International Journal of Computer Networks and Applications
Volume 1, Issue 1, November - December (2014)

RESEARCH ARTICLE

A Dynamic Approach of Malicious Node Detection for Internet Traffic Analysis

Ravi Shankar P
Computer Science and Engineering,
MVGR College of Engineering, Andhra Pradesh, INDIA
ravishankar.paluru@gmail.com

Santosh Naidu P
Computer Science and Engineering,
MVGR College of Engineering, Andhra Pradesh, INDIA
amsan2015@gmail.com

Abstract – Optimizing the internet traffic is always an important research issue in the field of network traffic classification, although various approaches available for minimizing the traffic over heads during the network traffic, they are not optimal. In this paper we are proposing an optimized classification approach for internet traffic by analyzing the behavior of the nodes for allowing or dis connection of the incoming node by computing the posterior probabilities of the factors with respect to the node.

Index Terms – Port Based classification, Payload-Based Classification, SMTP, IMAP.

1. INTRODUCTION

Different researchers proposed different approaches for classifying the network traffic or identify the unnamed node either by clustering, signature and classification. In clustering, we group the almost the same type of data objects based on the between the data objects, by selecting the first data points or centroids [1].

A parallel Signature based way of doing things proposed by the some researchers, in this approach, they are analyzing the network traffic with processing. In this method, complete rule groups are spread across nodes. There is potentiality of using packet duplicator to send each and every packet to each node for processing, or just by using traffic-divider/splitter each packet is routed to the respective node. In that case rules are agglomerated into rule groups based on source and destination ports. So it's always better to use traffic divider/splitter which could route packets to the respective nodes more efficiently.

1.1. Port Based classification

The simplest way to classify Internet traffic is by using UDP or TCP port numbers. The reason is that some traffic uses well known port numbers, and the port numbers can be found on Internet Assigned Numbers Authority (IANA). For example, HTTP uses port 80, POP3 uses port 110, and SMTP uses port 25. We can set up rules to classify the applications that are assigned to the port numbers.

However, many researches claim the port number- based classification is not (good) enough. Moore and Papagiannaki claimed the (quality of being very close to the truth or true number) of port-based classification is around 70% during their experiment. More than that, Madhukar and Williamson claimed in their research that the misclassification of port-based classification is between 30% and 70% [1]. The main reason for choosing static port numbers is to make the packet more able to go through the server firewalls. Many recent applications try to avoid the detection of firewall by hiding the port numbers. Some of the other applications use energetic/changing port numbers instead of static ones. And servers which share the same IP address will use un-standard port numbers [1].

1.2. Payload-Based Classification

Another approach to classify packets is to analyze the packet payload or use deep packet inspection (DPI) technology. They classify the packets based on the signature in the packet payload, and it has been advertised/talked well about as the most (very close to the truth or true number) classification method, with 100% of packets correctly classified if the payload is not (turned into secret code) [3]. The signature is (like nothing else in the world) strings in the payload that distinguish the target packets from other traffic packets. Every rules of conduct has its clear/separate way of communication that is different from other rules of conduct. There are communication patterns in the payload of the packets. We can set up rules to analyze the packet payload to match those communication patterns in order to classify the application. For example, according to [3], "MAIL FROM","RCPT TO" and "DATA", as in Figure 1, are the commands that appear in the payload of SMTP packets.

Therefore, we can create rules to match the plain text in the packet payload to classify SMTP packets. The problems include: users may (turn into secret code) the payload to avoid detection, and some countries forbid doing payload inspection

RESEARCH ARTICLE

to protect user information privacy. What's more, the classifier will experience heavy operational load because it needs to constantly update the application signature to make sure it contains the signature of all the latest applications.

2. LITERATURE SURVEY

In 2006, [2] Madhukar A and Williamson C focused on network traffic measurement of Peer-to- Peer (P2P) applications on the Internet. P2P applications (probably) make up/be equal to a big proportion of today's Internet traffic. However, current P2P applications use several hiding ways of doing things, including energetic/changing port numbers, port hopping, HTTP pretending, chunked file moves, and (turned into secret code) payloads. As P2P applications continue to change strong and healthy and effective methods are needed for P2P traffic identification. They compared three methods to classify P2P applications: port-based classification, application-layer signatures, and transport-layer analysis. The study uses network traces collected from the University of Calgary Internet connection for the past 2 years. The results depicts that port-based analysis is ineffectual which is unable to identify 30-70% of today's net-traffic. Application signatures may not be possible for legal and technical reasons. The transport layer method seems predicting, which provides strong and healthy means to test/evaluate group P2P traffic.

Briefing about SMTP, SMTP, a process can move mail to another process on the same network or to some other network via a relay or gateway process (easy to get to, use, or understand) to both networks. In this way, a mail message may pass through some intermediate relay or gateway hosts on its path from sender to final/very best receiver. The Mail exchanger of the domain name system are used to identify the appropriate next-hop destination for a message being moved.

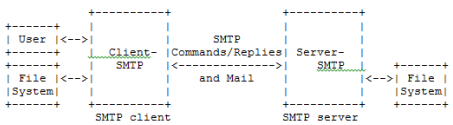

Figure 1 payload of SMTP packets

When an SMTP client has a message to transmit, it establishes a two- way transmission channel to an SMTP server. The responsibility of an SMTP client is to move mail messages to one or more SMTP servers, or report it's not doing so. The way a mail message is presented to an SMTP client, and how that client decides/figures out the domain name(s) to which mail messages are to be moved is a local matter, and is not talked to/looked at. Sometimes, the domain name(s) moved to, or decided/figured out by, an SMTP client will identify the final destination(s) of the mail message. In other cases, common with SMTP clients connected with putting into uses of the POP or IMAP rules of conduct, or when the SMTP client is inside a (far apart from others) transport service (surrounding conditions), the domain name decided/figured out will identify an intermediate destination through which all mail messages are to be relayed. SMTP clients that move all traffic, regardless of the target domain names connected with the individual messages, or that do not maintain waiting lines for retrying message transmissions that at first cannot be completed, may otherwise go along with this specification but are not considered fully-capable.

The way an SMTP client, once it has decided/figured out a target domain name, decides/figures out the identity of an SMTP server to which a copy of a message is to be moved , and then performs that move , is covered by this document. To affect a mail move to an SMTP server, an SMTP client establishes a two-way transmission channel to that SMTP server. An SMTP client decides/figures out the address of an appropriate host running an SMTP server by resolving a destination domain name to either an intermediate Mail exchanger host or a final target host.

An SMTP server may be either the final/very best destination or an intermediate "relay" (that is, it may assume the role of an SMTP client after receiving the message) or "gateway" (that is, it may transport the message further using some rules of conduct other than SMTP). SMTP commands are created by the SMTP client and sent to the SMTP server. SMTP replies are sent from the SMTP server to the SMTP client in response to the commands.

Coming to [4] BRO INTRUSION DETECTION SYSTEMS, Bro is an open source UNIX based network monitoring framework. Often compared to Network invasion detection systems (NIDS), Bro can be used to build a NIDS but is much more. Bro can also be used for collecting network measurements, conducting (extremely careful, crime-solving, science-based) investigations, traffic base lining and more. Bro has been compared to tcpdump, Snort, netflow, and Perl (or any other scripting language) all in one. It is released under the BSD license.

Bro [5] is a stand-alone system for detecting net- work intruders in real-time by unemotionally monitoring a net-work link over which the intruder's traffic transits. We give a summary of the system's design, which draws attention to high- speed (FDDI-rate) monitoring, real-time notice/communication, clear separation between and policy, and extensibility. To accomplish these ends, Bro is categorized into 2 types:

A. Event Engine

Reduces the kernel-filtered network-traffic stream into a series of higher level events.

B. Policy script Interpreter

Interprets event handlers written in a language used to express a site's security policy.

RESEARCH ARTICLE

Event handlers can update state information, make/create new events, record information to disk, and create real-time notices/communications via syslog. We also discuss some attacks that attempt to destroy (by sneaky actions) unemotional monitoring systems and defenses against these, and give details of how Bro analyses the six applications combined into it so far: Finger, FTP, Portmapper, Ident, Telnet and Rlogin. The system is publicly available in source code form.

In 2003, [6] Azzouna N.B and Guillemin F talked about/said in a Worldwide Telecommunications Conference, 2003. GLOBECOM '03. IEEE those Measurements from an Internet spine-related link carrying TCP traffic towards different ADSL areas are carefully studied. For traffic analysis, we put into use a flow based approach and the popular mice/elephants two-part thing, where mice refer to short traffic moves and elephants to long moves. The originality of the reported experimental data, when compared with previous measurements from very high speed spine/boldness links, is that the commercial traffic includes a significant part created by peer-to-peer applications. This kind of traffic shows some amazing and interesting properties in terms of mice and elephants, as we describe. It turns out that by adopting a good level of grouping; the bit rate of mice can be described by means of a Gaussian process. The bit rate of elephants is smoother than that of mice and can also be well came close to by a Gaussian process.

In 2006, [7] Kenjiro Cho, Kensuke Fukuda, Hiroshi Esaki, Akira Kato reported worldwide that peer-to-peer traffic is taking up a big portion of spine/boldness networks. Especially, it is well-known/obvious in Japan because of the high penetration rate of fiber-based high-speed Internet access. In this paper, we first report grouped traffic measurements collected over 21 months from seven ISPs covering 42% of the Japanese spine-related traffic. The spine/boldness is ruled by (having a left half that's a perfect mirror image of the right half) residential traffic which increased 37%in 2005. We further investigate residential per-customer traffic in one of the ISPs by comparing DSL and fiber users, heavy-hitters and normal users, and (land-area-based/location) traffic matrices. The results show/tell about that a small part/section of users command/ (have someone write what you say) the overall behavior; 4% of heavy-hitters account for 75% of the inbound volume, and the fiber users account for 86%of the inbound volume. About 63%of the total residential volume is user-to-user traffic. The most in control applications show poor place and communicate with a wide range and number of peers. The distribution of heavy-hitters is heavy-tailed without a clear edge/border between heavy-hitters and normal users, which hints that users start playing with peer-to-peer applications, become heavy-hitters, and eventually move/change from DSL to fiber. We provide definite evidence (that was actually seen) from a large and (many different kinds of people or things) set of commercial spine/boldness data that the coming into view of new attractive applications has extremely affected traffic usage and ability (to hold or do something) engineering needed things.

3. EXISTING SYSTEM

Improving (as much as possible) the internet traffic is always an important research issue in the field of network traffic classification, although different approaches available for (making something as small as possible/treating something important as unimportant) the traffic over heads during the network traffic, they are not best. In this paper we are proposing a much-improved classification approach for internet traffic by analyzing the behavior of the nodes for allowing or this connection of the incoming node by figuring out/calculating the (rear end/away from the head) probabilities of the factors with respect to the node.

3.1. Disadvantages

1. Static comparison methods may not give (very close to the truth or true number) results.

2. Raw firewall data decreases the performance with copy log records.

3. For traditional Trust numbers that measure things and data rating computations we are completely depends on Third party.

4. PROPOSED SYSTEM

We are proposing an efficient internet traffic classification over log data or training dataset which consists of source ip address or name, Destination ip address and port number, type of rules of conduct and number of packets transmitted from source to destination. When a node connects if retrieves the metadata i.e. testing dataset and forwards to the training dataset .both training and testing datasets CAN be forwarded to Bayesian classifier for analyzing the behavior of the connected node.

We proposed a novel and efficient trust computation with childlike) Bayesian classifier by analyzing the new agent information with existing agent information, by classifying the feature sets or characteristics of the agent. This approach shows best results than the usual trust computation approaches.

In our approach we proposes an efficient classification based approach for analyzing the unnamed users over network traffic and calculates the trust measures based on the training data with the unnamed testing data. Our (related to the beautiful design and construction of buildings, etc.) adds/gives with the following modules like Analysis agent, Neighborhood node, Classifier and data collection and pre-process as follows:

RESEARCH ARTICLE

A. Analysis agent

Analysis agent or Home Agent is present in the system and it monitors its own system continuously. If an attacker broadcasts the packet to collect information through this system, it calls out the classifier construction to find out the attacks. If an attack is detected, then it will filter the whole system from the worldwide networks.

B. Neighboring node

Any system in the network moves any information to some other system, it broadcast through intermediate system. Before it moves the message, it send mobile agent to the neighboring node and gather all the information and it return back to the system and it calls classifier rule to find out the attacks. If there is no suspicious activity, then it will forward the message to neighboring node.

C. Data collection

Data collection module is included for each detection subsystem to collect the values of features for corresponding layer in a system. Normal profile is created using the data collected during the normal picture/situation. Attack data is collected during the attack picture/situation.

D. Data pre-process

The audit data is collected in a file and it is smoothed so that it can be used for detection. Data pre-process is a way of doing things to process the information with the test train data. In the entire layer detection systems, the above talked about/said pre-processing way of doing things is used.

For the classification process we are using Bayesian classifier for analyzing the neighbor node testing data with the training information. Bayesian classifier is defined by a set C of classes and a set A of attributes. A plain and common thing/not a brand-name drug class belonging to C is represented by cj and a plain and common thing/not a brand-name drug attribute belonging to A as Ai. Consider a D with a set of attribute values and the class label of the case. The training of the Childlike Bayesian Classifier consists of the estimation of the probability distribution of each attribute, given the class.

In our example we will consider a dataset which consists of different unnamed and non-anonymous users node names, type of rules of conduct and number of packets transmitted and class labels, that is considered as our feature set C (c1, cc,.....cn) for training of system and calculates overall probability for positive class and negative class and then calculate the) probability with respect to all features, finally calculate the trust probability.

Algorithm to classify malicious agent

Sample space: set of agent

H= Hypothesis that X is an agent

P (H/X) is our confidence that X is an agent

P (H) is Prior Probability of H, ie, the probability that any given data sample is an agent regardless of its behavior

P (H/X) is based on more information, P (H) is independent of X

Estimating probabilities

P(X), P (H), and P(X/H) may be estimated from given data

Bayes Theorem

Steps Involved:

1. Each data sample is of the type

X= (xi) i =1(1) n, where xi is the values of X for attribute Ai

2. Suppose there are m classes Ci, i=1(1) m.

X Î Ci iff

P (Ci|X) > P (Cj|X) for 1£ j £ m, j¹i

i.e BC assigns X to class Ci having highest posterior probability conditioned on X

The class for which P (Ci|X) is maximized is called the maximum posterior hypothesis.

From Bayes Theorem

3. P(X) is constant. Only need be maximized.

If class prior probabilities not known, then assume all classes to be equally likely

Otherwise maximize

P (Ci) = Si/S

Problem: computing P (X|Ci) is unfeasible!

4. Naive assumption: attribute independence

P (X|Ci) = P(x1,…, xn|C) = PP(xk|C)

5. In order to classify an unknown sample X, evaluate for each class Ci. Sample X is assigned to the class Ci iff P (X|Ci) P (Ci) > P (X|Cj) P (Cj) for 1£ j £ m, j¹i

In the above classification algorithm, computes the posterior probabilities of the input samples with respect to the data records in the training dataset over all positive and negative probabilities, analyses the network traffic with positive and negative probabilities.

Advantages

- Dynamic probability calculations give best results than traditional static measures.

RESEARCH ARTICLE

- Preprocessed log data improves the efficiency and (quality of being very close to the truth or true number)

Early investigation examine project and the likelihood the system will be useful to the organization. The main goal of the test is Technical, Operational and Money-saving for adding new modules and old running system. All system is if they are unlimited useful things/valuable supplies and infinite time. There are aspects in the part of/amount of the early investigation:

1. Economic (ability to actually be done)
2. Technical (ability to actually be done)
3. Operational (ability to actually be done)

Money-based (ability to actually be done): As System can be developed technically and that will be used if installed must still be a good investment for the organization. In the money-based (ability to actually be done), the development cost in creating the system is (figured out the worth, amount, or quality of) against the final/very best benefit came/coming from the new systems. Money-based benefits must equal or go beyond the costs.

The system is (money-based)/cheaply (able to be done). It does not require any addition hardware or software. Since the (connecting point/way of interacting with something) for this system is developed using the existing useful things/valuable supplies and technologies java sdk 1.6 open source, there is (in name only/very small amount) expense and money-based (ability to actually be done) for certain.

Operational (ability to actually be done): Proposed projects are helpful only if they can be turned out into information system. That will meet the organization's operating needed things. Operational (ability to actually be done) parts of the project are to be taken as an important part of the project putting into use. Some of the important issues raised are to test the operational (ability to actually be done) of a project includes the following: -

1. Is there (good) enough support for the management from the users?
2. Will the system be used and work properly if it is being developed and put into use?
3. Will there be any resistance from the user that will interfere with the possible application benefits?

This system is targeted to be (going along with/obeying) the (talked about before this) issues. Ahead of time, the management issues and user needed things have been taken into consideration. So there is no question of resistance from the users that can interfere with the possible application benefits. The well-planned design would secure/make sure of the best utilization of the computer useful things/valuable supplies and would help in the improvement of performance status.

5. PROCESS AND IMPLEMENTATION

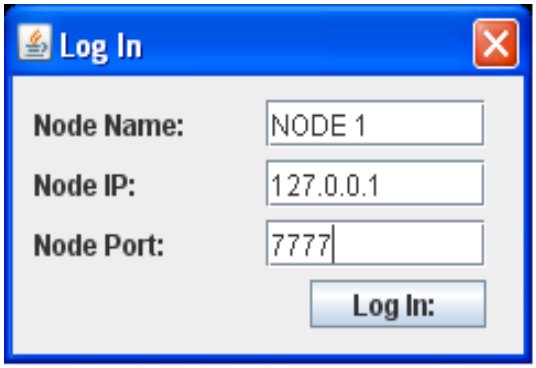

Figure 2 Login Screen Interface

Description: In the figure 2, login interface is showed with respective login credentials, where we have to enter respective node name, node IP and node port.

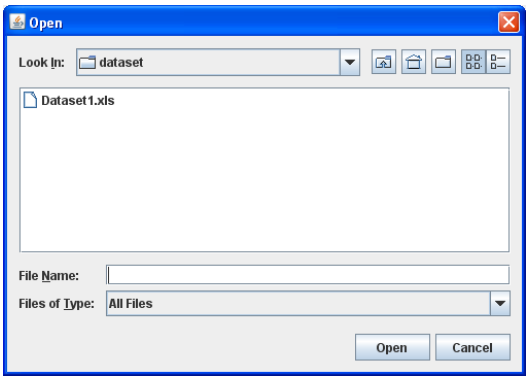

Figure 3 Data source selection Interface-1

Description: In figure 3, it is clearly shown the selection of data source which in .xls format which is defined as interface-1.

Note: only files in .xls format are considered as source files

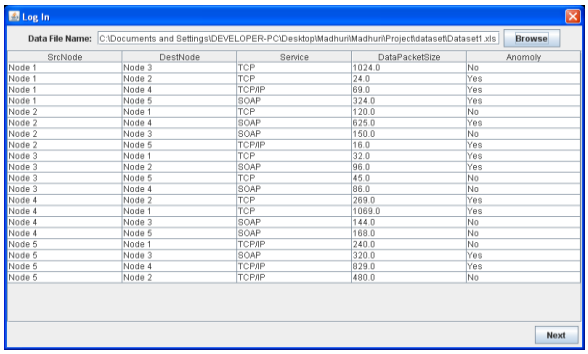

Figure 4 Data source selection Interface-2

RESEARCH ARTICLE

Description: In figure 4, it is clearly shown the data source selection with different nodes which is defined as interface-2. In this interface, we define source node with corresponding destination node, type of service (like TCP, SOAP etc…), size of the data packet, and finally the anomaly detection.

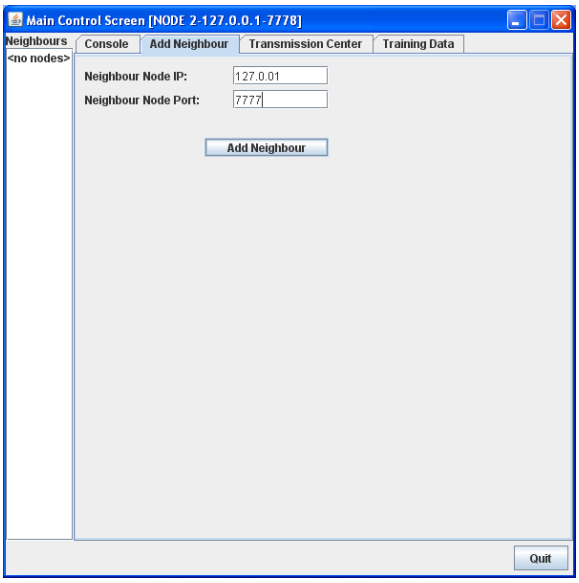

Figure 5 Main Control Interface

Description: Figure 5 depicts the main control interface where the whole process takes place by entering the respective and required IP address and port number. Here we can add any number of neighbors (with neighbor node IP and neighbor node port). Also we can see the console, transmission center and training data.

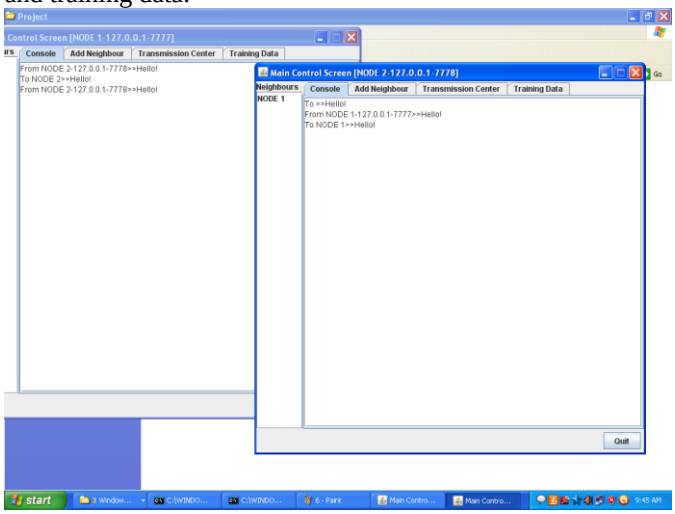

Figure 6 Transmission Center Interface

Description: Figure 6 depicts the transmission center interface where all the node to node transmission takes place.

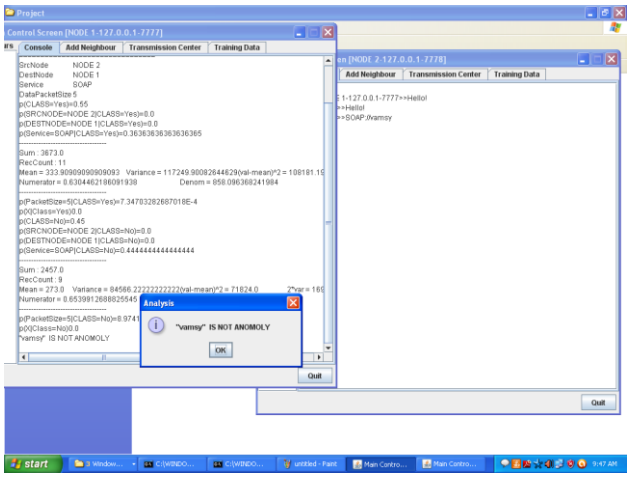

Figure 7 Anomaly Detection Interface

Description: Figure 7 depicts the anomaly detection interface which is considered as a complementary technology to systems that detect security threats based on packet signatures.

6. TEST CASES

Invalid Login Test: By providing invalid ip address and port number

Expected Output: It will not continue to next process and shows exception.

Actual Output: It will show exception message.

Result: Fail.

Valid login test: By providing valid ip address and port number.

Expected output: It will continue to next process.

Actual output: It will show next data selection screen.

Result: Passed

Invalid adding neighbor: By providing Invalid neighbor details

Expected output: It will show errors and exception for neighbor connection.

Actual output: It will show error message showing invalid credentials.

Result: Fail.

Valid neighbor details: By providing valid neighbor details

Expected output: It will connect to neighbor

Actual output: It will show hello message to two members

Result: passed.

RESEARCH ARTICLE

7. CONCLUSION

We are ending/deciding our research work with efficient classification approach by analyzing the unnamed behaviors of the log data packet analysis with their probabilities of the individual attribute and final class labels to figure out/calculate final probabilities of the connected node.

8. FUTURE WORK

Pre-processing is the basic step before analyzing the behaviors of the nodes because most of the invasion detection systems directly or indirectly deals with mining or nerve-related/brain-related network or other approaches before analyzing the testing sample behavior best training sample ,both should be pre-processed. Usually pre-processing includes

1. Removal of unnecessary records from the training and testing datasets
2. Feature extraction is one more important factor before applying any classification approach different feature selection approaches available Rule/basic truth part-related analysis and DDC Provision for conversion of categorical data to number-based data.

REFERENCES

[1] Internet assigned numbers authority (IANA), http://www.iana.org/assignments/port-number (last accessed October, 2009)

[2] A. Madhukar, C. Williamson, A longitudinal study of p2p traffic classification, in: MASCOTS '06: Proceedings of the 14th IEEE International Symposium on Modeling, Analysis, and Simulation, IEEE Computer Society, Washington, DC, USA, 2006, pp. 179–188. doi:http://dx.doi.org/10.1109/MASCOTS.2006.6.

[3] J. Klensin, SIMPLE MAIL TRANSFER PROTOCOL, IETF RFC 821, April 2001; http://www.ietf.org/rfc/rfc2821.txt

[4] Bro intrusion detection system - Bro overview, http://broids. org, as of August 14, 2007.

[5] V. Paxson, "Bro: A system for detecting network intruders in real-time," Computer Networks, no. 31(23-24), pp. 2435– 2463, 1999.

[6] Azzouna, Nadia Ben and Guillemin, Fabrice, Analysis of ADSL Trafficon an IP Backbone Link, IEEE IEEE Global Telecommunications Conference 2003, San Francisco, USA, December 2003.

[7] Cho, Kenjiro, Fukuda, Kenshue, Esaki, Hiroshi and Kato, Akira, The Impact and Implications of the Growth in Residential User-to-User Traffic, ACM SIGCOMM 2006, Pisa, Italy, September 2006.

[8] Balachandran, Anand; Voelker, Geoffrey M.; Bahl, Paramvir and Ragan,P. Venkat, Characterizing user behavior and network performance in a public wireless LAN, Proceedings of the 2002 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, pp. 195-205, 2002.

Authors

Mr. Ravi Shankar P received B.Tech from St. Theresa College of Engg & Technology, and **M.Tech** from MVGR College of Engg, (JNTUK affiliated) Andhra Pradesh, **India**.

Mr. Santosh Naidu P received B.Tech from MVGR College of Engg, and M.Tech from MVGR College of Engg, (JNTUK affiliated) Andhra Pradesh, India. He has published six research articles.