Framework for aiding tool for Network Observer

1- Introduction

Internet grows rapidly; a lot of application and a lot of users have been involved in Internet community. Internet now (2004) is not the same at 1990 and it definitely will differs at 2025.Studying the behavior of the Internet at 2025 requires simulating the current status of the Internet and predicting the future of it. To study the current status of the Internet we need great efforts from different specialist in different fields like network architecture, database systems, Electronic commerce, data mining and other fields. So integrated work between different specialists will help in getting the vision of the Internet at 2025.From this point EVERGROW has been started. . The intent of EVERGROW is to:

- Invent methods and systems to build infrastructure for measurement, mock up and analysis of network traffic
- Start addressing the opportunities presented by the internet at 2005
- Manage the physical network, the distributed relationships in overlay networks, and the services that end-users care about
- Discover new ideas, which will shape the Internet of 2025

Central laboratory of agriculture expert systems (CLAES) has developed a prototype for network traffic analysis and mining (NTAM) to be part of what is called a "Virtual Network Observatory". The Virtual Network Observatory (VNO) is a distributed computing and data storage facility; at the same time an archive of network data. Network Traffic measurement generates great amount of data. VNO not only enables storing of raw data but also storing of information, and knowledge discovered from this data.

In order to build NTAM prototype, we need data that represent a sample of network traffic data. In the next section, we will describe the sample data that has been used. In section 3, some tools that have used for network data analysis are reviewed. In section 4, we will show how the captured data will be prepared for analysis and mining. In section 5, we will describe the current system architecture and data flow. In section 6, we will demonstrate some of the experiments results of data-analysis and mining.

2. Data collection

We will use data that were collected by M2C repository "M2C is a project sponsored by the Dutch Telematica Institute". The repository can be found at this location: http://m2c-

a.cs.utwente.nl/repository/. The repository represent the measurement for network traffic, the measurements are performed by capturing the headers of all packets that are transmitted over the (Ethernet) “uplink” of an access network to the Internet, as outlined in Figure 1.

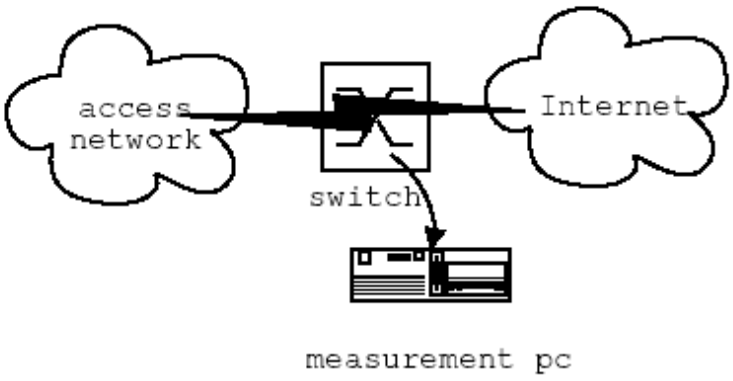

Figure 1: Measurement Setup

The raw data consisted of packet level transmission data including source and destination IP address and ports; flags, acknowledgements and packet sequence numbers; and window, buffer and optional information.

In M2C project the tool that has been used for capturing network traffic is the popular tcpdump tool. which can be downloaded from the Internet for free.

Tcpdump is the most used tool for network monitoring and data acquisition. Tcpdump uses libpcap (Packet Capture library), a system-independent interface for user-level packet capture. Tcpdump provides a standard packet capture interface, a common dump format, basic packet decoding features and can filter packets in various ways.

Platform: Unix (WinPCAP/WinDUMP is the porting to the Windows platform of libpcap/tcpdump)
Business Model: Open Source (BSD License)
Internet: http://www.tcpdump.org

3. Data Analysis tools for network traffic.

Capinfo

Capinfo is a tool for displaying statistics about network traffic from files saved with tcpdump or snoop.

Platform: Unix
Business Model: Open Source
Internet: http://tcpreplay.sourceforge.net

CoralReef

CoralReef is a comprehensive software suite developed by CAIDA to collect and analyze data from passive Internet traffic monitors, in real time or from trace files.
Platform: Unix
Business Model: Free for educational, research and non-profit purposes (see file Copyright in source package for details)
Internet: http://www.caida.org/tools/measurement/coralreef/

Ethereal

Ethereal is a GUI network protocol analyzer. It can examine data from a live network or from a capture file on disk. You can interactively browse the capture data, viewing summary and detail information for each packet.

Platform: Unix/Windows
Business Model: Open Source (GPL)
Internet: http://www.ethereal.com

Microsoft Network Monitor

Network Monitor captures network traffic for display and analysis. It allows you to perform tasks such as analyzing previously captured data in user-defined methods, extracting data from defined protocol parsers, and analyzing real-time traffic on your network.

Platform: Windows
Business Model: Commercial
Internet: MSDN Library: About network monitor 2.0

Tcptrace

tcptrace tool written at Ohio University, for analysis of TCP dump files. It can take as input the files produced by several popular packet-capture programs, including tcpdump, snoop, etherpeek, HP Net Metrix, and WinDump.

Platform: Unix/Windows
Business Model: Open Source (GPL)
Internet: http://www.tcptrace.org
Documentation: http://www.tcptrace.org/manual.html

Tcpflow

Tcpflow is a program that captures data transmitted as part of TCP connections (flows), and stores the data in a way that is convenient for protocol analysis or debugging.
A program like *tcpdump* shows a summary of packets seen on the wire, but usually doesn't store the data that's actually being transmitted. In contrast, tcpflow reconstructs the actual data streams and stores each flow in a separate file for later analysis.

Platform: Unix
Business Model: Open Source (GPL)
Internet: http://www.circlemud.org/~jelson/software/tcpflow/

4. Preparing data for analysis and mining.

The row data should be filtered to extract the main features that will affect data mining and data analysis process. From our observation, studying the nature of data and from the feedback of network engineer expert. Initially we choose the following features, [Source Address, Destination address, Source Port, Destination Port, packet length, and protocol], we well use these features in analysis of Network Traffic.

And in the future according to the view of network expert we may add another features. Further we may use annotated network traffic data and hope if they can help us in detecting and specifying the most correctly analysis feature rapidly, and reducing the time of data preprocessing for mining process.

5. Proposed System

The Proposed system consists of two phases, analysis phase, and mining phase. For analysis phase the expected output is visualized statistical analysis for network traffic data, and network data represented in a format that will be used in mining phase. For mining phase the expected output is a set of discovered knowledge from network data. The output from mining phase may be classification, Association Analysis, Sequencing Discovery, Regression, Time series. Figure 2 displays prototype for the proposed system.

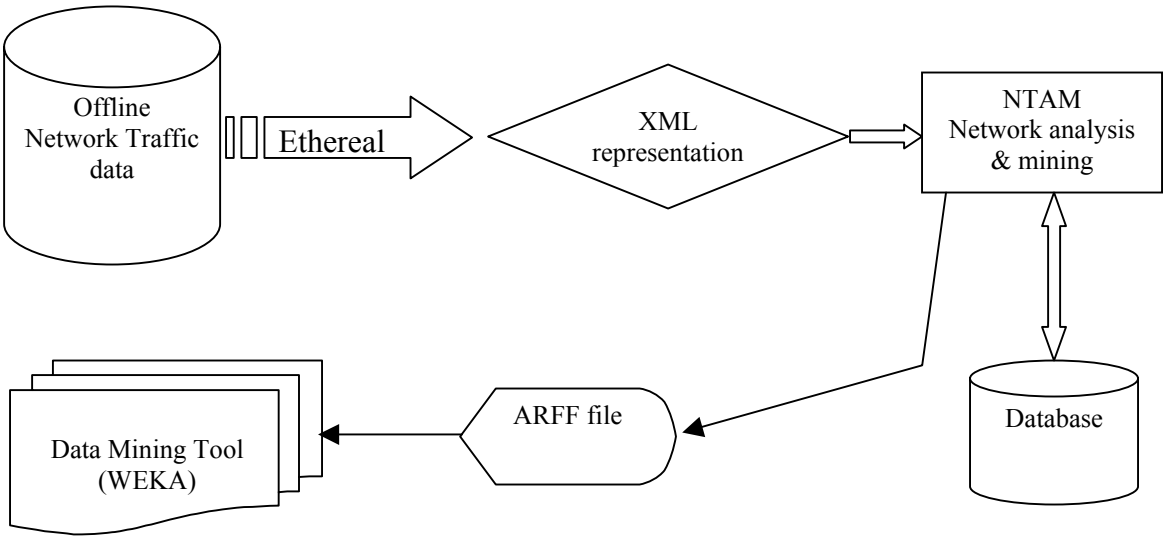

Figure 2 Framework for network analysis and mining.

As illustrated in figure 2, we use Ethereal "Network traffic analysis tool " for converting the tcpdump data into XML format, to facilitate the process of manipulating data. We use our prototype tool, Network Analysis & Mining Tool 'NTAM' for parsing the XML data, then NTAM will export interested feature of network data into database, instead of flat files to access data faster. After that NTAM can be used to export the data to ARFF file format to use it in mining phase.

An ARFF (Attribute-Relation File Format) file is an ASCII text file that describes a list of instances sharing a set of attributes. The Machine Learning Project at the Department of Computer Science of The University of Waikato developed ARFF files for use with the Weka machine learning software.

We didn't lose much time in reinventing the wheel to write a source code for data mining tool, instead we used an open source data-mining tool called Weka.

Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization. It is also well suited for developing new machine learning schemes. Weka is open source software issued under the GNU General Public License. Weka expected input data in an ARFF format, because it is necessary to have type information about each attribute, which cannot be automatically deduced from the attribute values. Before you can apply any algorithm to your data, it must be converted to ARFF form.

After feeding the network data to Weka we try to fit suitable machine learning algorithms to the data according to data mining task (Association, clustering, classification).

6. Experimental results

In analysis phase, for each feature of network traffic data The NTAM generates some statistics, as number of distinct values, number of unique values, minimum value and maximum value. After that the NTAM generates Bar chart diagram that represent the feature using and frequent in the collected data.

Destination port Analysis

According to IANA " Internet Assigned Numbers Authority " Port number divided into three ranges: the Well Known Ports, the Registered Ports, and the Dynamic and/or Private Ports.

- The Well Known Ports are those from 0 through 1023.
- The Registered Ports are those from 1024 through 49151
- The Dynamic and/or Private Ports are those from 49152 through 65535

Figure 3 illustrates which ports in **well-known ports** that has been used and number of use.

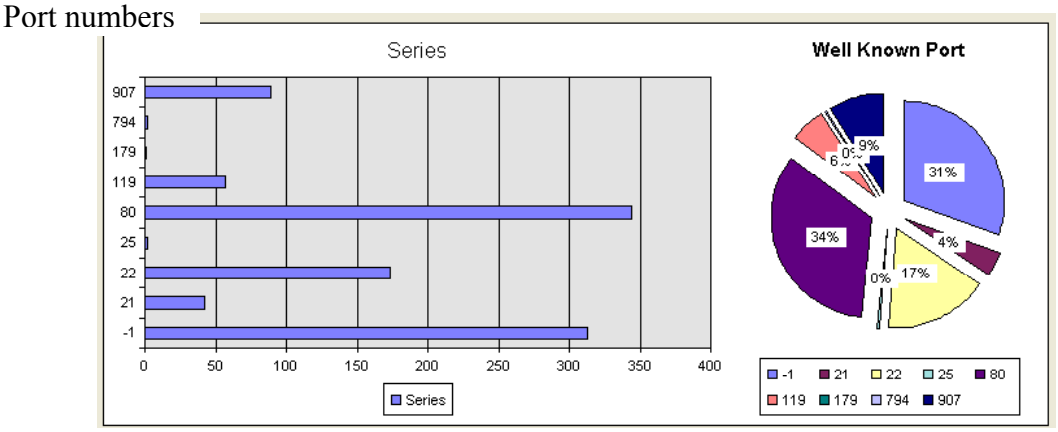

Figure 3 well-known ports for then collected packets

From this diagram Network observer will gives information about which destination port in well-known port that has been used and its number of use.

For example from this figure we observe that Port number 80 (World Wide Web HTTP) has the maximum occurrence and port 22 (SSH Remote Login Protocol) has next order in use. From this data network observer can observe that port number "-1" has been occurred more than 300 times, we set value –1 to behave the packet that has no destination port for ICMP protocol. Later on Network observer can find, which Source IP addresses almost, uses this port permanently. Also from analysis of destination port network observer can give a diagram represent **"registered Ports"** as in the figure 4

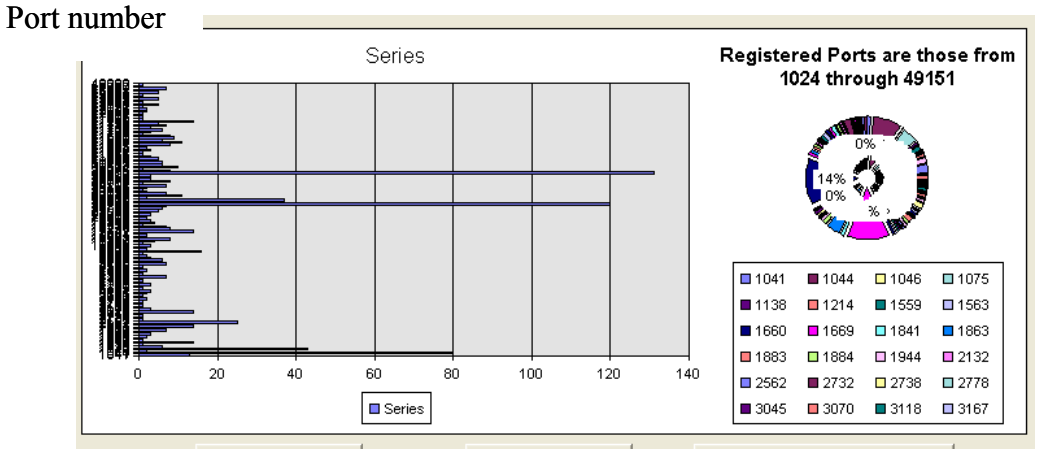

Figure 4 represent analysis of destination port for registered ports

Figure 4 illustrates that many ports has been used but with less frequent than the than the well-known ports.

Dynamic and/or Private Ports

Figure 5 illustrate analysis of " Dynamic and/or Private Ports "

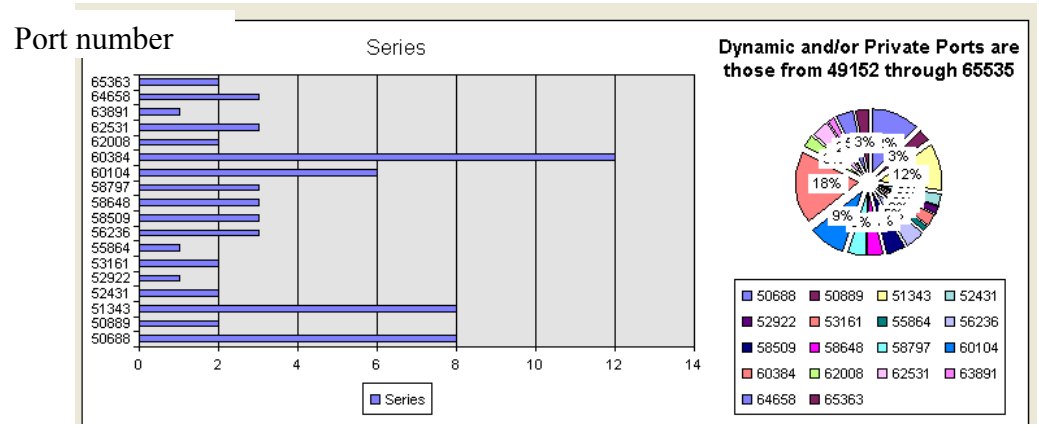

Figure 5 dynamic ports

If network expert compares figure 5 with figure 3 and figure 4, he can detect that, the number of destination port that has been used for dynamic port is less than well-known port and registered port. Further we observe that port number 60384 has the maximum number of use, but from IANA this is not assigned port, and unassigned port numbers should not be used. Network observer will be aware by which Source IP use this port and from the background knowledge it will decide if it intrusion or not. Each application uses a specific destination ports. So From analysis of used port we can predict which application that is mainly used. And this may help us in EVERGROW to get the vision about the application that will be used in Internet at 2025

Source port analysis

Analysis of Source port is the same as for destination port three diagrams depicts the IANA three types of port [well known port, registered port and dynamics port] as illustrated in the figure 6, 7,8.

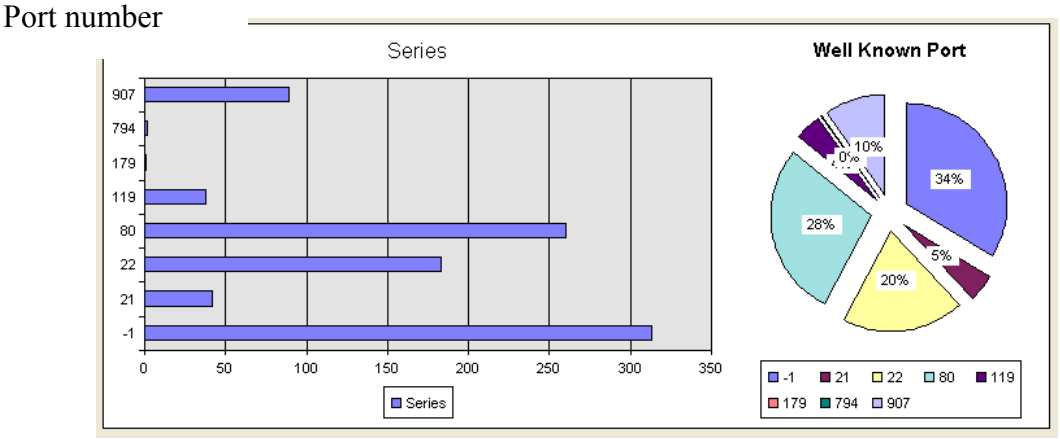

Figure 6 Well known ports that has been used by source port

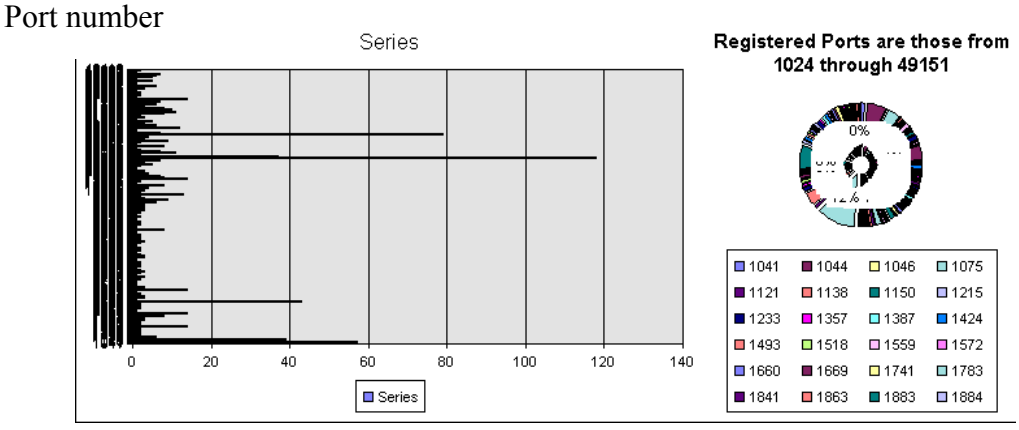

Figure 7 Registered ports that has been used by source port

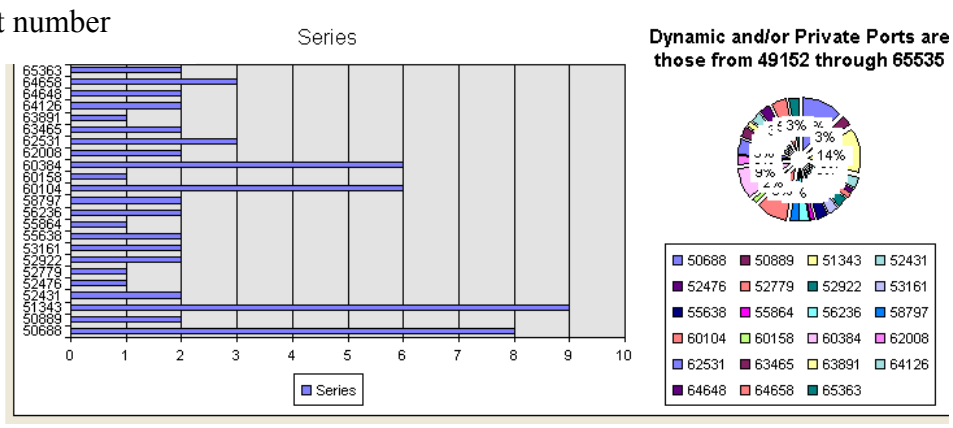

Figure 8 Dynamic ports that has been used by source port

Analysis of Source address and Destination address

Analysis for Source address and destination address is divided into two parts, part 1 is a diagram for the top 20 addresses, and part 2 is a diagram for all sources or destinations addresses.

The NTAM will help the network observer to see the top 20 IP (Source or destination) that has the most number of use, so according to network architecture and configuration. Network observer can view which Source or destination address is loaded, then from its previous knowledge of which application runs on this Source or destination address it can check if it is normal or not. Further it can recommend new configuration to the network .For example giving high priority for the loaded source address if this normal but if this abnormal it can prevent this source address from transmitting packets to the LAN.

Figure 9 illustrates which 20 Source Address has the great load in the collected packets.

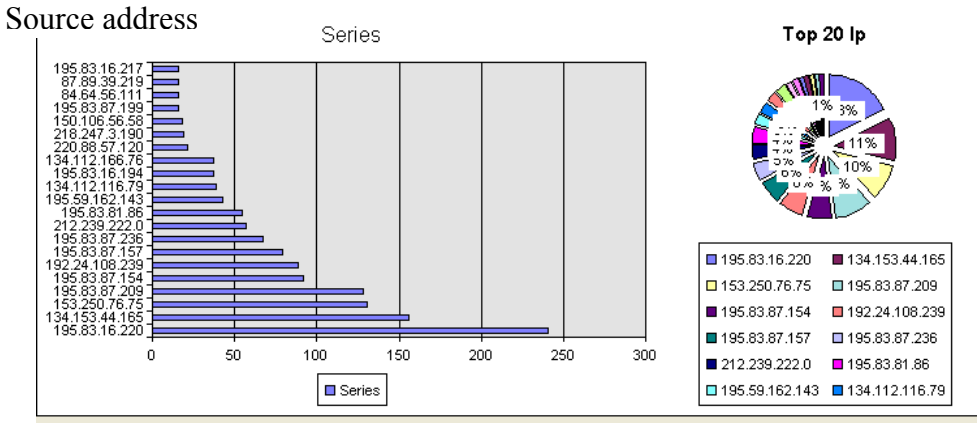

Figure 9 Top 20 source address

As illustrated in the figure IP (195.83.16.220) is most commonly used and it sends more than 200 packets from the collected packets.

Figure 10 displays a diagram represent all source addresses.

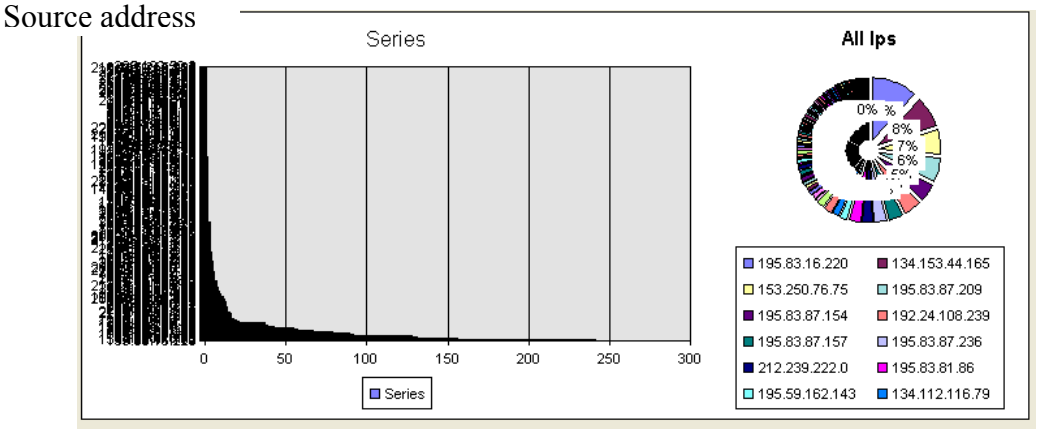

Figure 10 all source address

From this diagram we can detect that most source addresses send less than 10 packets and as it has been illustrated the top 20 Source addresses can be identified. This entire analysis figure will help network observer to produce information about enhancing network configuration and network performance.

Analysis of packet length

The analysis of packet length feature, will gives network observer the ability to knowing which packet length that is mainly used by he transmitted packets. The NTAM gives two-analysis diagram; Analysis of top 20-packet length, and analysis diagram for all packet length. Figure 11 displays analysis of top 20-packet length.

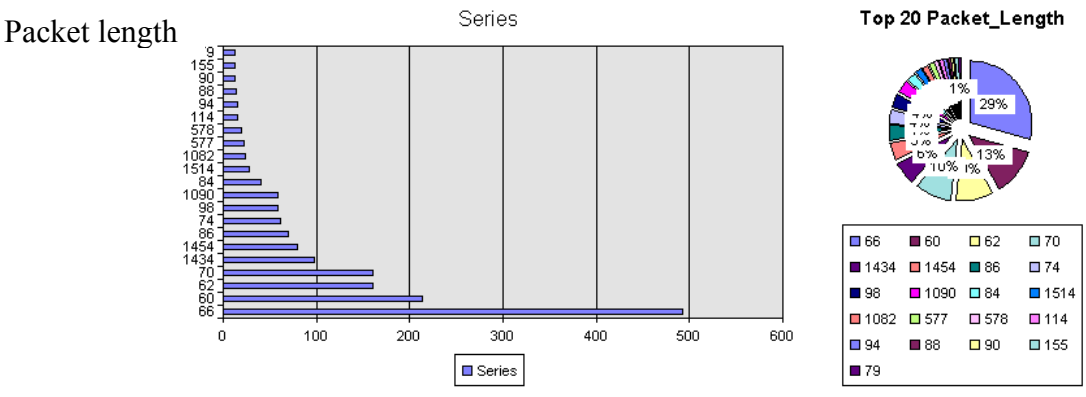

Figure 11 Top 20 Packet lengths

In figure 11 packets with length 66 bytes is mostly used by the colleted packets

From previous background knowledge of network observer it can decide if packet with length 66 byte is normal traffic or not, after that it can give information about which source address uses this packet length.

The NTAM also gives chart diagram for all used packets length. Figure 12 gives a view of all packet lengths that has been used.

Figure 12 gives analyses for all packet length.

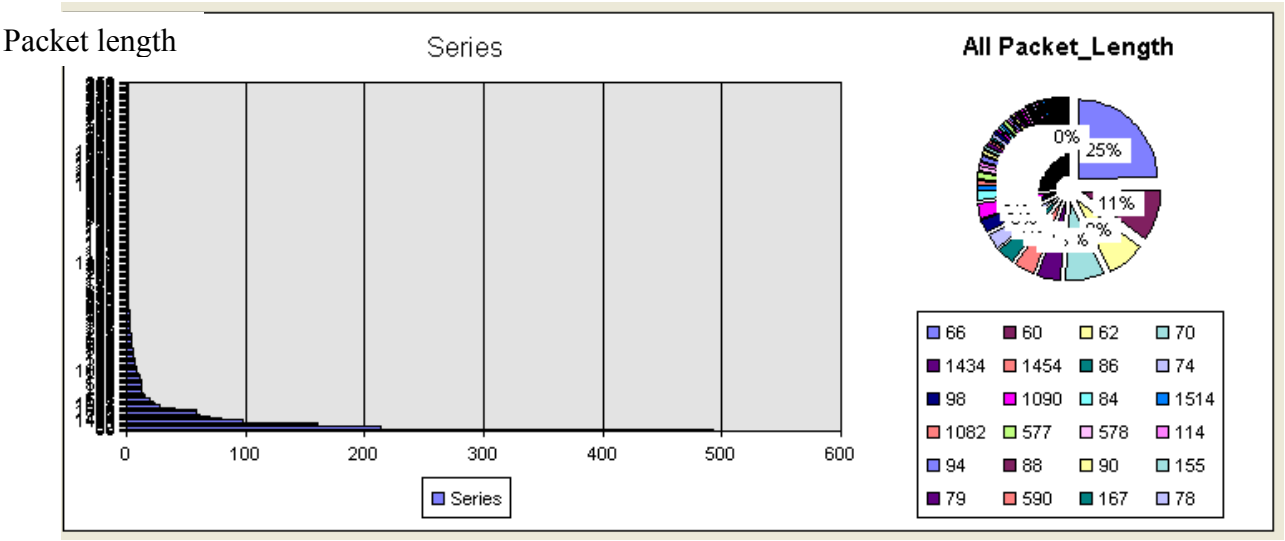

Figure 12 gives analyses for all packet length

Analysis of packets length may help network observer to detect new attacks for the network. For example network observer may observe the packets with specific length widely used by the network. And we assume that Network observer has background knowledge of packet length behaviors .So network observer can detect decide if the observed packet length is normal or abnormal.

Mining phase

In this phase we use Weka as a tool for data mining. The NTAM will generate files with ARFF format because Weka gets its input in ARFF file format. Weka generates Classification, clustering and association rules.

Mining Association rules between Source Address and destination port.

The task of association rule mining is to find certain association relationships among a source address and destination port. The association relationships are described in association rules. In association analysis the antecedent and consequent are sets of items (called itemsets) that are disjoint (do not have any items in common). Each rule has two measurements, support and confidence.

The support is simply the number of transactions that include all items in the antecedent and consequent parts of the rule.

The Confidence is the ratio of the number of transactions that include all items in the consequent as well as the antecedent (namely, the support) to the number of transactions that include all items in the antecedent.

In our experiment we use Weka's Apriori association rule algorithm, Apriori works with categorical values only. Therefore we will use ARFF file that contains source address and destination port in categorical format. The output of applying Apriori algorithm is in figure 13.

```
Apriori
=======

Minimum support: 0.05
Minimum metric <confidence>: 0.9
Number of cycles performed: 19

Generated sets of large itemsets:

Size of set of large itemsets L(1): 9

Size of set of large itemsets L(2): 2

Best rules found:

1. Destination_Port=38384 131 ==> Source_Address=153.250.76.75 131    conf:(1)
2. Source_Address=153.250.76.75 131 ==> Destination_Port=38384 131    conf:(1)
3. Source_Address=134.153.44.165 156 ==> Destination_Port=-1 155    conf:(0.99)
```

Figure 13 Outputs from Weka's Apriori

Description of Output

The minimum support, and Minimum confidence are 0.05, 0.9, and 2 large itemset. There are 3 best rules found, these rule will help network administrator to explore the relations between source address and destination port.

For example rule 3, if Source address 134.153.44.165.156 then destination port –1 with confidence 0.99. This means that Source address 134.153.44.165.156 always send packet which destination are not reachable. The packet with source address 134.153.44.165.156 may be intrusion.

Also use Wake's Apriori association rule algorithm, for mining association between **Source address** and **packet length**. And it detects the following rules

1. Source Address=134.153.44.165 156 ==> Packet_Length=70 154 conf:(0.99)
2. Packet_Length=70 160 ==> Source Address=134.153.44.165 154 conf:(0.96)

This result says that source address 134.153.44.165 always sends packets with length 70. Further mining associated relationships in overlay networks traffic; will discover new rules that may help Network observer in discovering new idea, which will shape the Internet of 2025.

Mining clustering in network traffic data

The task of clustering is to try to discover the inner nature of the network traffic data, and to divide the data into groups of similarity. Clustering is the partitioning of a data set into groups so that the points in the group are similar as possible to each other and different as possible from points in other groups. Weka provides five algorithms for mining clustering, EM, simpleKmeans, Cobweb, Farther First and make density based cluster.

We use The WEKA SimpleKMeans algorithm to perform the task of mining clustering. The WEKA SimpleKMeans algorithm uses Euclidean distance measure to compute distances between instances and clusters, it take the number of clusters as parameter in this example we choose 4 as number of clusters, number of clusters should be chosen correctly, network expert should specify the correct number. For example network data may be classified as normal and abnormal in this case number of classes will be 2, but here we assume that there exit four types of classes. Figure 14 illustrate part of the output from Weka' simpleKMeans algorithm.

```
Scheme:       weka.clusterers.SimpleKMeans -N 4 -S 10
Relation:     Network_Traffic
Instances:    2000
Attributes:   6
              Source_Address, Destination_Address, Source_Port, Destination_Port, Packet_Length
              Protocol

Cluster centroids:
Cluster 0
              Mean/Mode:  195.83.87.209 195.83.16.220 22 80 66 0x06
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 1
              Mean/Mode:  195.83.16.220 192.24.108.239 29769 22 66 0x06
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 2
              Mean/Mode:  153.250.76.75 195.83.87.157 80 38384 1434 0x06
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A
Cluster 3
              Mean/Mode:  134.153.44.165 195.83.87.236 -1 -1 70 0x01
              Std Devs:   N/A   N/A   N/A   N/A   N/A   N/A
Clustered Instances
0      1109 ( 55%)
1       414 ( 21%)
2       164 (  8%)
3       313 ( 16%)
```

Figure 14 outputs from Weka SimpleKmeans

Figure 14 shows the centroid of each cluster as well as statistics on the number and percentage of instances assigned to different clusters. Cluster centroids are the mean vectors for each cluster (so, each dimension value in the centroid represents the mean value for that dimension in the cluster). Thus, centroids can be used to characterize the clusters. For example, the centroid for cluster 1 shows that this

is a segment of cases representing source address 195.83.87.209, destination address 195.83.16.220, source port 22,destination port 80,packet length 66, protocol 0x06.

Clustering approaches for detecting outliers in large data sets for the purposes of fraud or intrusion detection appear in the literature, but these approaches are primarily based on ordered data. Knorr and Ng [1998] recently developed a distance-based clustering approach for outlier detection in large data sets. Ramaswarny, et al. [2000] defines a new outlier criterion based on the distance of a point to its kth nearest neighbor.

Mining classification in network traffic data

The objective of classification is to decide how new records of data will be classified. Machine learning software performs classification by learning discrimination rules from examples of correctly classified data. Weka provides a wide variety of classification algorithms. We use decision tree algorithm because they produce a result in a format familiar for human to understand it.

To perform a classification data we should have a set of training data. The 'quality' of the training data is one of the most important factors in achieving good classifier performance. The training data should be labeled; Classification requires previous knowledge from end user/analyst of how classes are defined. Unfortunately we haven't a real labeled network data. So we assume that data are labeled by packet protocol. In our example we take two analysis feature packet length and protocol. And weka builds a tree to explore relation between packet length and protocols this is a simple example of classification. Figure 15 illustrate output from wake's J48 classification algorithm.

```
=== Classifier model (full training set) ===
J48 pruned tree
------------------
Packet_Length <= 67: 0x06 (868.0/6.0)
Packet_Length > 67
|   Packet_Length <= 70: 0x01 (160.0/3.0)
|   Packet_Length > 70
|   | Packet_Length <= 96
|   |   | Packet_Length <= 87: 0x06 (228.0/20.0)
|   |   |   Packet_Length > 87
|   |   |   | Packet_Length <= 88: 0x11 (14.0/1.0)
|   |   |   | Packet_Length > 88
|   |   |   |   | Packet_Length <= 94
|   |   |   |   |   | Packet_Length <= 91
|   |   |   |   |   |   |   Packet_Length <= 90: 0x06 (14.0/3.0)
|   |   |   |   |   |   |   Packet_Length > 90: 0x11 (4.0)
|   |   |   |   |   |   |   Packet_Length > 91: 0x06 (16.0)
|   |   |   |   |   |   Packet_Length > 94: 0x11 (8.0/2.0)
|   |   Packet_Length > 96
|   |   |   Packet_Length <= 98: 0x01 (59.0/3.0)
|   |   |   Packet_Length > 98
|   |   |   |   Packet_Length <= 210
|   |   |   |   |   Packet_Length <= 194: 0x06 (165.0/26.0)
|   |   |   |   |   Packet_Length > 194: 0x11 (13.0/1.0)
|   |   |   |   Packet_Length > 210: 0x06 (451.0/9.0)

Number of Leaves  :   12
Size of the tree :    23
```

Figure 15 Outputs from Wake's J48 classification algorithm

Description of Output

Number of leaves of the tree 12, and size of the tree 23, the first rule in the discovered tree illustrate that all packet with length less than 67 uses 0X06 (TCP) Protocol. So if we build a model to classify new network traffic data. Each packet with length less than 67 bytes will be classified as TCP protocol

Wake's result includes a confusion metrics to evaluate classification result

From confusion metrics

Correctly Classified Instances	1916	95.8 %
Incorrectly Classified Instances	84	4.2 %

```
=== Confusion Matrix ===

   a    b    c    d    e   <-- classified as
 213    0    0    0    0 |    a = 0x01
   6 1670   11    0    0 |    b = 0x06
   0   57   33    0    0 |    c = 0x11
   0    8    1    0    0 |    d = 0x29
   0    1    0    0    0 |    e = 0x67
```

Conclusion

Network traffic analysis help network administrator to be aware of network traffic data. Data mining will give power to network traffic analysis tools .It will enrich analysis methods .we develop a prototype tool for offline network traffic analysis. And we apply our work to real network traffic data. A prototype gives network administrator a set of analysis chart describe the network data, network administrator analyze charts to get understand of nature traffic data.

Our vision in network traffic is to apply data mining in analysis of network data, so the prototype help in mining process by converting network data to ARFF format to use it by Data mining tools like weak Weak will help network analyst to discover patterns in network data.

Reference:

1. Kenjiro Cho, Koushirou Mitsuya and Akira Kato."Traffic Data Repository at the WIDE Project" USENIX 2000 FREENIX Track, San Diego, CA, June 2000.
2. Kenjiro Cho, Ryo Kaizaki and Akira Kato. "Aguri: An Aggregation-based Traffic Profiler In Proceedings of QofIS2001 (published by Springer-Verlag in the LCNS series). September 2001.
3. Barbara, D., N. Wu, and S. Jajodia [2001]. “Detecting Novel Network Intrusions Using Bayes Estimators”, Proceedings Of the First SIAM Int. Conference on Data Mining, (SDM 2001), Chicago, IL.
4. Knorr, E. M., and R. T. Ng [1998]. “Algorithms for Mining Distance-Based Outliers in Large Datasets”, VLDB'98, Proceedings of the 24th Int. Conference on Very Large Databases, Aug 24-27, 1998, New York City, NY, pp. 392-403.
5. Ramaswarny, S., R. Rastogi, and K. Shim, [2000]. “Efficient Algorithms for Mining Outliers from Large Data Sets”, Proceedings of the ACM Sigmod 2000 Int. Conference on Management of Data, Dallas, TX. http://www.iana.org/assignments/port-numbers.

http:/www.comlab.uni-rostock.de/research/tools.html

http://corky.net/2600/data-networks/packet-sniffer.shtml

http://m2c-a.cs.utwente.nl/tools/