CHAPTER 1

INTRODUCTION

1.1 MOTIVATION

With the enormous growth of internet for sharing information and e-commerce applications, network traffic has become a key issue in maintaining the performance of the network, application and web servers for Internet Service Providers (ISP). A deep recognition of network traffic behavior is required for various kinds of design and network operation. The key areas include designing phase of component and protocol design for performing modeling and simulation.

The expeditious growth of Internet acceptance is changing the networking industry. In general scenario, all have to make an access to a wide range of Internet services that consists of entertainment, news, providing business transactions in a secured manner and information exchanges. The divergent nature of internet ranging from audio to video streaming file or data transfers are dependent on a variety of network transport characteristics. In addition to that, enterprise level computing and resources related to network must be preserved from malicious attacks.

During the early stages of Internet development, decisions regarding forwarding the packet were based on simple look-up operations belonged to Layer 2(Data link layer) and Layer 3(Network layer) fields of a packet header. In general, today multiple look-up operations by using the Layer 3(Network layer) and Layer 4(Transport layer) fields are becoming common features in networking equipment.
Network traffic estimation originates from the fact of flow related to Origin-Destination (OD). Some of the problems related to network traffic are the involvement of traffic engineering, estimating traffic matrix, planning of capacity and forecasting issues. Anyhow the OD flow has not been extensively studied and a very little work related to OD flow is known.

One way to solve the problem of whole network traffic analysis is to identify the traffic acquired on different links of a network and it is also not independent. In fact, it is determined by using OD flows and a routing matrix. The OD flow refers to the collection of all traffic that enters into the network using a common ingress point and leaves out from a common egress point. Hence, instead of making an analysis of traffic involved on all links, the fundamental analysis can be used to provide a set of OD flows in whole network traffic.

Subsequently, when stated to analyze a high dimensional structure, the most commonly employed and powerful methodology is to provide an alternate lower dimensional approximation to the framework which preserves the highest required properties. In this way, the large dimension data are largely measured by way of small set of independent variables. Hence they can be well approximated by using lower dimensional representation.

The Figure 1.1 shows a typical deployment of OD flows in order to monitor the network traffic. It consists of three nodes connected through Local Area Network (LAN), a single dialup node, a node in the form of wireless media and a standalone computer. The network traffic is measured for all types of nodes. The look up table lists out the parameters used for measuring network traffic. They are IP address of local and remote machine, Port number and Login user name and notification of alarms to the administrator in case of excess usage.
The recent work on literature provides various mining mechanisms to monitor and manage network traffic from different peer networks that improves network performance appreciably. However, current exponential rise in the internet operations increase the flow of network streams to manifolds. The huge flow of network traffic provided the focus shift from
simple volume based analysis to high volume distributed network flow analysis.

The most distinct characteristic of data mining is that it deals with very large data sets (gigabytes or even terabytes). This requires the algorithms used in data mining to be scalable. However, most algorithms currently used in data mining cannot be applied for very large data sets because they were initially developed for other applications than data mining which involve small data sets. The study of scalable data mining algorithms for network traffic data has recently become a data mining research focus.

The usage of data flow collected from the site routers for various analysis i.e., network performance characterization, investigating computer security incidents and their prevention, network traffic statistics, and others. Currently, the data flow analysis is constructed as distributed system that collects data from multiple routers, both at the edge of the site networks as well as from the local routers and multilayer switches. The average per day volume of raw data is about 2GigaBytes (GB). Despite a high volume of collected information, some analysis is conducted in near real time to satisfy demands of user’s communities for quicker results.

The Figure 1.2 shows the pictorial representation of the traffic data used to mine. Initially the connection status regarding the network is initialized. Based on the connection status provided, the packet sizes are determined. The next step is to make provisions for the data arrangements for first connection and subsequently for all connections. The frequency patterns for both local and global parameters are analyzed. Finally, mining of network traffic data is performed using the threshold value.

In network traffic flow analysis the characterization has been a shift from volume-based analysis to network flow distribution-based analysis. The
traditional method for classifying network traffic is of two types. They are divided into port-based and payload-based. Due to the dynamic allocation of port and as it can share many applications; network traffic classification based on port is discarded. Because more and more network applications dynamically allocate port and a port can be shared by different applications, the port-based classification becomes less effective.

Figure 1.2 Mining Traffic Data

Payload-based classification method also has its limitation that classification rules must be updated frequently, whenever a small protocol change is applied to the network traffic data. With the help of machine learning the network traffic builds a classifier model using the features such as packet length and inter-arrival time. But due to the increased development
in high-speed network, classification of several thousands of flows requires significant memory consumption and Central Processing Unit (CPU) usage.

The decisive aim of network traffic flow is to create a framework and to gear up with a model which would make possible to reach their destination in the shortest possible time using the maximum bandwidth capacity. Any network traffic to be mined goes through the four phases as discussed below separately.

![Four phase model for network traffic estimates](image)

**Figure 1.3 Four phase model for network traffic estimates**

The four phase model for network traffic estimates are described as follows:

i. **Formation**: In this phase the traffic estimates, how many bounds can be generated. For this, the program needs the statistical data of both local and remote IP address.

ii. **Assessment**: After formation phase is derived, the different OD pair between the network traffic is found in step 1.

iii. **Split or Mode transition**: The system has to decide how much percentage of the bandwidth has to be split between the different modes of available ports.
iv. **Route Assignment**: Finally, routes are assigned to the different modes of machine (wired or wireless or standalone) on minimum threshold value.

This cycle is repeated until the solution converges. The pictorial representation of the four phase model for network traffic estimates is shown in Figure 1.3.

1.1.1 **Distributed Data Mining**

The role of distributed computing in data mining is very important due to the key role played by it in the recent data mining process for several reasons. The first reason is that the data mining work requires large amounts of resources in terms of storage space and computation time. In order to make the various systems scalable, it is highly required to develop mechanisms that distribute the work load among several sites or machines connected in different mode such as Local Area Network (LAN), Wide Area Network (WAN), and Metropolitan Area Network (MAN) in a flexible way. The second reason is that data distribution is to make it more efficient though they might be prone to security risks.

Distributed Data Mining (DDM) in Peer to Peer (P2P) network traffic is emerging as a new distributed computing paradigm for many novel network applications that involve exchange of information among a large number of peers with less centralized coordination. Huge data sets are being collected regularly from the P2P network traffic, but it is still extremely difficult to draw conclusions or make decisions based on the collective characteristics of such dynamically changing traffic volumes.

Analyzing massive data sets, which often span different sites, using traditional centralized approaches can be intractable. DDM is being fueled by
recent advances in grid infrastructures and distributed computing platforms. Huge data sets are being collected daily in different fields; e.g., retail chains, banking, biomedicine, astronomy, and so on, but it is still extremely difficult to draw conclusions or make decisions based on the collective characteristics of such disparate data. Four main approaches for performing DDM can be identified. A common approach is to bring the data to a central site, then apply centralized data mining on the collected data. Such approach clearly suffers from a huge communication and computation cost to pool and mine the global data. In addition, the data privacy cannot be preserved in such scenarios.

Due to the advancement in communication over wired and wireless networks there has been a great involvement of work in pervasive distributed computing environments. Some of the examples in distributed computing environment include internet, intranets, local area networks, ad hoc wireless networks, and sensor networks. These environments come with different distributed sources of computation. Data mining in the environments discussed above calls for proper utilization of these distributed resources.

Communication in P2P DDM will be very costly, if care is not taken to localize traffic, instead of relying on flooding of control or data messages. Regardless of any particular DDM approach, the distributed nature of the DDM concept itself usually entails a tradeoff between accuracy and scalability. In order to achieve better accuracy, the finer level of information exchanged between distributed nodes should become finer for different nodes and the different nodes should be increased. On the other hand, if better scalability is desired, the granularity should be coarser and the connectedness should be reduced.

One of the most popular methods used to implement the DDM operation is the use of Clustering. The work of clustering is to partition a set
of objects into cluster in such a way that objects belonging to the same cluster are more similar than that belonging to other clusters. Statistical clustering methods use similarity measures to partition objects whereas conceptual clustering methods cluster objects according to the concepts objects.

Clustering is one of the techniques recognized as an important area in research. Recent research shows that clustering involving in high-dimensional data has claimed that traditional clustering techniques may have the limitations of discovering meaningful clusters due to the varied dimensionality of data involved.

The dimensionality of data for the sake of mining is reduced by using appropriate techniques called feature transformation and feature selection as depicted in Figure 1.4. Feature transformation is performed using Principal Component Analysis (PCoA) and Singular Value Decomposition (SVD). The feature selection methods reduce the data dimensionality by selecting the most relevant attributes from the original data attributes. In this way only a particular subspace is selected to discover the clusters.

Figure 1.4 Clustering in Distributed Data Mining
Here, an approach for distributed data clustering, based on a structured P2P network architecture is introduced. The motivation behind DDM is to achieve flexible DDM model that can be designed which are suitable to different scenarios. The proposed model is a hierarchically distributed clustering for P2P networks traffic characteristics. It involves a hierarchy of P2P neighborhoods, in which peers in each neighborhood are responsible for building a clustering solution, using P2P communication, based on the data they have to access.

In addition to this, more recent contributions on DDM are fueled by advances in hierarchically distributed computing platforms. The distributed nature of the DDM makes a compromise of trade off between accuracy and scalability. This motivates further to improve the cluster model to distributed data clustering context for traffic analysis of hierarchical P2P network architecture. It provides better accuracy and scalability even for high volume network traffic data streams when monitored at various scenarios by providing higher performance level for the metrics towards F-measure, Precision. Distributed hierarchical clustering approach maintained a balanced network traffic distribution in P2P networks with efficient partitioning and minimized traffic entropies. The metrics true positive rate, false positive rate were measured using the traffic entropy which is the measure of abnormality rise in the improper distribution of high volume data streams.

### 1.2 RESEARCH FOCUS

This research work focuses on the problem of mining high volume distributed network traffic data in P2P networks by discovering useful relationships, and grouping of volume-ness input traffic data based on the nature of data like File Transfer Protocol (FTP), Transmission Control Protocol (TCP). The proposed work is a high speed content classification system which classifies traffic data flow based on the hierarchical nature of
peer networks. Clustering of unclassified flows is the focus of this research work on high stream traffic data in P2P networks. The proposed approach also presents an efficient clustering means to analyze the traffic data streams with two different natures referred to as symmetric and asymmetric nature.

This research work is concerned with the problem of mining network traffic data and groupings in large collections of data. Many algorithms related to mathematical transformations have been proven effectively at reducing the content of multilingual and transformation of unstructured data into a vector. Such techniques are highly required in fields which consist of information explosions, network traffic analysis and bio-informatics. In response, traffic mining methodology is being extended to improve performance and scalability.

As a part of an ongoing research project, a novel algorithmic approach for extracting traffic data from voluminous data streams has been developed. The approach is applicable to internet data servers, which can be automatically identified and converted into a common structure. Report on an extension of the traffic data clustering work, which clusters data flow hierarchically is based on the nature requested by users (demand, load). The new method represents the streaming hierarchical clustering of data streams, which can be seen as a subfield of the nascent discipline of “streaming AI”, “evolutionary clustering”, or “AI in hardware”.

The proposed system is a High Speed Content classification system that works in three stages to classify flows of TCP traffic. A TCP flow is half of a TCP conversation. A connection from a client to a Mail server with Simple Mail Transfer Protocol (SMTP) is an example of a flow. The connection from the Mail server back to the client is considered a separate flow. It extracts words then builds a vector representation and then scores
against known nature of the data flow. The scores of the completed flows are passed out of the system for evaluation.

The base data list from stage one used for counting in stage two. Each dimension is represented by 4 bits. Counting for a dimension saturates at 15. When the flow of the data stops completion the count vector which represents the flow is forwarded on to stage three. In the third stage a vector representing a flow is scored against vectors representing known traffic data. The vectors representing the data are also referred to as the Score Table (ST) and each and every time are reconfigurable at run time. The ST is derived from a set of traffic data streams. The output of the system is a set of scores and the count array of the flow.

Evaluation of the scores determines the classification of the flow against the known traffic data. However, simply classifying the flow as the data with greatest score is not adequate. A forced classification of all flows will be undesirable in most of the applications. A threshold provides a confidence level to the classification of flows. First, deep domain knowledge is needed to train and update the classifiers at regular intervals. Second, only specific applications are identified by the classifiers whereas the others are labeled as “unclassified”. Any traffic data, that is not classified, is considered unknown to the system. Clustering these unclassified flows is the focus of the Streaming Clustering work in this thesis along with traffic matrix estimation.

The importance of cluster analysis can be stated as follows:

i. Clustering analysis is based on similarity cluster objects.

ii. An object in a single cluster can be derived by using a set of measurements to other objects of clusters.
Clustering algorithms can be used to measure user behaviours in a network.

The users of network are grouped into clusters based on the similarity of their patterns.

Based on the formation of clusters, the network traffic is reduced for predicting the flow.

1.2.1 Traffic Matrix

In traffic matrix mode the Figure 1.5 shows the traffic requirements between every single source and destination node pair in the network. The traffic requirements are stored as a two dimensional matrix containing the directed traffic for each node pair. The source is always one node. If the end is one node it is a point-to-point relation. If the end is a group of nodes it will be described a defined traffic relation of a point-to-multipoint communication. The traffic matrix includes traffic for point-to-point and point-to-multipoint relations.

The traffic matrix represents the high involvement of the volume of traffic which flows between the source and destination group pairs in a network. In the internet terminology the nodes can refer to Points-of-Presence (PoPs), routers or links. In the existing Internet Protocol (IP) backbone networks, acquiring accurate estimates of traffic matrices is highly impossible. With the help of the knowledge provided by the traffic matrices a greater number of traffic engineering tasks could be greatly improved. Due to the above mentioned problem large number of network operators have identified and processed the need for the development of practical methods in order to obtain accurate estimates of traffic matrices.
Some of the real time applications using traffic matrix are logical topology design, forecasting, configuring the routing protocol and giving provision for service agreements. Using the existing current state of flow monitoring equipment the evolvement of direct measurement of traffic matrices across an entire ISP network is considered too cumbersome in terms of communication and computation overhead.

The research work presents an algorithm, called K-family for clustering of traffic data streams, to enhance the K-means paradigm to categorical domains. The algorithm has demonstrated a very good classification performance when tested with the well-known traffic data stream logs which is obtained from University of California, Irvine (UCI) or Machine learning repository.

![Traffic Matrix Representations](image)

**Figure 1.5 Traffic Matrix Representations**

1.3 CLUSTERING OF HIGH VOLUME DATA STREAMS

The recent advancement in hardware area has motivated to design the transactions of record automatically. This process leads to huge amounts
of data that grow rapidly at an unlimited rate. The data processing involved in it is referred to as data streams. Data stream is an ordered sequence of data records which is in structured form. The data streams possess certain characteristics such as fast arrival rate, ordered temporally, evolves over time, and unbounded nature. The arrival rate of data stream is thousands of data records per second, the concepts related to data streams evolve at varying rates over time. As it is unbounded it is highly infeasible to store all of its corresponding records in secondary storage due to the high dimensionality of data involved.

The area of data stream has been widely researched in current years due to the large number of relevant applications. Research work conducted on data streams clustering made an assumption in such a way that the cluster to be calculated for the entire data stream. The methods involved simply to view the data stream clustering of one-pass clustering algorithms. However such a task may be useful in many research areas as clustering problem needs to be defined carefully related to data stream.

The data stream can be viewed as an infinite process consisting of data which instantly changes over time. Due to this, the clusters may also change rapidly with time. The nature and working of the clusters may differ with both the time at which they are calculated and the time horizon over which they are measured. For example a user may wish to observe the clusters that occur in the last month, last year and so on. Those clusters may be drastically different. Hence, clustering algorithm that deals with the data stream must provide the flexibility to measure clusters over different time periods in an interactive manner.

The main objective of this work is related to data clustering which is a process related to data mining problem. High dimensional datasets are not uniformly distributed. As the high dimensional data cannot be classified,
clustering provides an attractive advantage as it is easily adapted to changes in the data and hence used to identify features that distinguish different clusters. Due to the involvement of high dimensionality clustering, the data set uncovers the distribution patterns. Subsequently data clustering provides with the method of partition, unlabeled data records from a high dimensional dataset into labeled clusters which is also called as classification.

The problem of high volume data streams has gained importance in recent years because of advances in hardware technology. These varied advances have made it easy to store, record and retrieve numerous transactions and activities in everyday life in a more automatic way.

The pervasive presence of high volume data streams in a more number of practical domains has generated a lot of research in this area. One of the important issues that have to be addressed is clustering of high volume data stream domain. With the advent of data summarization and detection of outlier systems, the problem of clustering is highly significant for the high volume data stream domain.

There is an extensively growing need for efficient algorithms to identify important trends and anomalies in network traffic data. For example, network administrators need to understand client behavior in order to plan the overall capacity of the network. Due to the increase in the network capacities traffic analysis tools face the problem of scalability that involves two parameters. The parameters are high packet arrival rates and absolute memory. Here, a hierarchical clustering technique is presented for detecting and measuring significant patterns regarding traffic flow. In general, this research work presents a novel way of exploiting the hierarchical structure of traffic attributes, such as IP addresses, in combination with categorical and numerical attributes.
A key challenge in clustering high dimensional network traffic data is the requirement to deal with numerous types of attributes such as numerical attributes with real values, categorical attributes with ranked or unranked nominal values and attributes with hierarchical structure. For example, byte counts are in the form of numerical, protocols are designed categorical and IP addresses have an organized hierarchical structure. The main issue for these schemes is how to represent a function consolidating hierarchical attributes to find meaningful clusters.

![Diagram](image)

**Figure 1.6 Clustering of high volume data streams**

This research work proposes a hierarchical approach for the sake of clustering network traffic data that exploits the hierarchical structure present in real life dataset as shown in Figure 1.6. In considering network traffic, a hierarchical relation between two IP addresses can represent flow of traffic. The attributes represented using the hierarchical representation of data provides more meaning to a form a general cluster, which can reflect a trend
in flow of traffic. A common scheme to consolidate such hierarchical attributes using the distance function of the clustering algorithm is proposed.

1.3.1 Clustering

The clustering problem is defined as follows, for a given transaction set of data points, the basic idea is to partition them into one or more groups of similar objects, where by the approach of similarity is defined by a distance function.

Traditional algorithms on clustering high data streams presume that the clusters are to be calculated over the entire data stream. Those methods simply view the data stream clustering problem as a measure of single pass clustering algorithms. However, such a method may be useful in many applications and a clustering problem needs to be evaluated carefully in the area of data streams. This is because a high volume data stream should be viewed and integrated as an infinite process consisting of data which continuously evolves or differs with time.

Depending on various clustering formulations that are mainly based on minimizing a formal objective function, perhaps the most widely used and studied for high dimension of data involved is K-means clustering. Clustering based on K-means is widely related to a number of other clustering methods and location problems. This method includes the Euclidean K-medians distance, where the main aim is to minimize the sum of distances to the nearest center, where the objective is to minimize the maximum distance from every point to its closest center. The K-means problem can be solved using the heuristics technique using simple iterative scheme in order to find a locally minimal solution. This algorithm is often called as K-means algorithm or K-means clustering algorithm.
1.3.2  K-Means Clustering

Algorithms can be classified into supervised or unsupervised learning algorithms. The clustering problem is designed using K-means which is one of the simplest unsupervised learning algorithms. The procedure is described as follows. It is classified using a given data set through a certain number of clusters. It is assumed as K clusters by way of fixing a priori value. The main concept is to derive K centroids defining one for each cluster. These centroid values should be placed in a manner such that it causes different results. So, the optimal choice is to keep them as much as possible far away from each other as shown in Figure 1.7.

![Figure 1.7 Traditional K-means clustering](image)

The second phase consists of taking each point which belongs to a given data set and combines it to the nearest centroid. When no point in the given data set is pending this results, in the completion of first step and in this way clustering is performed. At this juncture, need to re-calculate K new centroids of the clusters resulting from the previous step. Based on the K new centroids obtained a new processing has to be performed in between the same data set points and with the nearest new centroid.
A loop has been generated. As a result of this iteration, K centroids change their location step by step until no more changes are done. In other words centroids do not move anymore and remains fixed. Finally, this K-means algorithm aims at minimizing an objective function.

1.4 EXISTING SOLUTION

The existing work is concentrated on clustering network traffic called Echidna, a distance-based clustering scheme to summarize the main types of traffic flows that are observed in a given network. The algorithm takes each record and iteratively builds a hierarchical tree of clusters called a Cluster Features (CF) tree. As the tree is built, each record is inserted into the closest cluster using a combined distance function for all attributes, as described in the previous section.

The radius of a cluster determines either a record should be absorbed into the cluster or the cluster should be split. Once the cluster tree has been generated, a summarized report is created by applying heuristics to reduce the number of significant nodes even further to create a concise and meaningful report. All of these require one additional pass over the cluster tree.

1.4.1 Gaps in Existing Solution

The two main parameters in Echidna are to create a concise and meaningful report. All of these require one additional pass over the cluster tree. One choice for the threshold parameter T for the cluster radius is to set it to zero, and then, as memory runs out, increase T to reduce the size of the tree. As T is reduced, the clusters become more compact allowing less variation in records, which results in more clusters.
The choice of \( T \) is a trade-off between smaller radius values and higher memory requirements due to the increased number of clusters. Another parameter to choose is the branching factor \( B \). Since the branching factor determines the size of the tree and each node carries an overhead in terms of memory, it is tempting to choose a higher value. However, the Balanced Iterative Reducing and Clustering using Hierarchies (BIRCH) algorithm needs to search within each node to find the closest matching cluster, a higher branching factor will require more time for searching the closest cluster.

### 1.5 RESEARCH OBJECTIVES AND ORIGINAL CONTRIBUTIONS

#### 1.5.1 Objectives

The objectives of the research work include the following:

i. To verify traffic consistency and to introduce adaptive hierarchical partitioned clustering approach on high volume data stream in peer to peer network.

ii. To achieve accuracy for cluster formation and scalability for various network heights.

iii. To handle the dynamic nature of P2P network traffic and to allow nodes to join and leave the network by maintaining a balanced network in terms of partitioning and network height.

iv. To achieve a better computational efficiency for K-modes P2P Clustering algorithm.
1.5.2 Contributions

The contributions to the present work include the following:

i. Flexible to accommodate regular P2P networks with dynamic peer nodes (Partition Cluster Approach (PCA)).

ii. Adapt modularized networks data traffic through neighborhood (Hierarchical Cluster Approach (HCA)).

iii. Allows privacy within neighborhood boundaries.

iv. Allow nodes to join and leave dynamically in P2P network.

v. Maintain balanced peer network in terms of partitioning and network height.

vi. Optimal of P2P network height is evaluated.

vii. Cluster quality is measured with quality index.

viii. Traffic matrix estimation is made on the clustered data traffic in multiple time zones

ix. Balance Load-Demand in the network traffic.

x. Verification of routing protocol configuration.

1.6 RESEARCH METHODOLOGY

Among the different types of internet applications, Peer-to-Peer (P2P) network is one of the highest traffic network characterized by, which allows a major portion of the users to establish communication with each other. In network traffic flow analysis there has been a shift of focus from simple volume-based analysis to network flow distribution-based analysis. This in turn directly access and download files from the peers’ machine, and share the resources which are inevitable. The basic requirement regarding the internet
traffic is the classification of network and the performance of high volume data streams in peer to peer network.

Earlier model dealt with clustering of data stream which handled network traffic that comprised of numerical data set by using K-means algorithm whereas our proposed model works well for both numerical and categorical data set. In this work, the deployment of clustering techniques is categorized into two types, hierarchical and partition clustering and to identify interesting traffic patterns based on the volume of data from network traffic in an efficient manner.

The work consists of construction and evaluation of hierarchical and partition clustering algorithm for clustering of network traffic data streams in P2P network. Moreover, a new clustering algorithm which combines hierarchical and partition clustering approach namely, AHPC algorithm is proposed. The performances of AHPC using K-modes algorithm were discussed using network traffic for both normal and high volume data streams. The experimentation demonstrated the flexibility of the model, showing that it achieves comparable quality to its centralized counterpart while providing significant speedup by manipulating the neighborhood size and height parameters. The model shows good scalability with respect to network size and hierarchy height, degrading the distributed clustering quality significantly.

1.7 ORGANIZATION OF THE THESIS

The Organization of thesis is as follows. Chapter 1 provides an introduction about the network traffic scenario in the internet due to its increased usage. Then briefing about the data mining techniques, used to monitor and handle network traffic in different volumes of data streams in
peer to peer networks. Focus of clustering model for traffic monitoring and regularization approach is also discussed.

Chapter 2 gives an overview of various literatures dealing with distributed peer to peer network and the influence of data traffic in maintaining the performance. Clustering principles deployed for traffic analysis in distributed network of hierarchical peers were discussed to know the pros and cons of the data mining for effective flow of data stream across the network. Literary works related to hierarchical partitioning of peer to peer network for data exchange were also analyzed.

Chapter 3 describes the need and demand for developing and implementing a parallel k-clustering algorithm, to cluster traffic data streams in distributed peer to peer networks. The partitioning of the P2P network in hierarchical fashion is presented to effectively maintain traffic data streams. The efficiency of cluster mining for traffic analysis is interpreted with experimental simulation done on extracted traffic data streams from internet service providers. Cluster centroid is utilized to determine the distance between each data traffic clusters and its data size variance. The k-cluster algorithm seeks to minimize the inner cluster distance.

Chapter 4 discusses the effect of clustering of high-volume data streams in network traffic for normal data flow in the distributed hierarchical peer to peer network. It also analyzes traffic characteristics and network performance estimation of the data flow in terms of load balance and demand appropriation. Partitioning of peer networks with hierarchical distribution of the peer nodes are evaluated to balance the traffic imbalance generated in the network.

Chapter 5 analysis the performance of distributed hierarchical partitioning of peer nodes and the clustering efficiency of high volume traffic
data stream using hierarchical and k-cluster families. The estimates of cluster object cohesiveness is measured to identify the quality of traffic maintenance achieved through our work. The performance comparison of the proposed hierarchical clustering to existing k-means cluster in peer to peer network traffic is evaluated.

Chapter 6 presents the result and discussion of the proposed model of clustering scheme to effectively handle high volume network traffic data streams with efficient hierarchical partitioning and appropriate peer node distribution. It provides discussion about the effect of network size on clustering accuracy, scaling the hierarchy height, quality of clustering at different levels within a single hierarchy, accuracy of distributed cluster summarization and volume-ness effect on the traffic data stream. Experimentation is conducted to estimate the cohesiveness of cluster object to measure the quality of distribute hierarchical cluster to existing k-means cluster.

Chapter 7 concludes this research work with the briefing about the effect of the proposed Adaptive Hierarchical and Partition Clustering (AHPC) model for traffic data streams. The influence of partitioning and distribution of peer nodes on the balancing the high volume streams are discussed. It briefs about the applicability along with its contribution to the internet traffic scenarios at varied dynamic network conditions. In addition future directions are suggested to further investigate this cluster model for other data mining tasks in various online applications i.e., share trading, forex and commodity markets.