Classifying Dominant Congested Path Using Correlation Factors

[1] D. Prabavathi
M. Phil Scholar, Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

[2]Mrs. K. V. Sumathi
M.C.A., M.Phil,Assistant Professor,
Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

[3] Mrs. K. K. Kavitha
M.C.A., M.Phil., SET., (Ph.D).,
Vice Principal, Head of the Department of Computer Science,
Selvamm Arts and Science College (Autonomous)
Namakkal (Tk) (Dt) – 637003

Abstract: Traffic classification has wide applications in network management, from security monitoring to quality of service measurements. Recent research tends to apply machine learning techniques to flow statistical feature based classification methods. The nearest neighbor (NN)-based method has exhibited superior classification performance. It also has several important advantages, such as no requirements of training procedure, no risk of overfitting of parameters, and naturally being able to handle a huge number of classes. However, the performance of NN classifier can be severely affected if the size of training data is small. In this paper, we propose a novel nonparametric approach for traffic classification, which can improve the classification performance effectively by incorporating correlated information into the classification process. We analyze the new classification approach and its performance benefit from both theoretical and empirical perspectives. A large number of experiments are carried out on two real-world traffic data sets to validate the proposed approach. The results show the traffic classification performance can be improved significantly even under the extreme difficult circumstance of very few training samples.

*****

I. INTRODUCTION

Identifying the existence of a dominant path is useful for traffic engineering. For example, when there are multiple paths from one host to another and all are congested, improving the quality along a path with one dominant path may require fewer resources than those along a path with multiple congested links. Identifying whether a path has a dominant path also helps us understand and model the dynamics of the network since the behavior of a network with a dominant path differs dramatically from one with multiple congested links.

When a dominant path exists, identifying the existence of such a path requires distinguishing its delay and loss characteristics from those of the other links. Achieving this goal via direct measurements is only possible for the organization in charge of that network. However, commercial factors often prevent an organization from disclosing the performance of internal links. Furthermore, as the Internet grows in both size and diversity, one organization may only be responsible for a subset of links on an end–end path. Some measurement techniques obtain internal properties of a path by using ICMP (Internet Control Message Protocol) messages to query internal routers.

NETWORK TOMOGRAPHY

Network tomography infers internal link properties through end–end measurements. A rich collection of network tomography techniques have been developed in the past. Many techniques rely on correlated measurements (through multicast or striped unicast probes). More recently, several studies use uncorrelated measurements to detect lossy links, estimate loss rates, or locate congested segments that have transient high delays.

Measurement and inference of end-end path characteristics have attracted a tremendous amount of attention in recent years. Properties such as the delay and loss characteristics of the end-end path, the minimum capacity and available bandwidth of the path and the stationary of the network have been investigated. These efforts have improved our understanding of the Internet. They have also proved valuable in helping to manage and diagnose heterogeneous and complex networks. Furthermore, they have been exploited by several applications, such as server selection, overlay networks and streaming applications, to improve performance.

As we will see, the identification procedure only requires a short probing duration, in terms of minutes. The following are the main contributions of this paper:

- We present a formal yet intuitive definition of dominant congested link and provide two simple hypothesis tests to identify the existence of dominant congested link along a path.
- Our model-based approach fully utilizes the information from the probing packets and enables very fast identification. Validation using ns

simulation and Internet experiments demonstrate that this approach can correctly identify the existence of a dominant congested link within minutes.

- As a result of the identification procedure, we provide a statistical upper bound on the maximum queuing delay of the dominant congested link once we identify that a dominant congested link exists.

II. CONGESTION

Congestion, in the context of networks, refers to a network state where a node or link carries so much data that it may deteriorate network service quality, resulting in queuing delay, frame or data packet loss and the blocking of new connections. In a congested network, response time slows with reduced network throughput. Congestion occurs when bandwidth is insufficient and network data traffic exceeds capacity.

- Exponential backoff protocols that use algorithm feedback to decrease data packet throughput to acceptable rates
- Priority techniques to allow only critical data stream transmission
- Allocation of appropriate network resources in anticipation of required increases in data packet throughput

Congestion has been described as a fundamental effect of limited network resources, especially router processing time and link throughput. Traffic directing processes, performed by routers on the Internet and other networks, use a microprocessor. Cumulative router processing time greatly impacts network congestion. In fact, intermediate routers may actually discard data packets when they exceed its handling capability. When this occurs, additional data packets may be sent to make up for un received packets, which exacerbates the problem. Network congestion often leads to congestion collapse.

NETWORK CONGESTION

DO NOT confuse with flow control. Sometimes Internet IP packet routers get overloaded and congested,if that happens they will have to discard some packets. What could happen, if the sliding window packet retransmission software is too simple, is that it will immediately respond by retransmitting all the lost packets. This will make the congestion worse! All “good” implementations of TCP should respond more gentlyif packets timeout then the TCP sender will reduce the window size and delay before retransmitting, if it still has timeouts it delay even longer. It will only start increasing the window size and cutting the delay when it starts receiving acknowledgements again.

CONGESTION CONTROL OVERVIEW

Problem

When too many packets are transmitted through a network, congestion occurs At very high traffic, performance collapses completely, and almost no packets are delivered.

Causes: bursty nature of traffic is the root cause When part of the network no longer can cope a sudden increase of traffic, congestion builds upon. Other factors, such as lack of bandwidth, ill-configuration and slow routers can also bring up congestion.

Traffic shaping

Flow control policy

As burstiness of traffic is a main cause of congestion, it is used to regulate average rate and burstiness of traffic – e.g. when a virtual circuit is set up, the user and the subnet first agree certain traffic shape for that circuit. Monitoring traffic flow, called traffic policing, is left to the subset – Agreeing to a traffic shape and policing it afterward are easier with virtual circuit subnets, but the same ideas can be applied to datagram subnet at transport layer.

III. LITERATURE SURVEY

A MEASUREMENT STUDY OF AVAILABLE BANDWIDTH ESTIMATION TOOLS

Available bandwidth estimation is useful for route selection in overlay networks, QoS verification, and traffic engineering. Recent years have seen a surge in interest in available bandwidth estimation. A few tools have been proposed and evaluated in simulation and over a limited number of Internet paths, but there is still great uncertainty in the performance of these tools over the Internet at large.

It is important that unprivileged users be able to diagnose their paths. Performance depends on the interaction of the properties of the entire path and the application. Since operators do not share the users’ view of the network, they are not always well-placed to even observe the problem. Even when they are, they may be little better off than users.

USER LEVEL INTERNET PATH DIAGNOSIS

Focus on the problem of locating performance faults such as loss, reordering, and significant queuing at specific links, routers, or middle boxes (e.g., firewalls) along Internet paths. We consider this problem from the point of view of an ordinary user, with no special privileges, in a general setting where paths cross multiple administrative domains. We refer to this as the problem of user-level path diagnosis.

BLINC: MULTILEVEL TRAFFIC CLASSIFICATION IN THE DARK

To present a fundamentally different approach to classifying traffic flows according to the applications that generate them. In contrast to previous methods, our approach is based on observing and identifying patterns of host behavior at the transport layer. We analyze these patterns at three levels of increasing detail (i) the social, (ii) the functional and (iii) the application level.

SVM BASED NETWORK TRAFFIC CLASSIFICATION USING CORRELATION INFORMATION

Traffic classification is an automated process which categorizes computer network traffic according to various parameters into a number of traffic classes. Many supervised classification algorithms and unsupervised clustering algorithms have been applied to categorize Internet traffic. Traditional traffic classification methods include the port-based prediction methods and payload-based deep inspection methods. In current network environment, the traditional methods suffer from a number of practical problems, such as dynamic ports and encrypted applications.

IV. RESEARCH METHODOLOGY

MEASUREMENT METHODOLOGY

For each type of analysis, we use a variety of tools andmethods to collect and analyze network measurement data. However, the Pathneck tool and the measurement sources and destinations selection method are used in all the studies we present. We discuss them in this section. For the convenience of reference, Table I lists the definition of the terms used in
this paper.

Background on Pathneck

Pathneck is an active probing tool that allows end usersto efficiently and accurately locate the bottleneck link on an Internet path. Pathneck is based on a novel probing technique called Recursive Packet Train (RPT) which combines load and measurement packets. The load packets are UDP packets that are used to interact with background.

Reducing congestion networks

Congestion considerably slows down the overall network performance. Latency denotes the period of time spent by data on traversing a network segment. High latency, caused by a congested network, slows the speed of the enterprise network performance leading to unsatisfied and angry end users.

Here are 5 tips that can help to reduce the congestion in enterprise networks:

1. Conduct an analysis of the network traffic flows with the help of network monitoring tools. Setup a network sniffer to analyze network traffic, so that underlying troubles in network can be found out and submitted for resolution. While monitoring the network, look into the segments which generate the highest volume of traffic.
2. Network bottlenecks, which are the main reasons of congestion in a network, must be eliminated. Just like a traffic jam caused by a narrowing of a busy four-lane highway to just two lanes, a bottleneck, which is a network's segment unable to handle the amount of traffic coming from its connecting segments, causes intolerable amounts of congestion on a network. Bottlenecks can be eliminated by increasing the segment's bandwidth capacity so it matches the neighboring segments' maximum traffic flow. This can be accomplished by upgrading this network's segment or using different one.

EXISTING SYSTEM OF METHODOLOGY

In the existing systems, the identification of a dominant traffic path existence requires distinguishing its delay and loss characteristics from those of the other links. Achieving this goal via direct measurements is only possible for the organization in charge of that network. However, commercial factors often prevent an organization from disclosing the performance of internal links. Furthermore, as the Internet grows in both size and diversity, one organization may only be responsible for a subset of links on an end–end path. Some measurement techniques obtain internal properties of a path by using ICMP messages to query internal routers. Trace route and ping are two widely used tools in this category. Some more advanced techniques use ICMP messages to measure per-hop capacity or delay and pinpoint faulty links. These approaches, however, require cooperation of the routers (to respond to ICMP messages and treat them similarly as data packets). Contrary to direct measurements using responses from routers, a collection of network tomography techniques infers internal loss rate and delay characteristics using end-end measurements. Most tomography techniques, however, require observations from multiple vintage points. Disadvantages of this system are

- Loss of data is high,
- More delay time,
- Low available bandwidth,
- Low capability,
- Tight and narrow links.

PROPOSED SYSTEM OF METHODOLOGY

We propose a new framework, Traffic Classification using Correlation information (TCC), to address the problem of very few training samples. The correlation information in network traffic can be used to effectively improve the classification accuracy. To propose a novel non-parametric approach which incorporates correlation of traffic flows to improve the classification performance. To provide a detailed analysis on the novel classification approach and its performance benefit from both theoretical and empirical aspects.

DATA FLOW DIAGRAM

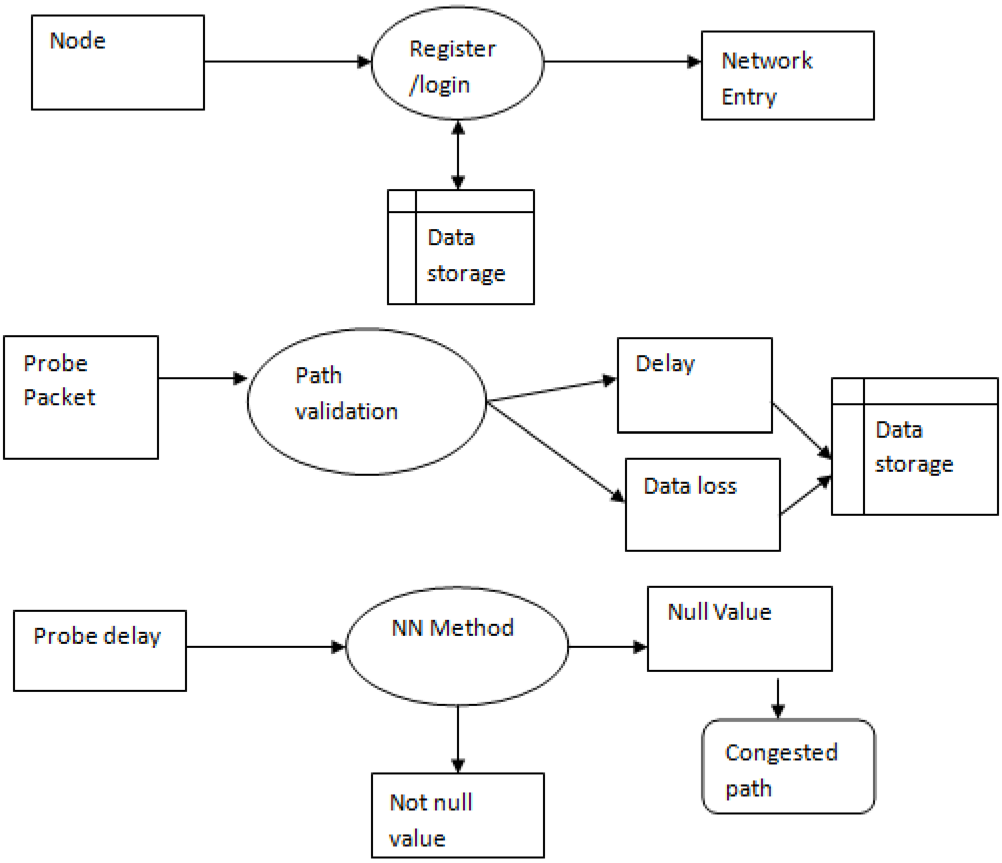

Figure3.1 Data Flow Diagram

NETWORK TRAFFIC CLASSIFICATION

Network traffic classification has drawn significant attention over the past few years. Classifying traffic flows by their generation applications plays very important role in network security and management, such as quality of service (QoS) control, lawful interception and intrusion detection. Traditional traffic classification methods include the port-based prediction methods and payload-based deep inspection methods. In current network environment, the traditional methods suffer from a number of practical problems, such as dynamic ports and encrypted applications. Recent research efforts have been focused on the application of machine learning techniques to traffic classification based on flow statistical features.

COMPUTATIONAL PERFORMANCE

The computational performance includes learning time,amount of storage, and classification time. First, the NN classifier does not really involve any learning process, which is shared with our proposed methods. However, other supervised methods, such as neural nets and SVM, need time to learn parameters for their classification model. Second, the proposed methods use the nearest neighbor rule which requires storage for all training data samples. However, the amount of storage is tiny if the training data size is small. Identifying the nearest neighbor of a given flow from among n training flows is conceptually straightforward with n distance calculations to be performed. The nearest neighbor rule is embedded in the proposed methods for traffic classification. With a small training set, the NN classifiers and the proposed methods classify very quickly. For instance, with 10 training samples for each class, the classification time of the proposed methods are about 2 seconds and 5 seconds for the *wide* dataset and *isp*dataset, respectively. The proposed methods, AVG-NN, MIN-NN, and MVTNN, have the same classification time, because they follow the same classification approach. Due to the extra aggregation operation, the classification time of the proposed methods is a little longer than NN but the classification accuracy of our methods is much higher than NN. If NN achieves the same accuracy to our proposed methods, it needs more training samples and must spend more classification time.

SYSTEM FLEXIBILITY

The proposed system model is open to feature extractionand correlation analysis. First, any kinds of flow statistical features can be applied in our system model. In this work, we extract unidirectional statistical features from full flows. The statistical features extracted from parts of flows can also be used to represent traffic flows inour system model. Second, any new correlation analysis to discover correlation information in traffic flows to improve the robustness of classification. In this paper, a 3-tuple heuristic based method is applied to discover flow correlation which aremodeled by BoFs.We presented the comprehensive analysis from theoretical and empirical perspectives, which is based on the BoF model instead of the 3-tuple method. Therefore, new correlation analysis methods will not affect the effectiveness of the proposed approach. In the future, we will work on developing new methods for flow correlation analysis.

CORRELATION ANALYSIS

We conduct correlation analysis using a 3-tuple heuristic,which can quickly discover BoFs in the real traffic data.

3-tuple heuristic: in a certain period of time, the flows sharing the same 3-tuple {dstip, dst port, protocol} form a BoF. The correlated flows sharing the same 3-tuple are generated by the same application. For example, several flows initiated by different hosts are all connecting to a same host at TCP port 80 in a short period. These flows are very likely generated by the same application such as a web browser. The 3-tuple heuristic about flow correlation has been considered in several practical traffic classification schemes. Ma *et al.* proposed Abpayload-based clustering method for protocol inference, in which they grouped flows into equivalence clusters using the heuristic. Canini*et al.*tested the correctness of the 3-tuple heuristic with real-world traces. In our previous work , we applied the heuristic to improve unsupervised traffic clustering. BoF to model the correlation information obtained by the 3-tuple heuristic and study the BoF model based supervised classification, which is different from the exiting works.

V. RELATED WORK

For instance, traffic classification is normally an essential component in the products for QoS control and The goal of network traffic classification is to classify intrusion detection .With the popularity of cloud traffic flows according to their generation applications. The computing, the amount of applications deployed on the current research of traffic classification concentrates on the Internet is quickly increasing and many applications adopt application of machine learning techniques into flow the encryption techniques. This situation makes it harder to statistical feature based classification methods.

A TRAFFIC CLASSIFICATION APPROACHWITH FLOW CORRELATION

Traffic Classification using **Correlation** information or***TCC*** for short. A novel nonparametric approach is also proposed to effectively incorporate flow correlation information into the classification process.

A Traffic Classification Approach with Flow Correlation

The presents a new framework which we call Traffic Classification using Correlation information or TCC for short. A novel nonparametric approach is also proposed to effectively incorporate flow correlation information into the classification process.

Correlation Analysis

The correlated flows sharing the same three-tuple are generated by the same application. For example, several flows initiated by different hosts are all connecting to a same host at TCP port 80 in a short period. These flows are very likely generated by the same application such as a web browser. The three-tuple heuristic about flow correlation has been considered in several practical traffic classification schemes proposed a payload based clustering method for protocol inference, in which they grouped flows into equivalence clusters using the heuristic. tested the correctness of the three-tuple heuristic with real-world traces.

Computational Performance

The computational performance includes learning time, amount of storage, and classification time. First, the NN classifier does not really involve any learning process, which is shared with our proposed methods. However, other supervised methods, such as neural nets and SVM, need time to learn parameters for their classification model. Second, the proposed methods use the nearest neighbor rule which requires storage for all training data samples. However, the amount of storage is tiny if the training data size is small.

System Flexibility

The proposed system model is open to feature extraction and correlation analysis. First, any kinds of flow statistical features can be applied in our system model. In this work, we extract unidirectional statistical features from full flows. The statistical features extracted from parts of flows can also be used to represent traffic flows in our system model. Second, any new correlation analysis method can be embedded into our system model. We introduce flow correlation analysis to discover correlation information in traffic flows to improve the robustness of classification. In this paper, a three-tuple heuristic-based method is applied discover flow correlations which are modeled by BoFs.

System Model

In thepre-processing, the system captures IP packets crossing a computer network and constructs traffic flows by IP header inspection. A flow consists of successive IP packets having the same 5-tuple: {srcip, src port, dstip, dst port, protocol}. After that, a set of statistical features are extracted to represent each flow. Feature selection aims to select a subset of relevant features for building robust classification models. Flow correlation analysis is proposed to

correlate information in the traffic flows. Finally, the robust traffic classification engine classifies traffic flows into application-based classes by taking all information of statistical features and flow correlation into account. We observe that the accuracy of conventional traffic classification methods are severely affected by the size of training data. When a small size

VI. CONCLUSION

To investigated the problem of traffic classification using very few supervised training samples. A novel nonparametric approach, TCC, was proposed to investigate correlation information in real traffic data and incorporate it into traffic classification. We presented a comprehensive analysis on the system framework and performance benefit from both theoretical and empirical perspectives, which strongly support the proposed approach. Three new classification methods, AVG-NN, MIN-NN, and MVT-NN, are proposed for illustration, which can incorporate correlation information into the class prediction for improving classification performance. A number of experiments carried out on two real-world traffic datasets show that the performance of traffic classification can be improved significantly and consistently under the critical circumstance of very few supervised training samples. The proposed approach can be used in a wide range of applications, such as automatic recognition of unknown applications from captured network traffic and semi-supervised data mining for processing network packets.

VII. FUTURE WORK

This work lends itself to extension in several directions. One open issue is determining congestion sharing in a multiple bottleneck scenario. Namely, sharing or not sharing is more than simply a binary variable. Consider two flows that share congestion at the access link of their common receiver; yet, one of them crosses a separate upstream bottleneck. In such a scenario, some kind of hierarchical congestion classification is desirable.

Another direction for future work is a more detailed investigation of the shape of the inter-arrival distributions. In particular, the envelopes formed by the tips of the spikes in Figure 1 trace out very regular curves. It would be informative to fit the spike train and the spike bump to well-known distributions and analyse the shape of their tails. This may lead to a better understanding of the distribution of the cross traffic burst. Furthermore, finding good models for the inter-arrival distribution in a flow would improve the ability to cluster flows that share the bottlenecks.

REFERENCE

[1] L. Bernaille, R. Teuxeira, I. Akodkenou, A. Soule, K. Salamatian, “Traffic Classification on the Fly”, ACM SIGCOMM Computer Communication Review, vol. 36, no. 2, April 2006.

[2] T. Karagiannis, K. Papagiannaki, and M. Faloutsos,“BLINC: Multilevel Traffic Classification in the Dark”, ACM Sigcomm 2005, Philadelphia, PA, USA, August 2005.

[3] Andrew W. Moore and KonstantinaPapagiannaki, "Toward the Accurate Identification of Network Applications", Passive and Active Measurements Workshop, Boston, MA, USA, March 31 - April 1, 2005.

[4] Cisco IOS Documentation, “Network-Based Application Recognition and Distributed Network-

[5] Based Application Recognition“, (as of February 2005).

[6] K. Lan, J. Heidemann, “On the correlation of Internet flow characteristics”, Technical Report ISI-TR-574, USC/Information Sciences Institute, July, 2003.

[7] Laurent Bernaille, RenataTeixeira, “Implementation Issues of Early Application Identification”, AINTEC 2007

[8] App-ID™ Application Classification, Technology Overview, Paloalto Networks, June, 2007

[9] Thomas Karagiannis, Andre Broido, Nevil Brownlee, kcclaffy, “Is P2P dying or just hiding?”, Proceedings of Globecom 2004, November/December 2004.

[10] S. Sen, O. Spatscheck, D. Wang, “Accurate, Scalable In¬Network Identification of P2P Traffic Using Application Signatures”, WWW 2004, New York, USA, May 2004.