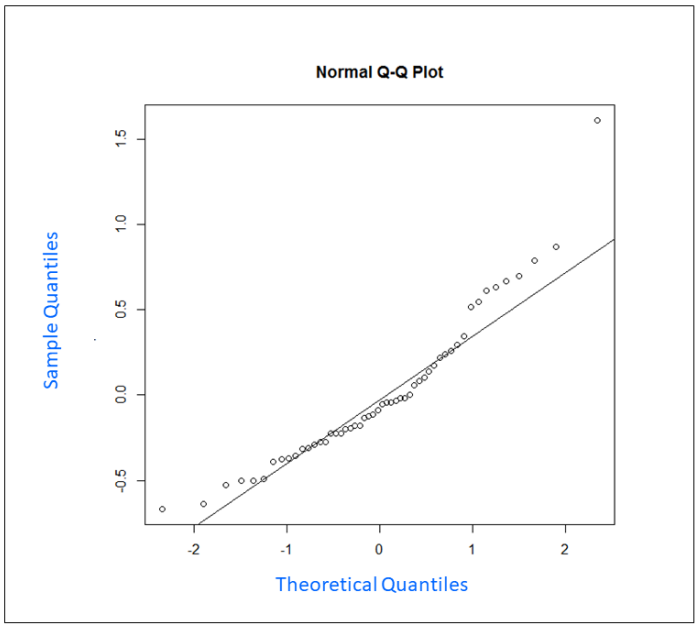


Figure 4.10: A normal Q-Q plot of the logistic transformed maximum proportion of topics showing that the distribution is approximately normal.

where w_t is log normally distributed. Here p_t is a C-ARIMA(p, d, q) process if and only if logit (p_t) is an ARIMA (p, d, q) process. The ARIMA (p, d, q) models are well-known statistical models for a non-stationary time series x_t, and can be mathematically expressed using backshift operator, B, as [SS06]:

$$\phi(B)(1 - B)^d x_t = \theta(B) w_t. \tag{4.33}$$

Fig. 4.12 shows that the logit (topic 1 proportion) has a decreasing trend. Therefore, an ADF regression with a trend has been applied. It is observed that the test statistic of -3.394 is not lower than the critical value at 5% level of significance for the logit series (see Table 4.11), suggesting that the series is not stationary. However, the first difference of the logit(p_t) series is stationary (see Table 4.12). Therefore, the logit(p_t) series is an integrated of order 1, i.e. $I(1)$, and, as a result, the ARIMA $(p, 1, q)$ models would be suitable.

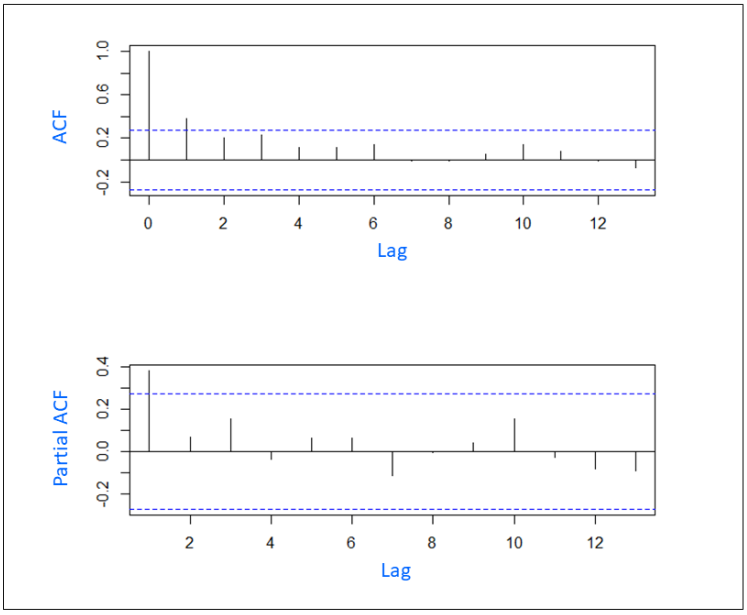


Figure 4.11: Sample ACF and PACF of the logistic transformed maximum proportion of topics series showing that the time series is stationary.

The ARIMA(0,1,1), ARIMA(1,1,0), ARIMA(2,1,0) and ARIMA(1,1,1) models have been fitted to the logarithm of odds series. The optimal model is an ARIMA(0,1,1) based on minimizing the AIC, and the estimated model is (see Table 4.13):

$$\text{logit}(p_t) = -0.731\text{logit}(w_{t-1}) + \text{logit}(w_t), \tag{4.34}$$

where $\text{logit}(w_t) \sim N(0, 2\tilde{\sigma}_w^2)$. The standardized residual in Fig. 4.13(upper panel) at week 20, exceeds 3 times the standard deviations, indicating excessive activities. The sample ACF of standardized residual and the Q-statistic show that the residuals are random (see Fig. 4.13, middle and lower panels, respectively). Also, it is observed from the histogram and the normal Q-Q plot of residuals that they are approximately normal except for an extreme value in the right tail (see Fig. 4.14).

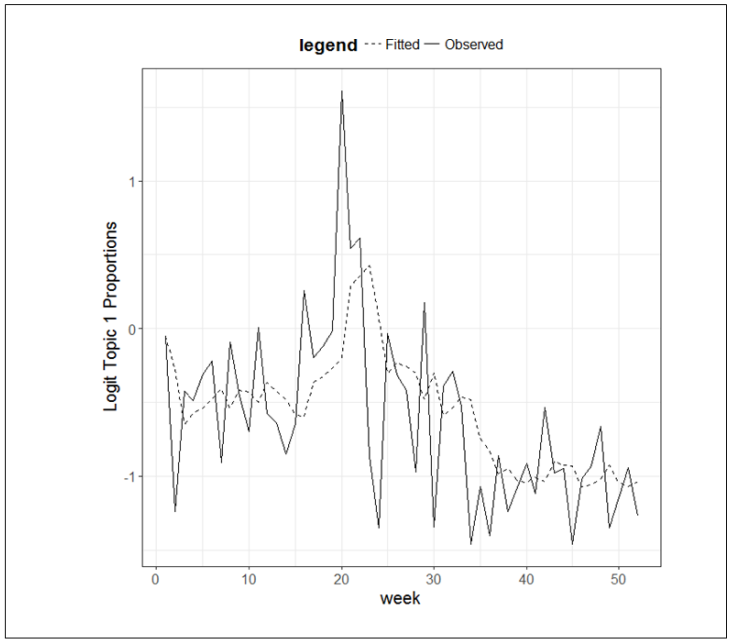


Figure 4.12: The observed and fitted logistic transformed proportion of topic 1 series for the ARIMA(0,1,1) fit to the logistic transformed proportion of topic 1 series.

4.11 Identifying Vertices with Excessive Messages using a Combination of 1-Nearest Neighbor (1NN) and K-Means

Let $\hat{\theta}_t(v)$ be the proportions of messages in the local region $G_t(v) = \Omega(N_1[v; G_t]; G_t)$ for each vertex v at time t, where $\boldsymbol{\theta} = [\theta_1, \theta_2, ..., \theta_k]^T$. After preprocessing, for each vertex and each week, the document term matrix (DTM) from the corpus obtained from unstructured email textual data has been estimated. The DTM has then been split into a training set (70%) and the test set (30%). First, the classical multidimensional scaling has been applied to both the training set and the test set to reduce the dimensionality. The collection of training documents have then been partitioned into K clusters/topics, where $K = 3$, using K-means clustering [HTF11],[JWHT13] and then classified using the 1-nearest neighbor (1NN)[HTF11], [JWHT13] to estimate $\hat{\theta}_t(v)$. Table 4.14 shows the proportion of messages

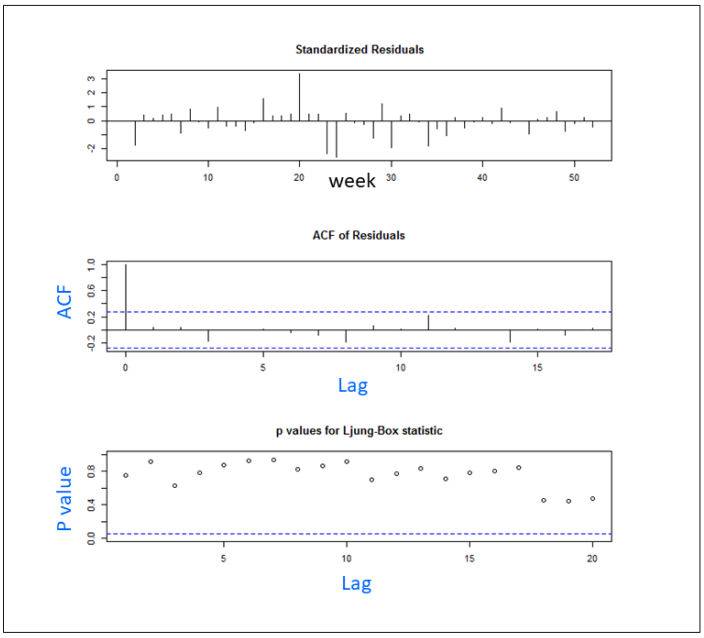


Figure 4.13: Time Plots of the standardized residuals, the ACF of standardized residual and the Q-statistic for the ARIMA(0,1,1) fit to the transformed topic 1.

for topic 1(θ_1), topic 2(θ_2), and topic 3(θ_3) for week 20, 21 and 22. Here θ_1 is the topic 1, which is associated with the data used, review and ID5. This is approximately similar to the topic 1 obtained from LDA model. It is observed that the ID1, ID5, ID7, ID10 and ID20 all have a maximum proportion of messages that are associated with the topic 1 (See Table 4.14). In order to verify this result, the betweenness associated with individuals obtained from the metadata and the maximum topic proportions obtained from the LDA using the textual data have been combined. Fig. 4.15 shows the time plot of the logistic transformed maximum topic proportion, obtained from the LDA (see the bottom of Fig. 4.15) and the time plots of betweenness of ID = 1, 5, 7, 10, 18, 20, 30 (upper panel). One could see that ID5, ID7, ID18, ID20, and ID30 have the maximum betweenness at week 20 and 21 (See Fig. 4.15). The maximum betweenness represents the maximum information flow through nodes or vertices. Therefore, it is apparent from this figure that the topic 1 (see Fig. 4.15,

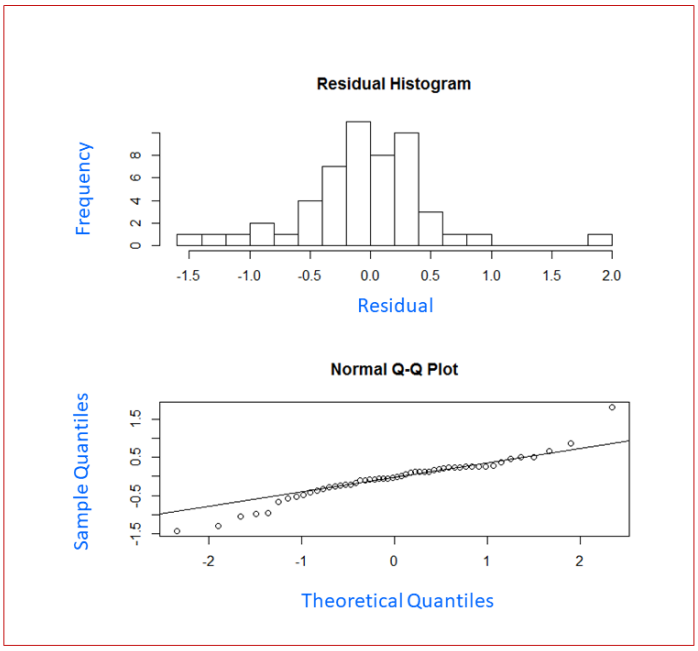


Figure 4.14: Histogram of the residuals (top), and a normal Q-Q plot of the residuals (bottom) for the ARIMA(0,1,1) fit to the logistic transformed proportion of topic 1 series.

is mostly discussed by these individuals.

4.12 Chapter Summary

In this chapter, the scan statistic has been implemented in the unstructured text data of the e-mail content. Such implementation of scan statistics on text data has not been done before. Although considerable efforts have been made to extract information from unstructured text data using text clustering, classification, and summarization, relatively few attempts have been carried out to investigate the excessive activities associated with a particular topic from the dynamic unstructured textual dataset. Pattern recognition and probabilistic modeling on unstructured records, therefore, would be very useful for further research to study fraudulent activities owing to excessive activities in communication networks. Here an unstructured text obtained from e-mail contents around the most likely (primary) cluster,

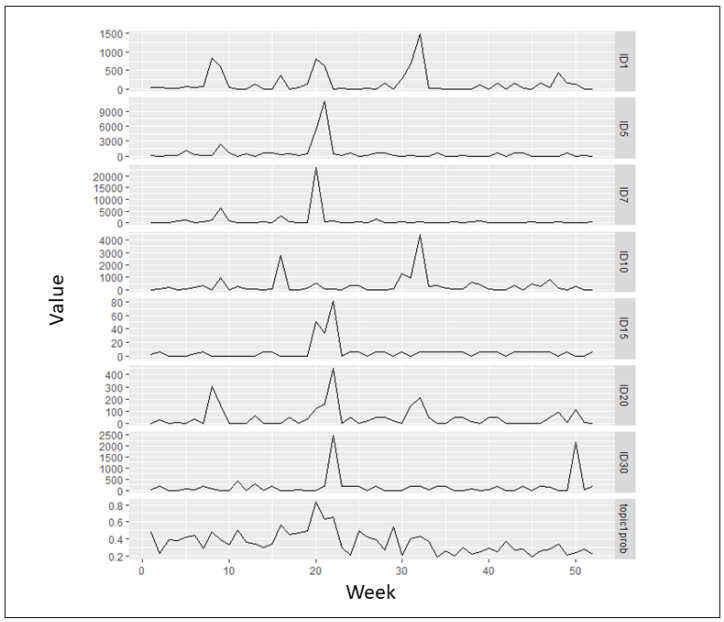


Figure 4.15: Comparison between time plots of betweenness for $k > 2$ for different IDs obtained from metadata, and the maximum topic proportion obtained from LDA using textual data, showing that the excessive topic activity relating to the topic 1, which is associated with ID = 5, 7, 18, 20 and 30 at around week 20.

from June 2003 to June 2004, has been analyzed. The major topics discussed in this period have been identified using text mining algorithms and probabilistic modeling, such as the latent Dirichlet allocation modeling. In the LDA modeling, the e-mail content of week 1 is represented as document 1, and the content of week 2 is document 2, and so on. Therefore, one can plot the evolution of topics across time, as the documents are time stamped. The optimal number of topics has been chosen to be three, based on the minimum AIC that gives the best fit to these data. The scan statistic based on normal distribution is then applied to the logistic transformed maximum topic proportions to obtain the excessive topic activities. A primary cluster from November 01 2003 until November 30 2003 has been identified, which is statistically significant with the maximum LLR = 10.24 and the MC p-value = 0.01.

The estimated maximum LLR of the primary topic cluster is greater than the standard

Monte Carlo critical value, 7.81, at the 0.0001 level of significance. It can be concluded that the scan statistic identifies statistically significant excessive topic activities, where mostly the topic 1 is discussed. In order to identify the individuals that are discussing topic 1, the betweenness associated with individuals from the metadata, and the maximum topic proportions obtained from the LDA using textual data are combined. It can be concluded that the topic 1 is mostly discussed by ID5, ID7, ID18, ID20, and ID30 at week 20 and 21.

Table 4.7: Topic probabilities by document obtained using the LDA method.

	Topic 1	Topic 2	Topic 3
1	0.4870072	0.12114695	0.39184588
2	0.2248749	0.08708504	0.68804002
3	0.3953967	0.05042092	0.55418234
4	0.3805993	0.20752240	0.41187828
5	0.4224959	0.20279146	0.37471264
6	0.4445123	0.18000407	0.37548361
7	0.2878730	0.44406349	0.26806349
8	0.4775672	0.33632871	0.18610406
9	0.3878587	0.45540839	0.15673289
10	0.3318318	0.32282282	0.34534535
11	0.5010635	0.23579459	0.26314190
12	0.3593337	0.23258176	0.40808450
13	0.3452609	0.13375932	0.52097977
14	0.2996117	0.37134682	0.32904149
15	0.3434343	0.37752525	0.27904040
16	0.5643217	0.13795748	0.29772080
17	0.4504685	0.22958501	0.31994645
18	0.4695216	0.20563272	0.32484568
19	0.4957613	0.22821295	0.27602577
20	0.8333829	0.07245083	0.09416625
21	0.6331434	0.20236898	0.16448765
22	0.6487928	0.23581982	0.11538739
23	0.2960980	0.66797526	0.03592672
24	0.2055872	0.66124220	0.13317060
25	0.4913804	0.33741282	0.17120674
26	0.4219797	0.35879763	0.21922267
27	0.3964082	0.45159877	0.15199299
28	0.2744511	0.49600798	0.22954092
29	0.5433990	0.29066375	0.16593727
30	0.2063568	0.65307906	0.14056416

Table 4.8: Topic probabilities by document: Continued.

	Topic 1	Topic 2	Topic 3
31	0.4036814	0.29296519	0.30335338
32	0.4279996	0.26344261	0.30855783
33	0.3720205	0.43192312	0.19605641
34	0.1886532	0.22590102	0.58544577
35	0.2555714	0.48885728	0.25557136
36	0.1977420	0.17560921	0.62664878
37	0.2976594	0.17610977	0.52623083
38	0.2239988	0.24150101	0.53450014
39	0.2536991	0.23175395	0.51454697
40	0.2856978	0.30749712	0.40680507
41	0.2466732	0.55972087	0.19360597
42	0.3687664	0.16404199	0.46719160
43	0.2730443	0.57349278	0.15346288
44	0.2794066	0.44432810	0.27626527
45	0.1883071	0.10742160	0.70427132
46	0.2648585	0.30398696	0.43115459
47	0.2827147	0.37664042	0.34064492
48	0.3394167	0.33512256	0.32546073
49	0.2053294	0.48880170	0.30586889
50	0.2409661	0.47216077	0.28687316
51	0.2805110	0.34058346	0.37890558
52	0.2193131	0.45579955	0.32488739

Table 4.9: Unit root tests on the transformed maximum topic proportion.

Unit root test	Value of test statistics	Lag Parameter	p-value
Augmented Dickey-Fuller Test	-3.7573	1	0.0284
Phillips-Perron Test	-4.9454	3	0.01
KPSS Test	0.43172	1	0.0635

Table 4.10: Temporal clusters of the topic proportion showing the estimated log likelihood ratio (LLR), standard Monte critical values (SMCV) and significance level (SL) obtained using SaTSscan software.

Cluster	Time Frame	LLR	p-value	SMCV (SL)
Primary	11/1/03 - 11/30/03	10.240	0.01	7.81 (0.001)

Table 4.11: Unit root tests for logit (p).

Test Statistics	Value of Test Statistics	CV (1%)	CV (5%)	CV (10%)
τ_3	-3.394	-4.04	-3.45	-3.15
ϕ_2	3.9337	6.50	4.88	4.16
ϕ_3	5.8733	8.73	6.49	5.47

Table 4.12: Unit root tests for the first difference logit (p) series.

Test Statistics	Value of Test Statistics	CV (1%)	CV (5%)	CV (10%)
τ_3	-6.2506	-4.04	-3.45	-3.15
ϕ_2	13.0303	6.50	4.88	4.16
ϕ_3	19.5415	8.73	6.49	5.47

Table 4.13: ARIMA(0,1,1), ARIMA(1,1,0) and ARIMA(1,1,1) model results fitted to the logit of topic 1 proportion series.

Parameter	ARIMA(0,1,1)	SE	ARIMA(1,1,0)	SE	ARIMA(1,1,1)	SE
ar1			-0.4843	0.1255	0.0912	0.1977
ma1	-0.7173	0.1106			-0.7781	0.1326
AIC	86.18		93.38		87.97	

Table 4.14: Proportion of massages obtained from the combination of K-means and nearest neighbor.

Week	ID	$\theta = (\theta_1, \theta_2, \theta_3)$
20	1	(0.481, 0.333, 0.037)
20	5	(0.446, 0.349, 0.059)
20	15	(0.6, 0.2, 0.000)
20	20	(0.533, 0.011, 0.367)
21	5	(0.553, 0.316, 0.079)
21	10	(0.667, 0.333, 0.000)
21	20	(0.524, 0.381, 0.048)
21	30	(0.667, 0.333, 0.000)
22	5	(0.364, 0.364, 0.136)
22	7	(0.667, 0.333, 0.000)

Chapter 5: Conclusions

5.1 Summary of Contributions

Around terabytes of unstructured electronic data are generated every day from twitter networks, scientific collaborations, organizational emails, telephone calls and websites. Fraudulent activities owing to excessive communication in communication networks continue to be a major problem in different organizations. In fact, retrieving information relating to detection of excessive activities is computationally intensive for large data sets. Therefore, one needs useful tools and techniques to analyze such a massive data set and detect anomaly. In a social network, anomalies can occur as a result of abrupt changes in the interactions among a group of individuals. Analyzing the excessive activity in a social network is thus important to understand the fraudulent behavior of individuals in a subregion of a network. The motivation of this research work is to investigate the excessive activities and make inferences in dynamic sub networks. Three major contributions have been presented for to detect anomalies of dynamic networks obtained from inter-organizational emails.

(i) **Implemented Scan Statistics with variable windows, and univeriate time series**: First, a temporal scan statistic was introduced to detect clusters using the maximum likelihood ratio as the test statistic, and betweenness as a locality statistic. Previous studies are mostly based on the fixed and disjoint windows, and on the assumption of short term stationarity of the series under null, which might result in loss of information and error in detecting excessive activities. In addition, the previous model assumed that the subgraphs are disjoint, and normalized the locality statistic twice to eliminate the trend and assumed short-time, near-stationarity for the null model. However, the scan statistics with fixed and disjoint scan window may not be

appropriate because of the occurrence of window overlaps, which may result in loss of information on the time axis. In this research work, I implement scan statistics with overlapping and variable window sizes to detect anomalies of dynamic networks obtained from organizational emails. I employ the maximum likelihood ratio (LLR) as the test statistic to rank the clusters, as the cluster size is not known. Furthermore, I assess the structural stability and apply differencing, seasonal adjustment to make the time series of scan statistics stationary and estimate the p-value using the Monte Carlo (MC) simulation and the extreme value distribution, such as Gumbel, as the exact sampling distribution of scan statistics under the null hypothesis is not known for most of the cases. In addition, as the unstructured data set size becomes larger, the formation of dynamic network structure is computationally intensive. Instead of applying temporal scan statistics for ego sub networks directly, I employ scan statistics of organizational emails with a two-step process, and use the likelihood function to rank the clusters. I initially estimate the maximum log-likelihood ratio (LLR) to obtain a primary cluster of communications (LLR = 82.07, p-value = 0.001) using the Poisson model on email count series, and then extract neighborhood ego subnetworks around the observed primary cluster to obtain more refined cluster (LLR = 644.11, p-value = 0.001) by invoking the graph invariant betweenness as the locality statistic using the binomial model. Furthermore, as an alternative approach, a univariate time series has been built using the graph edit distance (GED) between subgraphs. An autoregressive moving average (ARMA) process is then fitted to the time series, and the anomalies were assessed using residuals obtained from the fitted model and compared with the results obtained from the scan statistics.

(ii) **Developed multivariate time series model**: The second contribution is the development of multivariate time series models, vector autoregressive (VAR) models on the e-mail network obtained from the metadata. This represents the first known presentation of the VAR model to detect excessive activities. As the scan statistics measure the maximum of locality statistics, it can conceal the group of influential people in

the network. To overcome this limitation, the multivariate time series models have been used to identify the most influential node. I fit a VAR(1) model to the multivariate time series of subgraphs for each vertex using the graph edit distance to identify anomalies. This analysis considers multiple time series simultaneously, as the nodes or vertices of the subgraphs are interrelated. The objective of the VAR model is to identify the dynamic relationship between vertices. The excessive activities or anomalies associated with the nodes or vertices in email network have been assessed using residual threshold. One could clearly observe the residual greater than 5 times standard deviations at week 20 for all IDs, suggesting the occurrence of excessive activities at this time point. From this multivariate time series analysis, it can be concluded that the chatter has been initiated by ID = 5, and as k increases, the excessive chatter spread among other IDs. In addition, this analysis clearly demonstrates the social relationship between ID5 with ID1, ID7, ID10 and ID20.

(iii) **Implemented scan statistics on topic models of unstructured text**: The third contribution is the implementation of the scan statistics on topic models of the unstructured text data of e-mail content. Such implementation of scan statistics on text data has not been done before. Although considerable efforts have been made to extract information from unstructured textual data using text clustering, classification, and summarization, relatively few attempts have been carried out to investigate the excessive activities associated with a particular topic from the dynamic unstructured textual data set. Pattern recognition and probabilistic modeling on unstructured records, therefore, would be very useful for further research to study fraudulent activities owing to excessive activities in communication networks. Here I analyze an unstructured textual obtained from e-mail contents around the primary cluster, June 2003 to March 2004, and investigate the major topic discussed in this period using text mining algorithms and probabilistic modeling, such as latent Dirichlet allocation (LDA) modeling. I then use scan statistics to get excessive topic activities. It is observed that the topic-1, which is related to data used in a paper, has the largest

LLR in the neighborhood of the primary cluster using scan statistics.

(iv) One of the scientific challenges of this research includes understanding the distribution of the organizational email subnetworks. As in one dimension, the exact distribution of the scan statistic under the null hypothesis is only available for special cases, this research employs other methodologies, such as Monte Carlo (MC) simulations and the extreme value theory to estimate p-values. In addition, the size and power of the scan statistic have been estimated. Once the sampling distribution of scan statistic is determined, the inference on anomaly can be performed. The extreme value theory is a statistical model that is used to model the extreme data in a given period of time, and is based on the location-scale family. Gumbel distribution is the most well-known distribution that belongs to this family, and has been widely used in engineering. The present work also applies the Gumbel distribution to approximate the p-value. Another challenge is the choice of local statistic as it provides important structural location of a node and its neighborhood. Here I apply graph invariant, betweenness, as a measure to identify local structure and anomaly in social networks. For building time series of graphs, the challenges are to compute the graph distance metrics, which are computationally intensive, and to fit time series model to assess anomalies based on residuals.

The conclusions of this dissertation work are summarized below.

(i). The Poisson process model is employed to identify the most likely cluster of emails using the email count data as a step-1 scan statistic process for a 10-year period. The most likely statistically significant temporal primary cluster is detected using the maximum LLR as the scan statistic (LLR = 82.07, p = 0.001).

(ii). The binomial model applied to network data in step-2 scan statistic process, using betweennes as the locality statistics and the maximum LLR as the scan statistic, detects the more refined purely temporal clusters around the primary cluster, obtained from the step-1 scan process.

(iii). A univariate time series has been built using the graph edit distance metric to compare subgraphs between two consecutive weeks. An autoregressive moving average process is then fitted to the GED time series, and the anomalies were assessed using residuals from the fitted model exceeding a threshold, and compared with the results obtained from the scan statistics. The residual analysis demonstrates the statistically significant excessive activity, consistent with the results obtained using scan statistics.

(iv). A vector autoregressive model has been fitted to the email ego-centered subnetworks. From this multivariate time series analysis, it is concluded that the excessive chatter has been initiated by ID5, and as k increases, the excessive chatter has been observed to diffuse among ID1, ID7, ID10 and ID20. In addition, this analysis clearly demonstrates the social relationship among ID1, ID5, ID7, ID10 and ID20.

(v). An excessive topic activity or topic chatter from the 4^{th} week of October to 4^{th} week November 2003 have been identified by combining the LDA and the scan statistics. The primary cluster consists of mostly the topic-1, which is related to data used in a paper. Furthermore, it is concluded that ID1, ID5, ID7, ID10 and ID20 are mostly discussing topic 1.

5.2 Future Work

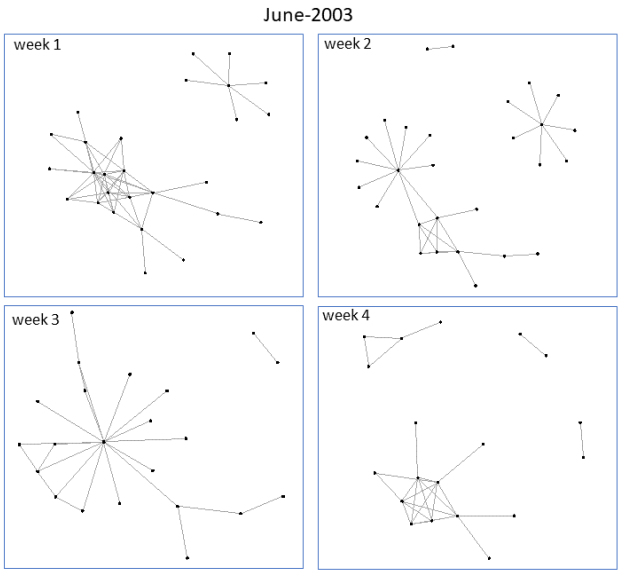
This dissertation work opens up many potential areas of the future research. These methodologies can be applied to other networks, such as Twitter, Facebook, and telecommunication networks. In particular, the detection of multiple excessive activities would be an area of interest. These methods can be extended to investigate multiple clusters using scan statistics with variable window sizes and overlapping windows.

The over-dispersion of the Poisson counts may occur in count data. In such cases, the negative binomial distribution can be employed. Furthermore, one can employ the modified ARMA models, such as integer valued ARMA (INARMA), INGARCH models to fit the over-dispersed and correlated count data to get the critical cluster.

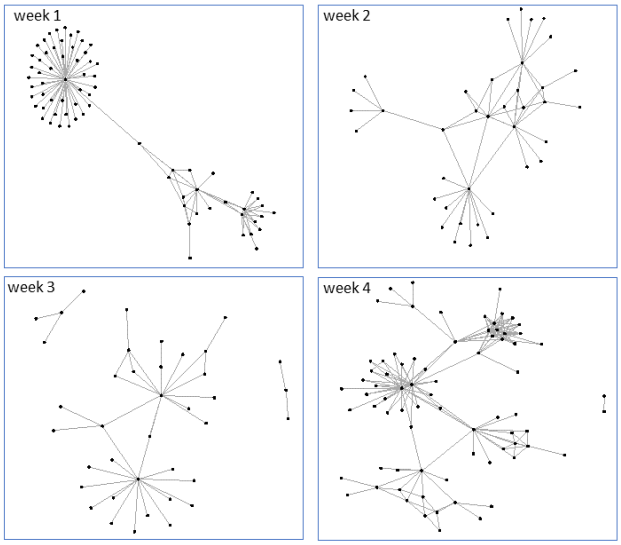
The VAR model, in chapter 4, could be extended in different ways. It would be possible to construct a VARMA model to detect the excessive activities. If the multiple time series is co-integrated, one could use vector error correction model (VECM) instead of the VAR models. An interesting area for further research would be forecasting the excessive activities using the VAR models as a part of prospective surveillance.

Chapter 6: Appendix A

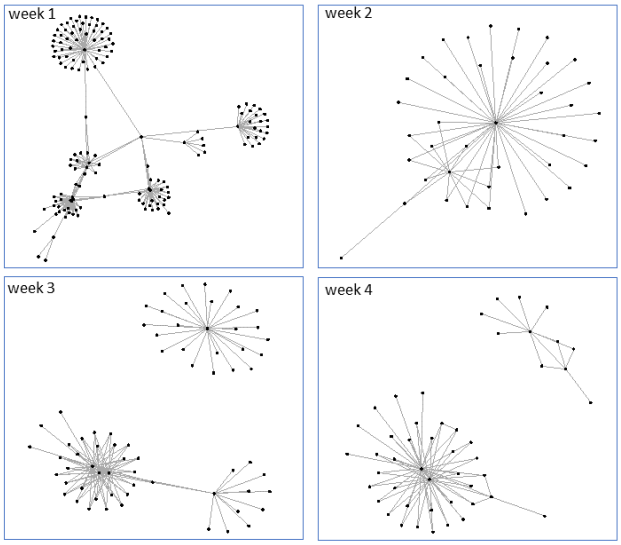
The weekly networks (week 1 to week 4) from June 2003 to June 2004, extracted from the e-mail edge lists, for the 52 week period are given below. Such networks have been constructed around the most likely cluster, obtained from the two-step scan process.

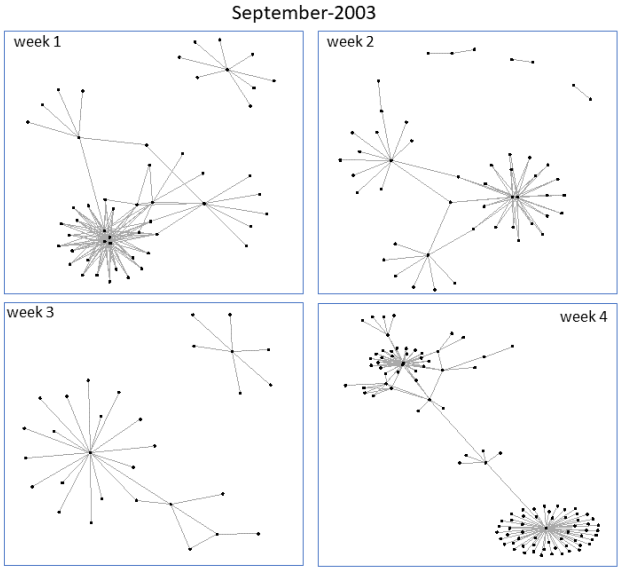


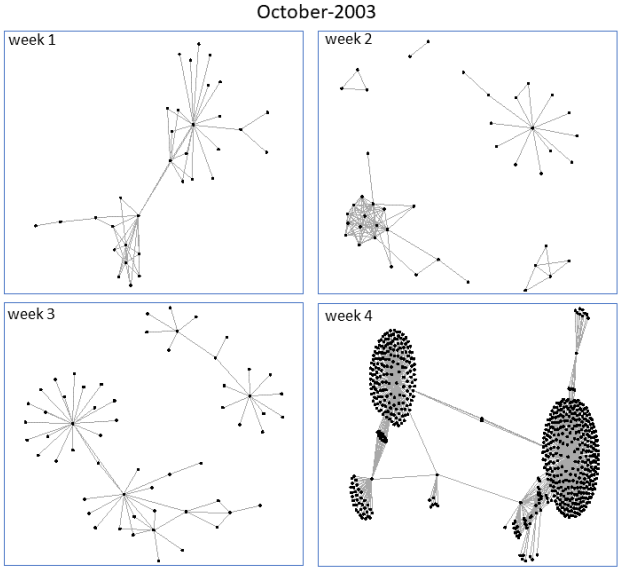
July-2003

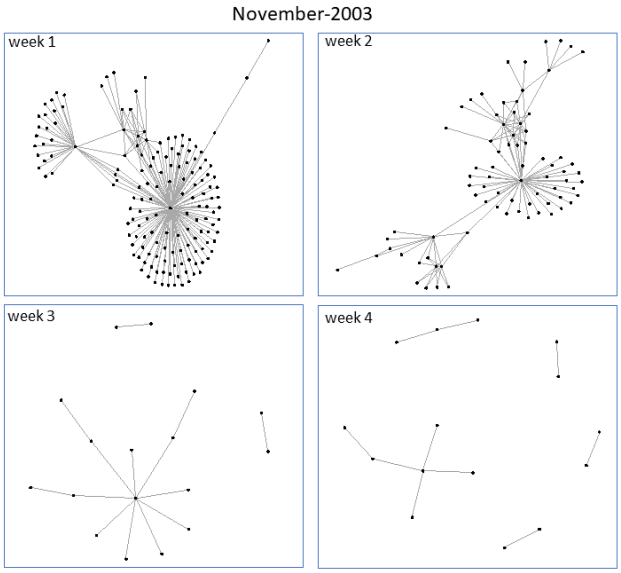


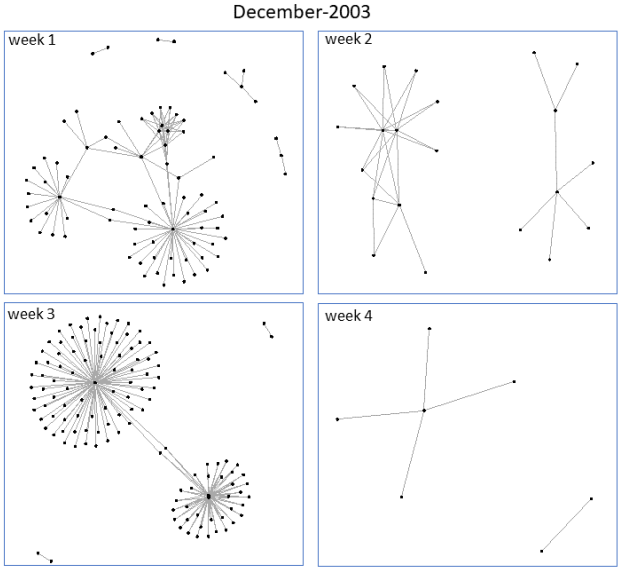
August-2003

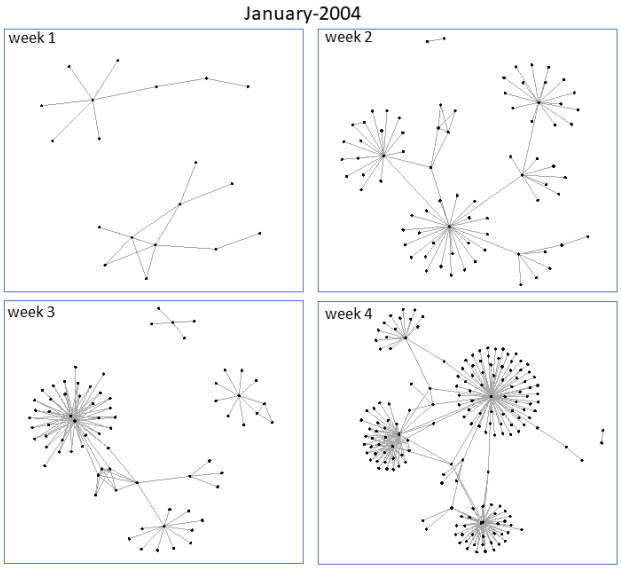


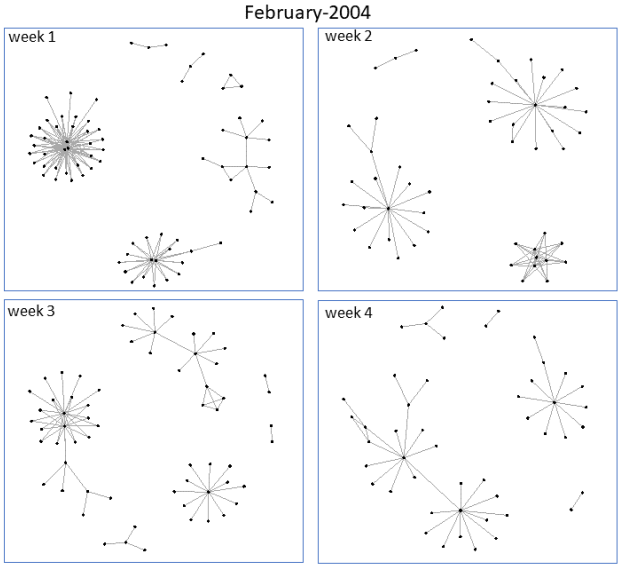

September-2003
week 1
week 2
week 3
week 4

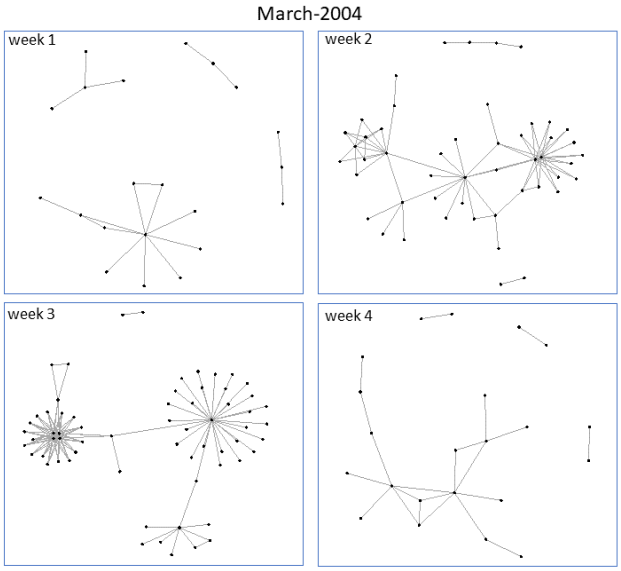

October-2003
week 1
week 2
week 3
week 4

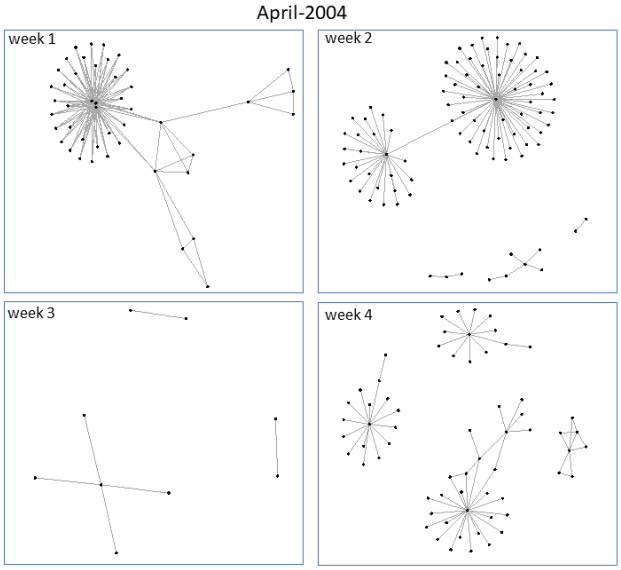

November-2003
week 1
week 2
week 3
week 4

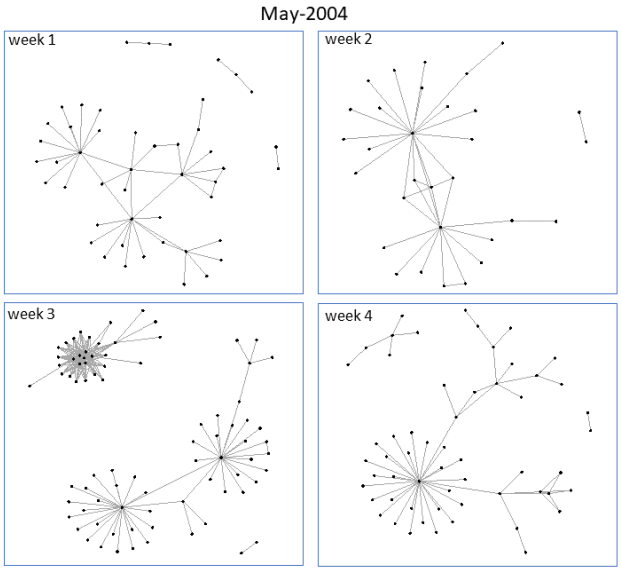

December-2003
week 1
week 2
week 3
week 4

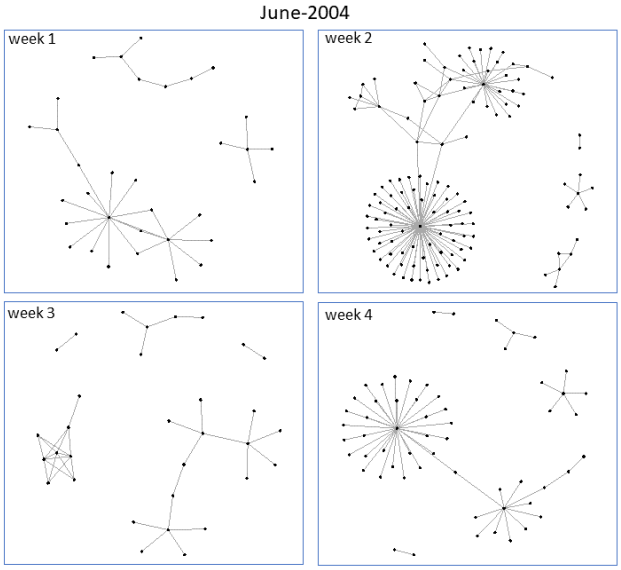

January-2004
week 1
week 2
week 3
week 4


February-2004
week 1
week 2
week 3
week 4


March-2004
week 1
week 2
week 3
week 4


April-2004
week 1
week 2
week 3
week 4


May-2004
week 1
week 2
week 3
week 4


June-2004
week 1
week 2
week 3
week 4

Chapter 7: Appendix B

Here the partial codes in R for size and power calculations and the maximum log likelihood estimation using non-parametric method have been provided.

1. **Size and Power Calculations**

```
poisson2 <-function(M)
email <- rpois(153,158.8301)
llr = rep(0,M)
tim = rep(1,n)
SumE = 0
SumT = 0
llr = 0
for(i in 1:n)
SumE = email[i]+SumE
SumT =tim[i]+SumT
lambda1Tot = (SumE/SumT)
lambda0 = lambda1Tot
lambda1 = 158.88
A2 = log(lambda1/lambda0)
A1=(lambda0-lambda1)
llr=((SumT*A1)+(A2*SumE))
return(abs(llr))
M=1000
p2 = rep(0,M)
for(i in 1:M)
```

```
p2[i] <-poisson2(i)
cv <- quantile(p2,c(0.9,0.95,0.975,0.99,0.995,0.999))
M = 1000
Im = 0
p2 = rep(0,M)
for(i in 1:M)
p2[i]< -poisson2(i)
if (p2[i] > cv[3])
Im = Im+1
alphahat < -Im/M
poisson3 < -function(M,lamda1)
email < -rpois(n,lamda1)
llr = rep(0,M)
tim = rep(1,n)
SumE = 0
SumT = 0
llr = 0
for(i in 1:n)
SumE = email[i]+SumE
SumT = tim[i]+SumT
lambda1Tot = (SumE/SumT)
lambda1 = lambda1Tot
lambda0 = 158.8301
A2 = log(lambda1/lambda0)
A1 =(lambda0-lambda1)
llr=-((SumT*A1)+(A2*SumE))
return(abs(llr))
M = 1000
```

lambda1 = seq(from=158.81, to=170.9, by=0.1)

betahat = rep(0,length(lambda1))

powerhat = rep(0,length(lambda1))

for (j in 1:length(lambda1))

Im = 0

lamda1=lambda1[j]

p3 = rep(0,M)

for(i in 1:M)

p3[i] <-poisson3(i,lamda1)

if (p3[i] <cv)

Im = Im+1

betahat[j]<-Im/M

powerhat[j]<-1-betahat[j]

2. **Maximum likelihood estimate: Non Parametric**

z <- ufit(y,x = x,lmode = x[1],lc = TRUE,type = "b")

plot(x,y)

lines(z,col="red")

plot (z-h, do.points = TRUE,col.hor="black",col.vert="black",

add = FALSE,xlab='x',ylab='f')

text(1, 0.025, paste("Mode at =", x[1]),cex=1.0)

plot(z-x,z-y)

lines(z,type='l',lwd=2)

text(12,2.5e-8, paste("Mode at =", x[1]),cex=1.0)

qplot(zx, zy,data=df1,geom=c("point","line"))

+annotate("text",x=10,y=2.5e-08,label="Mode at 1",cex=5)

z <- ufit(y,x=x,lmode=x[32],lc = TRUE,type="b")

```
plot(x,y)
lines(z,col="red")
plot(zh,do.points=TRUE,col.hor="black",col.vert="black",
add=FALSE,xlab='x',ylab='f')
text(1, 0.025, paste("Mode at =", x[32]),cex=1.0)
text(25,0.0e+00, paste("Mode at =", x[32]),cex=1.0)
qplot(zx, zy,data=df1,geom=c("point","line"))
+annotate("text",x=25,y=2e-08,label="Mode at 32",cex=5)
qplot(zx, zy,data=df1,geom=c("point","line"))
for (i in 1:32)
{
z <- ufit(y,x=x,lmode=x[i],lc=TRUE,type="b")
plot(x,y)
lines(z,col="red")
plot(zh,do.points=TRUE,col.hor="black",col.vert="black",
add=FALSE,xlab='x',ylab='f')
text(-0.05, 0.05, paste("Mode at =", x[i]),cex=1.0)
plot(zx, zy)
lines(z,type='l',lwd=2)
text(-0.05, 0.05, paste("Mode at =", x[i]),cex=1.0)
print(-sum(log(zy)),digits=20) }
```

Bibliography

Bibliography

[Adl84] R. J. Adler. The supremum of a particular gaussian field. *Annals of Probability*, 12:436–444, 1984.

[ADW94] C. Apte, F. Damerau, and S. M. Weiss. Towards language independent automated learning of text categorization models. In *Proc. ACM-SIGIR conference on Information Retrieval*. Springer-Verlag New York, 1994.

[Aka73] H. Akaike. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory*, Budapest:267–281, 1973.

[Alm83] S. E. Alm. On the distribution of the scan statistic of a poisson process. *Probability and Mathematical Statistics*, 5:1–10, 1983.

[Arr05] M. A. Arranz. Portmanteau test statistics in time series. *Tol-Project, www.tol-project.org*, 2005.

[BD93] B.E. Brodsky and B.S. Darkhovsky. *Nonparametric Methods in Change-Point Problems*. Springer-Science + Business Media, B.V., 1993.

[BDGH93] A. Banerjee, J. Dolado, J. Galbraith, and D. Hendry. *Co-Integration, Error-correction and the Econometric Analysis of Non-Stationary Data*. Oxford University Press Oxford, 1993.

[BKNS00] M. M. Breunig, H. P. Kriegel, R. T. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *Procedings of the 2000 ACM SIGMOD International Conference on Management of Data*. SIGMOD, 2000.

[BNJ03] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Bro08] C. Brooks. *Introductory Econometrics for Finance*. Cambridge University Press, Cambridge, UK, 2008.

[BVAMF07] C. Barcelo-Vidal, L. Aguilar, and J. A. Martin-Fernandez. Compositional time series: a first approach. *Procedings of the 22nd International Workshop of Statistical Modeling, Bercelona, Spain*, pages 81–86, 2007.

[CD04] W. W. Chen and R. S. Deo. A generalized portmanteau goodness of fit test for time series. *Econometric Theory*, 20:382–416, 2004.

[CG92] G. Casella and E. I. George. Explaining the gibbs sampler. *American Statistian*, 46:167–174, 1992.

[CH74a] T. Calinski and J. Harabazs. A dendrite method in cluster analysis. *Communications in Statistics*, 3:1–27, 1974.

[CH74b] D. R. Cox and D. V. Hinkley. Theoretical statistics. *Chapman and Hall, London*, 1974.

[Che02] Y. T. Chen. On the robustness of ljung-box and mcleod-li q tests: A simulation study. *Economics Bulletin*, 3:1–10, 2002.

[Cle93] W. S. Cleveland. *Visualizing Data*. Hobart Press, Summit, New Jersey, 1993.

[CM09] P. S. P. Cowpertwait and A. V. Metcalfe. *Introductory time series with R*. Springer, Dordrecht, 2009.

[DBDK02] P. Dickinson, H. Bunke, A. Dadej, and M. Kraetzl. Median graphs and anomalous change detection in communication networks. In *Information, Decision and Control Conference*. IDC-2002 Adelaide, 2002.

[DH97] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. Cambridge University Press, Cambridge, United Kingdom, 1997.

[DLR77] A. P. Demster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

[DPL+15] R. Dalinina, V. A. Petryshyn, D. S. Lim, A. J. Braverman, and A. K. Tripati. Application of clustering techniques to study environmental characteristics of microbialite-bearing aquatic systems. *Biogeosciences Discussions*, 12:10511–10544, 2015.

[EG87] R. F. Engle and C. W. J. Granger. Co-integration and error correction: Representation, estimation and testing. *Econometrica*, 55:251–276, 1987.

[Fre79] L. C. Freeman. Centrality in social networks' conceptual clarification. *Social Networks*, 1:215–239, 1979.

[Fri03] M. Frisen. Statistical survaillance: Optimality and methods. *International Statistical Review*, 71:403–434, 2003.

[GB99] J. Glaz and N. Balakrishnan. *Scan Statistics and Applications, Statistics for Industry and Technology*. Springer Science + Business Media, LLC, New York, 1999.

[Gen02] J. E. Gentle. *Elements of Computational Statistics*. Springer, 2002.

[Gen15] M. Genin. Discrete scan statistics and the generalized likelihood ratio test. *Rev. Roumaine Math. Pures Appl.*, 60:83–92, 2015.

[Gla79] J. Glaz. Expected waiting time for a visual response. *Biological Cybernetics*, 35:39–41, 1979.

[Gla81] J. Glaz. Clustering of events in a stochastic process. *Journal of Applied Probability*, 18:268–275, 1981.

[GNW01] J. Glaz, J. Naus, and S. Wallenstein. *Scan Statistics, Springer Series in Statistics*. Springer, New York, 2001.

[GPW09] J. Glaz, V. Pozdnyakov, and S. Wallenstein. *Clustering of events in a stochastic process, Statistics for Industry and Technology*. Winsdell Publishing Company, 2009.

[Gra81] C. W. J. Granger. Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, 16:121–130, 1981.

[Gra17] J. Grandell. *Formulas and Survey Time Series Analysis*. Avd. Matematisk Statistik, www.Math.kth.se/mastat/gru/sf2943/tsform, Visited Fall 2017, 2017.

[GS04] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101:5228–5235, 2004.

[GW16] S. Goswami and E. J. Wegman. Comparison of classification methods for forensic research. *Journal of Statistical Science and Application*, 4:65–84, 2016.

[HHWB00] S. Hawkins, H. X. He, G. J. Williams, and R. A. Baxter. Outlier detection using replicator neural networks. In *Proc. Fifth International Conference and Data Warehousing and Knowledge Discovery (DaWaKO2)*. Aix en Provence, France, 2000.

[HN75] R. Huntington and J. Naus. A simpler expression for k^{th} nearest neighbor coincidence probabilities. *The Annals of Probabilities*, 3:894–896, 1975.

[Hof99] T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in information retrieval*, pages 50–57, 1999.

[Hoh08] E. L. Hohman. *A Dynamic Graph Model for Representing Streaming Text Documents,*. a PhD Dissertation, 2008.

[HQ79] E. J. Hannan and B. G. Quinn. The determination of the order of an autoregression. *Journal of the Royal Statistical Society, Series B*, 41:190–195, 1979.

[HR05] R. A. Hannenman and M. Riddle. *Introduction to social network methods*. University of California, Riverside CA, 2005.

[HTF11] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning Data Mining Inference and Prediction*. Springer Series in Statistics Second Edition, 2011.

[HXD03] Z. He, X. Xu, and S. Deng. Discovering cluster-based local outliers. *Pattern Recognition Letters*, 24:1641–1650, 2003.

[Joa98] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning*. Springer Verlag Heidelberg DE, 1998.

[JSTW18] D. R. Jeske, N. T. Stevens, A. G. Tartakovsky, and J. D. Wilson. Statistical methods for network surveillance. *Applied Stochastic Models in Business and Industry*, pages 1–21, 2018.

[Jun16] I. Jung. Scan tests. In *Handbook of Spatial Epidemiology, Eds. A. B. Lawson, S. Banerjee, R. P. Haining and M. D. Ugarte*. Chapman and Hall, CRC Press, Taylor and Francis Group, 2016.

[JWHT13] G. James, D. Witten, T. Hastie, and R. Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer Series in Statistics, 2013.

[KB92] S. Karlin and V. Brendel. Chance and statistical significance in protein and dna sequence analysis. *Science*, 257:39–49, 1992.

[KHK09] M. Kulldorff, L. Huang, and K. Konty. A scan statistic for continuous data based on the normal probability model. *International Journal of Health Geographics*, 8:58, 2009.

[KNT00] E. M. Knorr, R. T. Ng, and V. Tucakov. Distance-based outliers: algorithms and applications. *International Journal on Very Large Data Bases*, 8:237–253, 2000.

[Kre02] V. E. Krebs. Mapping networks of terrorist cells. *Connections*, 24:43–52, 2002.

[Kul79] M. Kulldorff. A spatial scan statistics. *Communications in Statistics-Theory and Methods*, 26:1481–1496, 1979.

[LK04] H. Lutkepohl and M. Kratzig. *Applied Time Series Econometrics*. Cambridge University Press Cambrigde, 2004.

[LK05] A. Lazarevic and V. Kumar. Feature bagging for outlier detection. In *Proc. 11^{th} ACM SIGKDD international conference on Knowledge Discovery in Data Mining*, pages 157–166, 2005.

[LM81] W. Li and A. McLeod. Distribution of the residual autocorrelations in multivariate arma time series models. *J. Roy. Statist. Soc. Ser. B*, 43:231–239, 1981.

[Lov68] J. B. Lovins. Development of a stemming algorithm. *Mechanical Translation Computational Linguistics*, 11:22–31, 1968.

[Lut06] H. Lutkepohl. *New Introduction to Multiple Time Series Analysis*. Springer-Verlag New York, 2006.

[Mar12] D. J. Marchette. Scan statistics on graphs. *WIREs. Comput. Stat.*, 4:466–473, 2012.

[MBB17] S. Moritz and T. Bartz-Beislstein. inputets: Time series missing value inputation in r. *The R Journal*, 9:207–218, 2017.

[MC85] G. W. Miligan and M. C. Cooper. An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, 50:159–179, 1985.

[MGW92] B. Masand, L. Gordon, and D. Waltz. Classifying news stories using memory based reasoning. In *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM Press New York, 1992.

[MM16] W. L. Martinez and A. R. Martinez. *Computational Statistics Handbook with MATLAB*. CRC Press, 2016.

[MS02] C. D. Manning and H. Schutze. *Foundations of Statistical Natural Language Processing*. The MIT Press, Cambridge, Massachusetts, 2002.

[Nau65] J. Naus. The distribution of the size of the maximum cluster of points on a line. *Journal of the American Statistical Association*, 60:532–538, 1965.

[Nau82] J. Naus. Approximation for distribution of scan statistics. *Journal of the American Statistical Association*, 77:377–385, 1982.

[NCS02] B. V. North, D. Curtis, and P. C. Sham. A note on the calculation of emperical p values from monte carlo procedures. *American JOurnal of Human Genetics*, 71:439–441, 2002.

[New63] G. F. Newell. Distribution for the smallest distance between any pair of k^{th} nearest-neighbor random points on a line. In *Time Series Analysis, Proceedings of the Conference*. Brown University (Ed., M. Rosen-blatt), New York: Academic Press, 1963.

[New10] M. E. J. Newman. *Networks: An Introduction*. Oxford University Press, Oxford OX2 6DP, 2010.

[NP95] S. Ng and P. Perron. Unit root tests in arma models with data dependent methods for the selection of the truncation lag. *Journal of the American Statistical Association*, 90:268–281, 1995.

[NP01] S. Ng and P. Perron. Lag length selection and the construction of unit root tests with good size and power. *Econometrica*, 69:1519–1554, 2001.

[PCMP05] C. E. Priebe, M. J. Conroy, D. J. Marchette, and Y. Park. Scan statistics on enron graphs. *Computational Mathematical Organization Theory*, 11:229–247, 2005.

[Pfa06] B. Pfaff. *Analysis of Integrated and Cointegrated Time Series with R*. Springer-Verlag New York, 2006.

[Pfa08] B. Pfaff. Var, svar and svec models: Implementation within r package vars. *Journal of Statistical Software*, 27:1–32, 2008.

[PGKS05] V. Pozdnyakov, J. Glaz, M. Kulldorff, and J. M. Stllele. A mertingale approach to scan statistics. *Ann. Inst, Statist. Math.*, 57:21–37, 2005.

[Pin05] B. Pincombe. Anomaly detection in time series of graphs using arma processes. *ASOR Bulletin*, 24:2–10, 2005.

[POH98] C. E. Priebe, T. Olson, and D. M. Healy. A spatial scan statistic for stochastic scan partitions. *Journal of the American Statistical Association*, 92:1476–1484, 1998.

[PPM+10] C. E. Priebe, Y. Park, D. J. Marchette, J. M. Conroy, J. Grothendieck, and A. L. Gorin. Statistical inference on attributed random graphs: Fusion of graph features and content: An experiment on time series of enron graphs. *Computational Statistics and Data Analysis*, 54:1766–1776, 2010.

[PS00] M. D. Porter and R. Smith. Network neighborhood analysis. In *Intelligence and Security Informatics (ISI)*. IEEE International Conference, 2000.

[Ren94] E. Rennison. Galaxy of news: An approach to visualizing and understanding expansive news landscapes. In *ACM Symposium on User Interface Software and Technology*, 1994.

[Rob67] T. Robertson. On estimating a density which is measurable with respect to a σ-lattice. *The Annals of Mathematical Statistics*, 38:303–358, 1967.

[Rub87] D. B. Rubin. *Multiple imputation for nonresponse in surveys*. John Wiley and Sons, 1987.

[Sch78] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6:461–464, 1978.

[Sch89] G. W. Schwert. Tests for unit roots: A monte carlo investigation. *Journal of Business and Economic Statistics*, 7:147–160, 1989.

[SG07] M. Steyvers and T. L. Griffiths. Probabilistic topic models. *Handbook of latent semantic analysis*, 427:424–440, 2007.

[Sim80] C. A. Sims. Macroeconomics and reality. *Connections*, 48:1–48, 1980.

[SKM10] R. S. Sparks, T. Keighley, and D. Muscatello. Early warning cusum plans for surveillance of negative binomial daily disease counts. *Journal of Applied Statistics*, 37:1911–1929, 2010.

[SKR99] P. J. Shoobridge, M. Kraetzl, and D. Ray. Detection of abnormal change in dynamic networks. In *Information, Decision and Control Conference (IDC'99)*. IDC'99 Adelaide, 1999.

[SM83] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill Inc. New York, 1983.

[Sol08] J. L. Solka. Text data mining: theory and methods. *Statistics Surveys*, 2:94–112, 2008.

[SPST+01] B. Schlkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Estimating the support of a high-dimensional distribution. *Neural Computation*, 13:1443–1471, 2001.

[SS06] R. H. Shumway and D. S. Stoffer. *Time series analysis and its applications with R examples*. Springer Texts in Statistics) 2nd Edition, 2006.

[SSBM96] G. Salton, A. Singhal, C. Buckley, and M. Mitra. Automatic text decomposition using text segments and text themes. In *Proceedings of the seventh ACM conference on Hypertext*. ACM Press New York, 1996.

[Sta10] T. Stadnitska. Deterministic or stochastic trend: Decision on the basis of the augmented dickey-fuller test. *Methodology*, 6:83–92, 2010.

[Str99] D. W. Stroock. *Probability theory: an analytic view*. Cambridge University Press, 1999.

[Sut12] C. D. Sutton. *Nonparametric Statistics: Order Statistics, Quantiles and Coverages*. Class notes (Stat 657), George Mason University, 2012.

[SWY75] G. Salton, A. Wong, and L. S. Yang. Spatial scan statistics adjusted for multiple clusters. *Journal of the American Society for Information Science*, 18:613–620, 1975.

[SZ10] X. Shao and X. Zhang. Testing for change points in time series. *J. Am. Stat. Association*, 105:1288–1240, 2010.

[SZY+14] D. Savage, X. Zhang, X. Yu, P. Chou, and Q. Wang. Anomaly detection in online social networks. *Social Networks*, 39:62–70, 2014.

[Tsa14] R. S. Tsay. *Multivariate time series analysis with R and financial applications*. Wiley, 2014.

[VA80] P. Vacek and T. Ashikaga. An examination of the nearest neighbor rule fot imputing missing values. *Proc. Statist. Computing Sect*, pages 326–331, 1980.

[Wal87] K. F. Wallis. Time series analysis of bounded economic variables. *Journal Time Series Analysis*, 8:115–123, 1987.

[Weg70] E. J. Wegman. Maximum likelihood estimation of a unimodal density function. *The Annals of Mathematical Statistics*, 41:341–364, 1970.

[Weg90] E. J. Wegman. Hyperdimensional data analysis using parallel coordinates. *Journal of American Statistical Association*, 85:664–675, 1990.

[Weg11] E. J. Wegman. Order restricted inference: computational algorithms. *Joint Statistical Meetings, Miami, Fl*, 2011.

[Wei07] C.H. Weiss. Controlling correlated processes of poisson counts. *Quality and reliability engineering international*, 23:389–400, 2007.

[Wei09] C.H. Weiss. Modelling time series counts with overdispersion. *Statistical Methods and Applications*, 18:507–519, 2009.

[WN74] S. Wallenstein and J. Naus. Probabilities for the size of the largest clusters and smallest intervals. *Journal of the American Statistical Association*, 69:690–697, 1974.

[WN87] S. Wallenstein and N. Neff. An approximation for the distribution of the scan statistic. *Statistics in Medicine*, 12:1–15, 1987.

[WT09] C.H. Weiss and M. C. Testik. Cusum monitoring of first-order inter-valued autoregressive processes of poisson counts. *Journal of quality technology*, 41:349–400, 2009.

[ZAK10] Z. Zhang, R. Assuncao, and M. Kulldorff. Spatial scan statistics adjusted for multiple clusters. *Journal of Probability and Statistics*, 64:2379–2389, 2010.

[ZCP+15] W. Zhao, J. J. Chen, R. Perkins, Z. Liu, W. Ge, Y. Ding, and W. Zou. A heuristic approach to determine appropriate number of topics in topic modeling. *BMC Bioinformatics (suppl 13)*, 16:S8, 2015.

[ZW06] E. Zivot and J. Wang. *Modeling financial time series with S-PLUS*. Springer, New York, NY, 2006.

Curriculum Vitae

Suchismita Goswami earned B.Sc (Horns) in Mathematics from Krishnath College, Calcutta University and M.Sc. in Applied Mathematics from Vidyasagar University, India. She worked as a lecturer in Berhampore college, India, where she taught linear algebra and calculus after her graduation. She then moved to Stony Brook, NY and enrolled in the MS program in Applied Mathematics and Statistics (Operation Research) at SUNY Stony Brook in 2000, and graduated in 2001. She completed an MS in Epidemiology and Biostatistics in 2009, and MS in Statistical Science in 2013 from George Mason University. She enrolled in the PhD program in Computational Science and Informatics (CSI) at George Mason University in Fall 2013. She worked at the World Health Organization, Washington, DC as a summer intern on the development database and statistical modeling of Malaria. She has presented her work in national and international conferences. She chaired sessions on Time Series Models and Bioinformatics at an SDSS conference in 2018. She published a journal paper, entitled Comparison of Different Classification Methods on Glass Identification for Forensic Research Journal of Statistical Science and Application, April 2016, Vol. 4, No. 03-04, 65-84 doi: 10.17265/2328-224X/2015.0304.001, and another journal paper (under progress), Detection of Excessive Activities in Time Series of Graphs, Journal of Applied Statistics, communicated February 2019.