# Characteristics of Network Traffic Flow Anomalies

Paul Barford and David Plonka

## I. Introduction

One of the primary tasks of network administrators is monitoring routers and switches for anomalous traffic behavior such as outages, configuration changes, flash crowds and abuse. Recognizing and identifying anomalous behavior is often based on ad hoc methods developed from years of experience in managing networks. A variety of commercial and open source tools have been developed to assist in this process, however these require policies and/or or thresholds to be defined by the user in order to trigger alerts. The better the description of the anomalous behavior, the more effective these tools become. In this extended abstract we describe a project focused on precise characterization of anomalous network traffic behavior.

The first step in our project is to gather passive measurements of network traffic at the IP *flow* level. IP flow level data as defined in [1] is a unidirectional series of IP packets of a given protocol traveling between a source and a destination IP/port pair within a certain period of time. While flow level data is certainly not as precise as passive measurements of packet level data, we demonstrate that it is sufficient for exposing many different types of aberrant network traffic behavior in close to real time. It also has the benefit of generating much smaller data sets than packet level measurements which can become a significant issue in large, heavily used networks.

We use the FlowScan [2] open source software to gather and analyze network flow data. FlowScan takes Netflow [3] feeds from Cisco or other Lightweight Flow Accounting Protocol (LFAP) enabled routers, processes the data and then it in an efficient data structure. FlowScan also has a graphical interface which is currently the principal means for anomaly identification by network managers. FlowScan is currently deployed at the border router at the University of Wisconsin - Madison as well as over 100 other sites nation wide.

P. Barford is a member of the Computer Science Department at the University of Wisconsin, Madison. D. Plonka is a member of the Division of Information Technology at University of Wisconsin, Madison. E-mail: pb@cs.wisc.edu, plonka@doit.wisc.edu

FlowScan has been used effectively at UW - Madison to identify a variety of traffic anomalies for the past two years. To begin our analysis, we cluster these anomalies into three groups based on similarities in observed flow behavior. The groups include network operation anomalies, flash crowd anomalies and network abuse anomalies. Experience has shown that the key to identifying each of these types of anomalies is to use *combinations* of flow measurements which are explained in Section IV. We present examples of each of the aforementioned anomalies in this abstract, and our future task is to analyze and characterize collections of each type of anomaly.

Our anomaly analysis process will be focused on precisely identifying both similarities and differences within each anomaly group. Our goal is not simply to cluster anomalies with similar statistical features but actually to characterize the features of each anomaly group rigorously. Our study benefits greatly from the fact that we have and continue to build an archive of flow data for which anomalies have already been identified by network managers (through ad hoc methods) Our analysis approach will employ a variety of tools including simple statistics, time series analysis and wavelet analysis to characterize anomaly features. We anticipate that each anomaly group will exhibit some invariant characteristics; our hope is that this will be sufficient to differentiate each anomaly group such that anomalies can be accurately identified through automated methods in near real time. Finally, we intend to gather flow data from at least 10 other institutions to see if similar anomalies are observed at other sites.

## II. Related Work

Network traffic properties have been intensely studied for quite some time. Examples of analysis of typical traffic behavior can be found in [4], [5]. More detailed characterizations and models of network traffic including the identification of *self-similar* properties can be found in [6], [7]. A variety of analysis methods have been used in these and other studies including time series techniques and wavelet analysis [8]. The majority of this work has been focused on the typical, packet level behavior (a notable exception being [9]). Our focus is at the flow level and on characterizing anomalous behavior.

Fault and general anomaly detection techniques in networks have also been widely treated due to their impor-

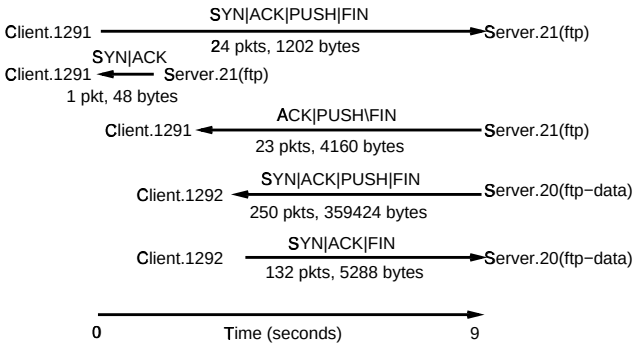


Fig. 1. Flow level breakdown of a simple FTP transfer

tance in network management. Examples include work by Katzela and Schwartz which focuses on methods for isolating failures in networks [10], Feather *et. al* which shows that faults can be detected by statistical deviations from regularly observed behavior [11], Brutlag which applies thresholds to time series models to detect aberrant network behavior [12], and Hood and Ji who describe an adaptive monitoring system which is able to detect unknown or unseen faults [13]. Most of this work focuses on how to detect accurately deviations from normal behavior, whereas our work is focused on analyzing and characterizing statistically specific types of anomalous behavior.

Many papers have been written on detection of nefarious behavior such as denial-of-service (DoS) attacks and port scan attacks which have increasing over the past few years. This includes papers on clustering methods [14], neural networks [15] and Markov models [16] to recognize intrusions. Recent work by Moore *et. al* has shown that flow-based methods can be effective for identifying DoS [17]. Related to this is the development of intrusion detection tools such as Bro [18] which provide a framework for defining policies to detect attacks. Our work complements this work by providing detailed statistical descriptions of a variety of anomalous behaviors.

One area not particularly well treated in the literature is characterizations of *flash crowd* behavior. While content delivery companies have installed vast infrastructures to deal with large populations of users suddenly requesting the same content in a very short time interval (such as the famed Victoria Secret webcast), little has been done in the way of characterizing this behavior. New mechanisms involving *cooperative pushback* are being proposed for detection and control of this type of problem [19].

## III. Measurement of Flow Data

FlowScan collects Netflow data exported by Cisco routers in a network. Netflow data includes source and destination AS/IP/port pairs, packet and byte counts, flow start and end times and protocol information. This data is exported either on timer deadlines or when certain events occur; whichever comes first. Thus, a single transaction, such as the FTP transfer shown in Figure 1, is represented as multiple data flows between the two hosts.

FlowScan maintains a set of counters based upon the attributes of each flow reported by a router. The attributes include IP protocol (ICMP, TCP, UDP), well known service (such as FTP or HTTP) based on source/destination port, CIDR block of local IP address and source/destination AS number. This time series data is written periodically into an efficient database which is used for both archiving and as an interface to the graphical back end which displays aggregate flow data.

Visualizations of both inbound and outbound traffic flows are given by FlowScan for data aggregated over five minute intervals, and are displayed by bits/packets/flows per second over a given time period. An example of packets per second broken out by protocol type is shown in Figure 2. While this level of reporting is coarse-grained enough so that short time scale behavior will be missed, it is sufficient for observing many traffic flow anomalies. Of course, aggregation of this data is possible and is used to visualize long term trends in network use.

FlowScan has been deployed at our site for the past two years. During this time a great deal of operational expertise has been developed in identifying specific traffic anomalies from graphs of traffic flows. This expertise has been developed by first observing a significant difference in traffic flow and then tracking down the source of the anomaly using other tools such as SNMP network monitors. Experience has enabled classes of anomalies to easily be distinguished from typical traffic based on graphs of traffic flows. Since we are collecting data from an operational network, each anomaly is confirmed, diagnosed and logged in detail by network managers.

## IV. Anomaly Identification

Visual analysis of traffic flow anomalies has lead to grouping anomalies into three general categories. These categories are useful for describing general anomaly characteristics however, they may or may not continue to be useful after we complete our characterization work.

**Network Operation Anomalies:** These include network device outages, significant differences in network behavior caused by configuration changes (*e.g.* adding new equipment or imposing rate limits) and plateau behavior caused by traffic reaching environmental limits. Anomalies in this category are distinguished visually by steep, nearly instantaneous changes in bit rate followed by bit rates which are stable but at a different level over a time

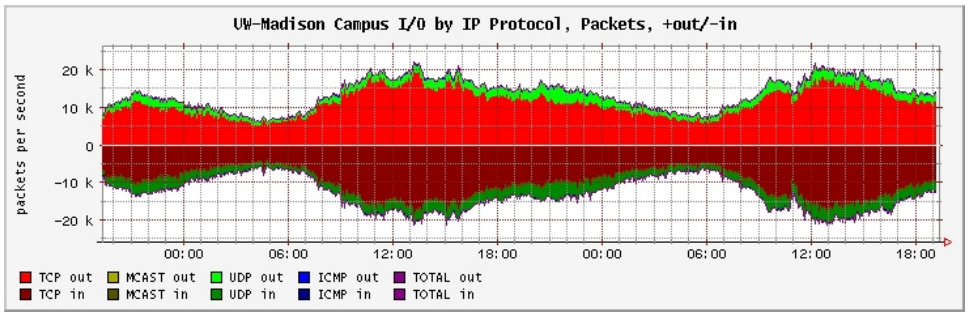


Fig. 2. Example of FlowScan output: Packet count per second broken down by protocol for a typical 48 hour period

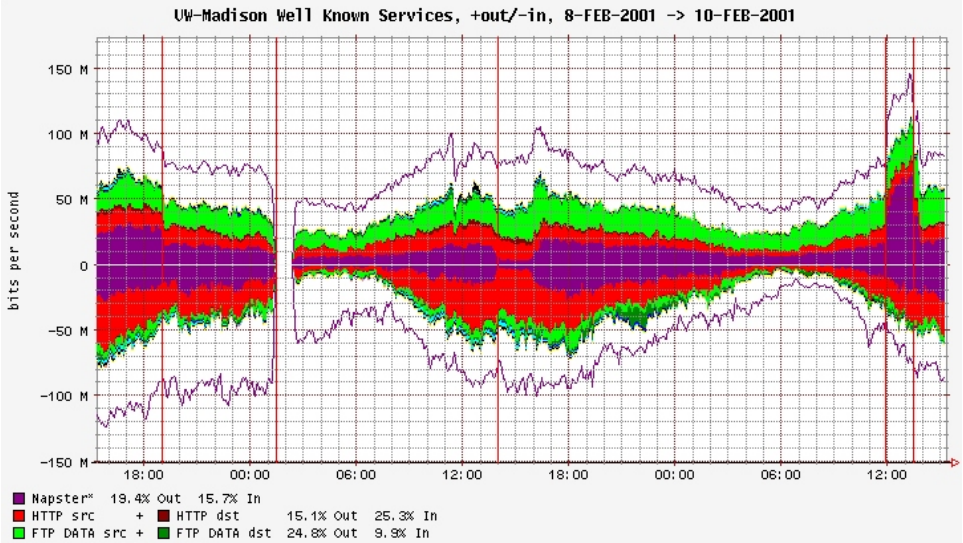


Fig. 3. Example of implementation of rate limits on Napster traffic

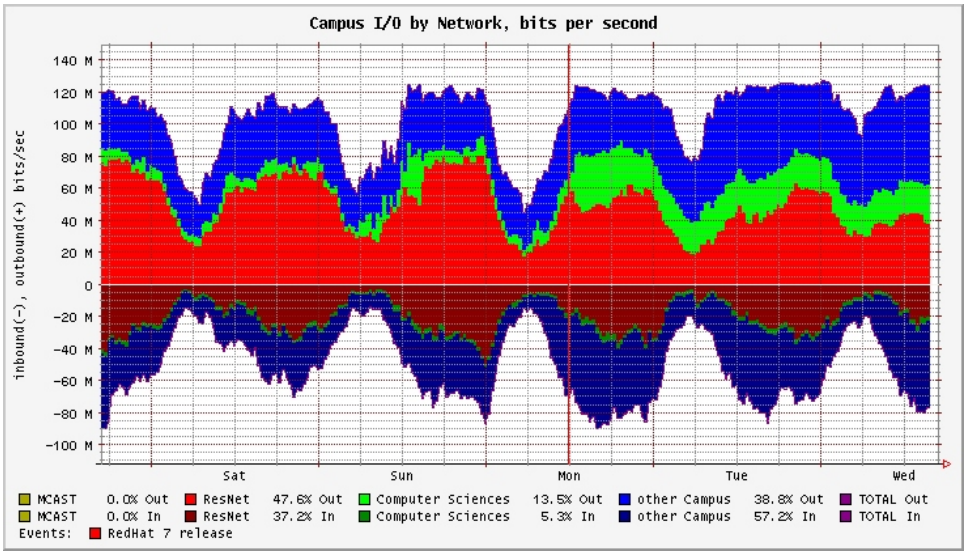


Fig. 4. The release of Linux Redhat 7.0: an example of flash crowd behavior

period. An example of a network operation anomalies can be seen in Figure 3. This figure shows five minute averages for bits per second transferred into and out of our network broken out by application. Five distinct anomalies are identified by the vertical lines in the graph. They were diagnosed as a network outage which occurred just after 1:00am, a Napster server outage which occurred at 2:00pm, and three instances of turning on/off rate limiters on Napster traffic for the network.

**Flash Crowd Anomalies:** In our environment, anomalies in this category are typically due to either a software release (*e.g.* UW is a RedHat Linux mirror site) or external interest in a Web site due to some kind of national publicity. Flash crowd behavior is distinguished by a rapid rise in traffic flows of a particular type (eg. FTP flows) or to a well known destination with a gradual drop off over time. An example of a flash crowd anomaly can be seen in Figure 4. This figure shows hourly bit rate averages over a five day period broken out by local source/destination. The anomaly identified in this graph is the large increase on Monday in traffic flowing out of the Computer Science department. In this instance, the CS department hosts a mirror site for RedHat Linux and Monday was when the 7.0 release occurred.

**Network Abuse Anomalies:** Two types of network abuse that can be identified using flows are DoS flood attacks and port scans. These types of abuse are observed multiple times per week in our network. Network abuse anomalies are distinct from network operation and flash crowd anomalies in that they are not always readily apparent in bit or packet rate measurements. However, flow count measurements clearly indicate abuse activity with many distinct source address/port pairs since each connection appears as a separate flow. An example of a network abuse anomaly can clearly be seen in Figure 5. This figure shows five minute averages for flows per second into and out of our network broken out by protocol. The anomalous behavior is clearly evident in the spike of flows into the network during a half hour period just before noon.

## V. Anomaly Characteristics

One of the principal distinctions of our project is our intention to analyze rigorously and characterize network traffic flow anomalies. While anomaly detection has been addressed in many prior projects, we are aware of no other work which has statistically characterized different types of network traffic flow anomalies. One advantage we have in this process is our ability to identify specific network traffic anomalies in a ex post facto manner and relate them directly to FlowScan measurements. This enables us to gather and classify potentially large sets of data in each of

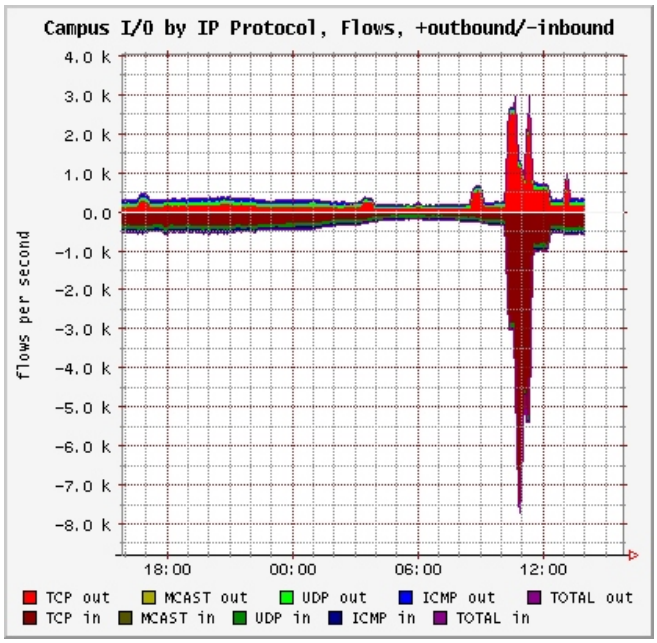


Fig. 5. An example of detecting a denial of service attack

our anomaly categories. We currently have a small archive of flow data anomalies at the five minute time aggregates, and we are in the process of building up the archive at this time.

The first step in our analysis process will be to isolate each of the anomalies in our data sets and to group them into our three general categories. Simple statistical analysis techniques will then be applied to each of the anomalies. These include finding moments, plotting distributions and looking for distributional models to describe the anomalies. This level of analysis may or may not lead identification of significant similarities or differences within and/or between categories.

Our next step will be to apply time series analysis techniques to the anomaly data. This will include analyzing stationarity, correlation structures and testing various time series models to see if any are accurate statistical representations of our anomaly data. We expect these analyses to give insight to the nature of anomalies and possibly to provide predictive capability if good models can be developed, however the distinctive shapes of each type of anomaly warrant further investigation.

The final step in our characterization process will be to apply wavelet analysis to the anomaly data. Wavelets are functions which divide data into frequency components enabling analysis of each component according to its scale. Wavelets have advantages over standard Fourier analysis for data sets which have sharp spikes such as is seen in our anomaly data. We expect wavelet analysis to shed significant light on the structures of each anomaly and to pro-

vide us with additional models for identifying and grouping anomalies.

## VI. Conclusions and Future Work

In this extended abstract we describe our project to characterize network traffic flow anomalies. The goal of our work is to identify precisely the statistical properties of anomalies and their invariant properties if they exist. If we are successful in this effort, our results can be coupled with flow monitoring tools to generate more accurate real time alerts when anomalies occur.

At the time of writing we are in the process of building an archive of anomalies based on IP traffic flow measurements taken from the border router for our campus network. We are in the early stages of applying various statistical analysis techniques to the data.

After completing the current round of analysis we intend to extend this project in a number of directions. We plan to evaluate whether or not we are better able to distinguish anomalies by taking measurements from FlowScan at one minute intervals. This will give us a more accurate representation of behavior but at the cost of much larger data sets. We also plan to extended our anomaly data collection process across multiple sites. FlowScan is already widely deployed and multiple sites have already tentatively agreed to participate. Not only will this give us larger datasets but will also enable us to investigate correlations of behavior across sites.

## References


[1] K. Claffy, G. Polyzos, and H.-W. Braun, "Internet traffic flow profiling," Tech. Rep. TR-CS93-328, University of California San Diego, November 1989.

[2] D. Plonka, "Flowscan: A network traffic flow reporting and visualization tool," in *Proceedings of the USENIX Fourteenth System Administration Conference LISA XIV*, New Orleans, LA, December 2000.

[3] Cisco's IOS Netflow Feature, ," http://www.cisco.com/wrap/public/732/netflow.

[4] R. Cáceres, "Measurements of wide-area Internet traffic," Tech. Rep. UCB/CSD 89/550, Computer Science Department, University of California, Berkeley, 1989.

[5] V. Paxson, *Measurements and Analysis of End-to-End Internet Dynamics*, Ph.D. thesis, University of California Berkeley, 1997.

[6] V. Paxson and S. Floyd, "Wide-area traffic: The failure of poisson modeling," *IEEE/ACM Transactions on Networking*, vol. 3(3), pp. 226–244, June 1995.

[7] W. Willinger, M. Taqqu, R. Sherman, and D. Wilson, "Self-similarity through high-variability: Statistical analysis of Ethernet LAN traffic at the source level," *IEEE/ACM Transactions on Networking*, vol. 5, no. 1, pp. 71–86, February 1997.

[8] P. Abry and D. Veitch, "Wavelet analysis of long range dependent traffic," *IEEE Transactions on Information Theory*, vol. 44, no. 1, 1998.

[9] K. Claffy, *Internet Traffic Characterization*, Ph.D. thesis, University of California, San Diego, 1994.

[10] I. Katzela and M. Schwartz, "Schemes for fault identificaiton in communicaitons networks," *IEEE/ACM Transactions on Networking*, vol. 3(6), pp. 753–764, December 1995.

[11] F. Feather, D. Siewiorek, and R. Maxion, "Fault detection in an ethernet network using anomaly signature matching," in *Proceedings of ACM SIGCOMM '93*, San Francisco, CA, September 2000.

[12] J. Brutlag, "Aberrant behavior detection in time series for network monitoring," in *Proceedings of the USENIX Fourteenth System Administration Conference LISA XIV*, New Orleans, LA, December 2000.

[13] C. Hood and C. Ji, "Proactive network fault detection," in *Proceedings of IEEE INFOCOM '97*, Kobe, Japan, April 1997.

[14] J. Toelle and O. Niggemann, "Supporting intrusion detection by graph clustering and graph drawing," in *Proceedings of Third International Workshop on Recent Advances in Intrusion Detection RAID 2000*, Toulouse, France, October 2000.

[15] K. Fox, R. Henning, J. Reed, and R. Simonian, "A neural network approach towards intrusion detection," Tech. Rep., Harris Corporation, July 1990.

[16] N. Ye, "A markov chain model of temporal behavior for anomaly detection," in *Workshop on Information Assurance and Security*, West Point, NY, June 2000.

[17] D. Moore, G. Voelker, and S. Savage, "Inferring internet denial-of-service activity," in *Proceedings of 2001 USENIX Security Symposium*, Washington, DC, August 2001.

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

[19] R. Manajan, S. Bellovin, S. Floyd, V. Paxson, S. Shenker, and J. Ioannidis, "Controlling high bandwidth aggregates in the network," ACIRI Draft paper, February 2001.