Table 1: Key results: Comparing Web and P2P traffic

Characteristics	Web	P2P	Section
Flow Size	Introduces many mice but few elephant flows. Model: hybrid Weibull-Pareto distribution.	Introduces many mice and elephant flows. Model: hybrid Weibull-Pareto distribution.	3.1
Flow Inter-arrival time	Typically short inter-arrival time. Distribution is long-tailed. Model: two-mode Weibull distribution.	Typically long inter-arrival time. Distribution is heavy-tailed. Model: hybrid Weibull-Pareto distribution.	3.2
Flow Duration	Typically short-lived. Model: two-mode Pareto distribution.	Typically long-lived. Model: hybrid Weibull-Pareto distribution.	3.3
Flow Concurrency	Most hosts maintain more than one concurrent flow. Hosts maintain concurrent flows with a few distinct hosts.	Many hosts maintain only one flow at a time. Hosts that maintain more than one flow do so by connecting with many distinct hosts.	4.1
Transfer Volume	Large transfers are dominated by downstream traffic. Heavy-hitters account for a large portion of total transfer and their transfers follow a power-law distribution.	Large transfers happen in either upstream or downstream direction. Heavy-hitters account for a huge portion of total transfer and their transfers follow a power-law distribution.	4.2
Geography	Most external hosts are located primarily in the same geographic region.	External peers are globally distributed.	4.3

Table 2: Key results: Comparing Gnutella and BitTorrent traffic

Characteristics	Gnutella	BitTorrent
Flow Size	Both small and large flows are observed. Elephants are relatively more frequent. Distribution is heavy-tailed. Model: hybrid Lognormal-Pareto distribution.	Small flows are prevalent. Elephants are less frequent, but comparatively large. Distribution is heavy-tailed. Model: hybrid Lognormal-Pareto distribution.
Flow Duration	Typically short-lived. Distribution is heavy-tailed.	Typically long-lived. Distribution is long-tailed.
Flow Concurrency	Peers mostly connect to a single host at a time.	Peers maintain many concurrent flows with a large number of distinct hosts.
Transfer Volume	Transfers are extremely asymmetric and dominated by single direction traffic. Heavy hitters account for less volume of traffic.	Transfers are comparatively less asymmetric and more balanced. Heavy-hitters contribute more traffic volume.
Geography	External peers are mostly concentrated in the same geographic region.	External peers are from regions with broadband connectivity.

2. METHODOLOGY

2.1 Trace Collection and Traffic Identification

The network traffic traces used in this work were collected from the commercial Internet link[1] of the University of Calgary, a large research-intensive university with 28,000 students and 5,000 employees. We used `lindump`[2] running on a dual processor 1.4 GHz Pentium system with 2 GB memory and 70 GB disk space to capture TCP/IP packets via port mirroring.

Identifying P2P traffic correctly in the traces is a challenge. One approach, which has been used in some recent P2P characterization studies [17, 21, 24], is to map network traffic to applications using well-known port numbers. However, many P2P applications including BitTorrent and Gnutella use dynamic port numbers. This necessitated the use of payload signatures [11, 20] to identify applications.

We used `Bro` [15], an open source Network Intrusion Detection System, to perform the payload signature matching. The built-in payload "signature matching engine" in `Bro` was used to perform the mapping of network flows to application types. We used the signatures described by Sen *et al.* [20] and Karagiannis *et al.* [11]; details of our payload-based identification scheme can be found in [6]. We identify the start of a TCP flow using connection establishment semantics (i.e., `SYN-SYNACK-ACK` packet transmissions) or by the first packet transmission observed between hosts, and end of a TCP flow after observing a `FIN` or `RST` packet. By default, `Bro` considers a flow terminated if it is idle for more than 900 seconds.

[1]At the time of trace collection, the Internet link was a 100 Mbps full-duplex connection.

[2]`http://awgn.antifork.org/codes/lindump.c`

The payload-based identification technique requires traces with relevant application-layer headers. The signature strings for some P2P applications (e.g., Gnutella) can be buried deep inside a packet [6]; therefore, successful string matching requires full-packet payloads. This poses another challenge: the huge storage space required for full-packet trace collection from a high-speed Internet connection for an extended interval (e.g., a day or a week). For our work, we used non-contiguous one-hour traces collected between April 6 and April 30, 2006. The traces were collected each morning (9-10 am) and evening (9-10 pm) on Thursday through Sunday every week (i.e., eight one-hour traces per-week). Although discontinuous traces limit the analysis of long-term traffic behavior, we expect the traces to capture morning/evening and weekday/weekend trends. Our methodology also captured behavioral aspects related to the academic calendar.

2.2 Trace Summary

The traces contain 1.12 billion IP packets totalling 639.4 Gigabytes (GB) of data. In this paper, attention is restricted to only TCP/IP packets because these account for 84.4% of the total packets and 92% of the total bytes in the traces. Furthermore, Web and P2P applications such as Gnutella and BitTorrent use TCP in most cases. In total, we consider 23.3 million TCP flows with 946 million IP packets and 588.3 GB of data.

Table 3 shows the breakdown by application type. Web and P2P dominate in terms of bytes. Although P2P accounts for only 2.8% of the total flows, it accounts for 33.1% of the total bytes. The Unknown category includes HTTPS (port 443), flows without payloads, and flows unclassified by `Bro`. The Others category bundles together the remaining traffic; the main contributors (by bytes) are email (5%), file transfer (3%), and streaming (2%) applications.