Is P2P dying or just hiding?

Thomas Karagiannis
UC Riverside
tkarag@cs.ucr.edu

Andre Broido, Nevil Brownlee, kc claffy
CAIDA, SDSC, UC San Diego
{broido,nevil,kc}@caida.org

Michalis Faloutsos
UC Riverside
michalis@cs.ucr.edu

***Abstract*— Recent reports in the popular media suggest a significant decrease in peer-to-peer (P2P) file-sharing traffic, attributed to the public's response to legal threats. Have we reached the end of the P2P revolution? In pursuit of legitimate data to verify this hypothesis, we embark on a more accurate measurement effort of P2P traffic at the link level. In contrast to previous efforts we introduce two novel elements in our methodology. First, we measure traffic of all known popular P2P protocols. Second, we go beyond the "known port" limitation by reverse engineering the protocols and identifying characteristic strings in the payload. We find that, if measured accurately, P2P traffic has never declined; indeed we have never seen the proportion of p2p traffic decrease over time (any change is an increase) in any of our data sources.**

***Index Terms*—traffic measurements, peer-to-peer, file-sharing**

I. Introduction

Recently, popular media sources have reported a sharp decline in peer-to-peer (P2P) traffic during the last year [5] [21], with user population dropping by half. This assertion is in direct contrast to the constant increase of P2P activity over the last years. The decline has been attributed to legal issues most loudly articulated by the Recording Industry Association of America (RIAA). RIAA reports suggest that the overwhelming threats of copyright lawsuits and fines have stalled the growth of file-sharing networks. Have we reached the end of the P2P revolution as we know it?

In this paper we challenge the stated P2P reports and their conclusions, emphasizing the fact that measurements of P2P traffic are problematic. First, measurement methodologies of these analyses are usually not disclosed. Second, the studies are limited to only a small set of two or three traditional file-sharing[1] networks, and yet they unabashedly draw conclusions for the future of P2P file-sharing networking as a whole.

In reality, current file-sharing networks (including private P2P networks) provide users with a variety of options. Furthermore, an increasing number of P2P networks intentionally camouflage their traffic. Newer versions of P2P protocols can flexibly use any port number, even port 80, traditionally used for Web traffic. We no longer enjoy the fleeting benefit of first-generation P2P traffic, which was relatively easily classified due to its use of well-defined port numbers.

This paper sheds doubt on the claim that P2P traffic is declining. To do so, we develop a framework and heuristics to measure camouflaged P2P traffic. We also provide the first estimate of the percentage of P2P traffic under non-specified ports for eight different P2P protocols. We use data collected at two different OC48 (2.5Gbps) links of Tier1 Internet Service Providers (ISPs) in 2002 through 2004. Our specific contributions include:

- In our traces, P2P traffic volume has not dropped since 2003. Our datasets are inconsistent with claims of significant P2P traffic decline.
- We present a methodology for identifying P2P traffic originating from several different P2P protocols. Our heuristics exploit common conventions of P2P protocols, such as the packet format.
- We illustrate that over the last few years, P2P applications evolved to use arbitrary ports for communication.
- We claim that accurate measurements are bound to remain difficult since P2P users promptly switch to new more sophisticated protocols, e.g., BitTorrent.

In general we observe that P2P activity has not diminished. On the contrary, P2P traffic represents a significant amount of Internet traffic and is likely to continue to grow in the future, RIAA behavior notwithstanding.

II. Previous work and related statistics

P2P measurement studies have thus far been limited, usually focused on topological characteristics of P2P networks based on flow level analysis [23], or investigating properties such as bottleneck bandwidths [22], the possibility of caching [17], or the availability of content [3]. In general, the analysis and modeling community tends to neglect P2P traffic and/or assume that it generally behaves like other traffic.

P2P traffic is a significant fraction of total workload. Telefonica, Spanish and South American telecom/ISP reports over 50% of IP traffic in their network being P2P [9]. According to Sprint's IP Monitoring Project [25], for August 2002, for the majority of the monitored links in New York and San Jose, P2P traffic is approximately 20% of the total volume. In April 2003, 20-40% of total bytes corresponds to P2P traffic. Sprint analysis [8] uses Coral Reef [16] application port tables. Their data can be interpreted as P2P activity increasing or being stable in 2002-2003.

Over the same time interval, the *other TCP* traffic category in Sprint's network increased. This category includes TCP traffic that cannot be classified using known port numbers, which may imply that P2P traffic is shifting from known to arbitrary ports. This increase in unclassified traffic is consistent with comments in [13], where the authors observe an increase in unclassified TCP and web traffic when certain port numbers (Fasttrack ports) are rate-limited, implying use of nonstandard port numbers by P2P applications.

[1]We will use the terms *P2P* and *file-sharing* interchangeably although file-sharing is only a subset (but typically vast majority) of P2P traffic.

TABLE I

ONE-HOUR OC-48 TRACES ANALYZED

Set	Date	Link	Start	Direction	Src.IP	Dst.IP	Src. AS	Dst. AS	Flows	Packets	Bytes	Mean Util.
D04N	2002-08-14	B1	10:00	Northbound (0)	469 K	963 K	4270	1596	18 M	294 M	164 G	365 Mbps (14.6%)
D08N	2003-05-07	B1	10:00	Northbound (0)	189 K	725 K	2408	614	7 M	93 M	57 G	125 Mbps (5%)
D09N	2003-05-07	B2	10:00	Northbound (1)	632 K	2241 K	3505	229	30 M	459 M	293 G	651 Mbps (26.2%)
D09S	2003-05-07	B2	10:00	Southbound (0)	295 K	1307 K	599	3752	23 M	308 M	169 G	376 Mbps (15.1%)
D10N	2004-01-22	B2	14:00	Northbound (1)	812 K	2181 K	4544	411	24 M	413 M	288 G	639 Mbps (25.7%)
D10S	2004-01-22	B2	14:00	Southbound (0)	279 K	4177 K	2893	3596	19 M	253 M	117 G	260 Mbps (10.5%)

TABLE II

STRINGS AT THE BEGINNING OF THE PAYLOAD OF P2P PROTOCOLS. THE CHARACTER "0X" BELOW IMPLIES HEX STRINGS.

P2P Protocol	*String*	*Trans. prot.*	*Def. ports*
eDonkey2000	0xe3, 0xc5	TCP/UDP	4661-4665
Fasttrack	"GIVE" / 0x270000002980	TCP / UDP	1214
BitTorrent	"0x13Bit"	TCP	6881-6889
Gnutella	"GNUT", "GIV" / "GND"	TCP /UDP	6346-6347
MP2P	GO!!, MD5, SIZ0x20	TCP	41170 UDP
Direct Connect	"$MyN","$Dir" / "$SR"	TCP/UDP	411-412

III. TRACES AND ANALYZED PROTOCOLS

Table I lists general workload dimensions of our datasets: counts of distinct source and destination IP addresses and the numbers of flows, packets, and bytes observed. The processing of our traces was performed by the Coral Reef suite[16].

We analyze one-hour packet traces that are part of CAIDA's Backbone Traffic Data Kit (BTDK). Our dataset notation follows [2]. We use traces captured in August 8, 2002 (dataset D04), May 5, 2003 (D08 and D09) and January 22, 2004 (D10) by state-of-the-art Dag 4 monitors [19] and packet capture software from the University of Waikato and Endace [12]. We monitored traffic of two OC-48 (2.5Gbps) San Jose-Seattle links of two US commercial Tier 1 backbones. Our monitors capture 44 bytes of each packet, which includes IP and TCP/UDP headers, and initial 4 bytes of payload for some packets. These bytes are mostly present for Backbone 1 (**B1**) data where 66% of packets have 40-byte headers. However, approximately 75% of the packets of Backbone 2 (**B2**) are encapsulated with an extra 4-byte MPLS label which leaves no space for payload bytes.

Utilization in B2 traces averaged 25% of link capacity for northbound (San Jose to Seattle) direction. For southbound direction, utilization is slightly lower in our 2004 trace compared to 2003 (14% to 10%). These percentages reflect a typical approach for large backbone providers who overprovision capacity [8]. For B1 traces, utilization was around 15% for August 2002. However, our May 2003 B1 trace shows low utilization, approximately 5%. Thus, for traffic comparison purposes we only use our B2 traces.

We study eight of the most popular P2P protocols: *eDonkey* [10] (statistics referring to eDonkey, also include the *Overnet* and *eMule* [11] networks), *Fasttrack* which is supported by the well known Kazaa, *BitTorrent* [4], *OpenNap* and *WinMx* [28], *Gnutella*, *MP2P* [20], *Soulseek* [24] and *Direct Connect* [7].

IV. LIMITATIONS AND METHODOLOGY

Our goal is passive monitoring of P2P traffic. As flows cross our monitored link, our objective is to determine if a specific flow is P2P. Our analysis is based on identifying characteristic bit strings in packet payload, which in principle represent control traffic of P2P protocols. This section describes limitations that inhibit robust estimation of P2P traffic volume at the link level. In addition, we present our methodology to identify P2P flows.

A. Limitations

There were several issues that we had to take into consideration throughout our study. While some are data related others originate from the nature of P2P protocols. Specifically, these limitations are the following:

44-byte packets: CAIDA monitors capture 44 bytes [2] of each packet (see section III), which leaves 4 bytes of TCP packets to be examined (TCP headers are typically 40 bytes for packets that have no options). While our payload heuristics would be capable of effectively identifying all P2P packets if the whole payload was available, this 4-byte payload restriction limits the number of heuristics that can undoubtedly pinpoint P2P flows. For example, BitTorrent string *"GET /torrents/"* requires 15 bytes of payload for complete matching. Our 4-byte view of *"GET "* could potentially indicate a non-P2P web HTTP request. On the other hand, UDP header is only 8 bytes, which leaves enough payload bytes for effective string matching. However, approximately 85%-90% of all packets are transfered with TCP.

MPLS: 60%-80% of the packets in our B2 traces are encapsulated with 4-byte MPLS (Multiprotocol Label Switching) headers. MPLS is used by the ISP for routing and traffic engineering purposes. MPLS decreases the number of packets that can be matched against our string table since for a significant amount of traffic there is no payload (4-byte MPLS header + 40-byte TCP header).

HTTP requests: A number of P2P protocols uses HTTP requests to transfer files similar to web traffic. In both cases (P2P and web) the first four bytes of the payload indicate the HTTP method or code used (e.g., the four bytes of the payload would be "GET ", "HTTP" etc.). In these cases, packets could be either HTTP or P2P . Thus, a number of possible characteristic bit strings are rejected.

ISP caching: To alleviate the effect of P2P traffic, ISPs lately employ caching of P2P content [14]. P2P caching (similar to web caching) is capable of reducing upstream traffic yielding large savings for the ISPs[3]. Naturally, P2P requests that are served by these caches do not reach the backbone. Thus, caching results in a limited view of P2P usage especially when comparing with past years where such practices were not applied.

Encryption: Increasing number of P2P networks relies on encryption and ssl to transmit packets and transfer files. Payload string matching misses all P2P encrypted packets.

P2P versus copyrighted traffic: Typically, the majority of P2P traffic is related to copyrighted material. Although we cannot

[2]Privacy issues prohibit the examination of more bytes of user payload.

[3]ISPs are usually charged based on the traffic they send upstream to their own providers. In general, ISPs prefer to keep traffic generated by their customers within the boundaries of their own ASes.

necessarily equate P2P with copyrighted traffic, the above statement is largely believed to be valid. Thus, RIAA and other legal agencies monitor P2P traffic as an indication of illegal activity. These agencies are usually interested in the number of distinct users and downloaded works, while we examine P2P IP population and traffic volume as metrics for quantifying P2P trends. Our study cannot identify the trends in the use of copyrighted material.

Link utilization and time of the day: While our traces are collected during business hours, the January 2004 trace is likely to have more home user traffic, due to its capture later in the week and in the day. Two traces are never alike and present different characteristics. However, general conclusions can be reached with careful statistical analysis. We address this issue extensively in section V-C.

In fact, many limitations to this analysis, as with virtually all Internet measurement studies, are neither new nor unique to Internet science: "$\Delta\iota\varsigma\ \epsilon\iota\varsigma\ \tau o\nu\ \alpha\upsilon\tau o\nu\ \pi o\tau\alpha\mu o\nu\ o\upsilon\kappa\ \alpha\nu\ \epsilon\mu\beta\alpha\iota\epsilon\iota\varsigma$" [4].

B. Methodology

Our analysis is based on identifying specific *bit strings* in the packet payload. Since documentation for P2P protocols is generally poor, we empirically derived a set of distinctive bit strings for each case by monitoring both TCP and UDP traffic using tcpdump[26] after installing various P2P clients. Table II lists a subset of these strings for some of the analyzed protocols for TCP and UDP. Due to space limitations, we do not present our whole list of used bit strings [5]. Note that for TCP, we only use 4-byte long bit strings, since the available TCP payload in our traces is at best 4 bytes. This constraint restricts the number of bit strings that can effectively identify P2P packets.

We classify packets in flows. Flows are defined by the 5-tuple source IP, destination IP, protocol, source port and destination port. We use 64 seconds for flow timeout which is a common practice in measurement community [6], i.e., if there are no packet arrivals for a specific flow for a time period of 64 seconds, the flow expires.

To address the limitations described in the previous section, we apply four different methodologies to estimate P2P traffic. These methodologies are the following in increasing levels of aggressiveness as to which flows are considered P2P :

M1: If a source or destination port number of a flow matches one of the "well-known" port numbers (Tab.II) the flow is flagged as P2P.

M2: We compare the payload (if any) of each packet in a flow against our table of strings. In case of a match between the 4-byte payload of a packet and one of our bit strings, the flow is flagged as P2P with the proper protocol (e.g., Fasttrack, eDonkey, etc.). If none of the packets match, then the flow is considered a non-P2P flow. In case of eDonkey, since our strings consist of one byte at the beginning of the packet, we require that either all packets in a UDP flow or 10% of the packets in TCP flows with more than 20 packets match our strings[6].

M3: If a *UDP* flow is flagged as P2P from *M2*, both source and destination IPs of this flow are hashed into a table of IPs. All flows (TCP or UDP) that contain an IP from this IP table are flagged also as P2P even if for these flows there is no payload match. This kind of IP tracking is performed only for host IPs that we have identified as P2P from *M2* to avoid recursive misclassification of non-P2P flows as P2P .

M4: If a TCP flow is flagged as P2P, both source and destination IPs of this flow are hashed into a second table of IPs. All flows that contain an IP from the second IP table are flagged as "possible P2P" even if for these flows there is no payload match. Similar to *M3*, flows are classified as "possible P2P" only if IPs have been identified as P2P from *M2*.

Note that the sequence *M1* through *M4* includes all previous conditions, in the sense that P2P flows identified by *M1* as P2P will also be flagged as P2P in *M2*, *M3* and *M4*. *M2* through *M4* attack the trade-off of underestimating versus overestimating by assessing both extremes. In all P2P networks, P2P clients maintain a large number of connections open even if there are no file transfers. Thus, there is increased probability that a host identified as P2P from *M2* will participate in other P2P flows. These flows will be flagged either as P2P or "possible P2P" in *M3* or *M4* respectively. On the other hand, a P2P user may be browsing the web or sending email while connected to a P2P network. Thus, we exclude from *M3* and *M4* all flows whose source or destination port implies web, mail, ftp, ssl, dns (i.e., ports 80, 8000, 8080, 25, 110, 21, 22, 443, 53) for TCP and online gaming and dns (e.g., 27015-27050, 53) for UDP to minimize false positives [7]. In addition, *M3* and *M4* allow us to partially overcome the MPLS and encryption limitations described in section IV-A.

In general, we believe that *M3* will provide for an estimate closer to the real intensity of P2P traffic while *M2* and *M4* may be considered as loose lower- and upper bounds of its volume. Since there is enough UDP payload to safely identify all P2P UDP flows, our knowledge of the IPs participating in these flows facilitates identification of corresponding TCP flows in *M3*. Even though some P2P protocols do not use UDP (e.g., Soulseek, BitTorrent), the rest use both transfer protocols making classification easier. Note also that the purpose of our study is not to precisely quantify the percentage of P2P traffic in the backbone, but instead to affirm or refute claims on the trends of P2P file-sharing usage during the past few years.

V. Analysis of P2P traffic

We now present P2P traffic characteristics for our traces. We describe bitrates for the total volume of P2P protocols as identified by the methodology in the previous section for the analyzed protocols. In addition, we report statistics of P2P activity in numbers of participating IPs and ASes as seen by each step of our methodology. Finally, we demonstrate trends of individual P2P networks.

[4] "You cannot enter the same river twice", Heracletus of Ephesus, 500 BC.

[5] The whole list of bit strings can be found in [15].

[6] eDonkey transfers blocks of bytes that start with a specific byte (see Table II). Thus, this byte will mark the beginning of all blocks by being their first byte in the first packet of each block.

[7] Since nothing prevents P2P clients from using these ports also, excluding specific protocols by looking at port numbers may result in underestimating P2P traffic.

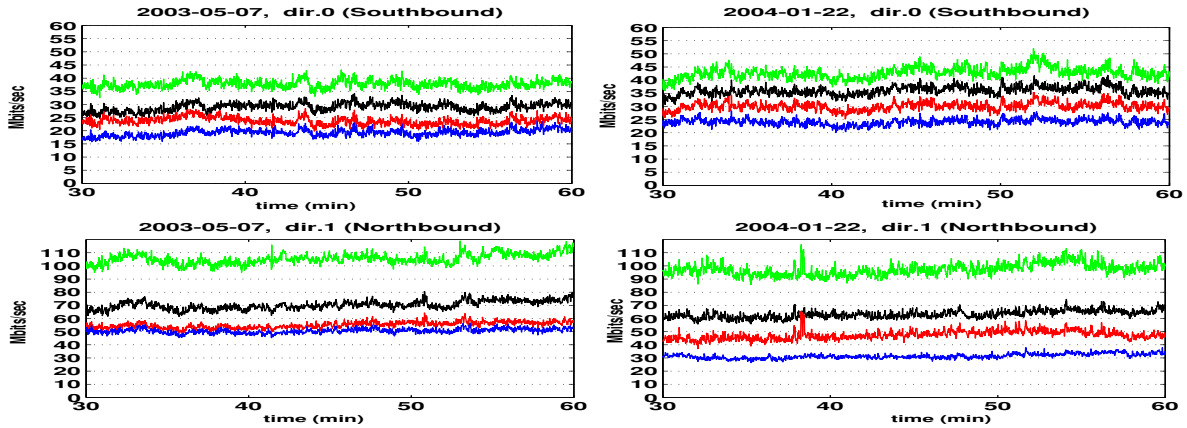

Fig. 1. Bitrate of total P2P traffic as seen by our methodology for May 2003 and January 2004 for backbone 2. Each line presents a more aggressive estimate (*M1* - *M4*) of P2P traffic starting from the lower line. *M2* and *M4* provide loose lower and upper bounds whereas *M3* is a more realistic estimate of P2P traffic. Plots are presented chronologically from left to right, and categorized by direction (S or N) from top to bottom. P2P traffic is comparable across the year with an increase in absolute numbers in direction 0, where utilization has decreased in 2004 and a small decrease in direction 1 (approximately 1% change.)

A. Bitrate of P2P traffic

We demonstrate that the percentages of P2P traffic on the observed links are comparable between 2003 and 2004.

Fig. 1 shows the total P2P bitrate for both May 2003 (D09) and January 2004 (D10) for B2 for the last 30 minutes of the traces. These bitrates correspond to the sum of the bitrates of the eight analyzed protocols described in section II. The upper portion of the figure presents direction 0 (S) while the bottom shows direction 1 (N) of B2 traces. From left to right we present the traces chronologically. Each plot shows 4 different bitrates. Starting from the lower line, each rate represents a more aggressive estimate of total P2P bitrate computed by:

- *M1* (i.e., only the "well-known" port numbers),
- *M2* (i.e., payload heuristics),
- *M3* (i.e., P2P UDP IP tracking),
- *M4* (i.e., P2P TCP/UDP IP tracking)

The axes of each plot for the same direction are presented in consistent scaling to facilitate comparison. We consider *M2* and *M4* as lower and upper bounds respectively of the true intensity of P2P traffic. We choose *M2* instead of *M1* as a lower bound, since *M1* is port-based and greatly underestimates P2P traffic; in other words, the volume of P2P traffic is at least the amount shown by *M2*. We present *M1* as a reference to indicate what the port-based estimation would be in each case.

These traces demonstrate that the level of P2P traffic volume in January 2004 is similar to 2003. Definitely, Fig. 1 does not contribute to the claims of significantly declining P2P usage trends. More specifically, on the average for direction 0 (S) of the link, total P2P traffic increased from May 2003 to 2004 even for standard port numbers. Using the payload heuristic (*M2*), there was an increase from 23.8 to 30.1 Mbps (6.5% to 11.6% of the whole observed traffic), and according to *M3* from 26 to 33.4 Mbps (8% to 14%). On the other hand, for direction 1 (N), there was a small decrease from 55.2 to 47.5 Mbps (8.4% to 7.4%) according to *M2* and from 70.3 to 63.2 Mbps (10.7% to 9.9%) with *M3*.

Our ability to match packet payload depends upon whether IP packets are encapsulated with MPLS. The percentage of MPLS encapsulated packets is reflected in the relative difference between *M1* through *M4* lines in each plot. For instance, *M2* estimate for D09N (lower left plot) is very close to *M1* (i.e., bitrate computed by port numbers only). On the contrary, for D10N (lower right plot) the difference between *M1* and *M2* is larger. These differences reflect MPLS traffic percentages on the link. MPLS traffic was 80% in D09N versus 63% in D10N. Thus, more payload was available in D10N, allowing for classification of a larger number of flows as P2P, pushing *M2* away from *M1* (implying increasing number of flows using nonstandard port numbers). For direction 0, MPLS traffic percentages were 77% for 2003 (D09S) vs. 63% for 2004 (D10S).

B. Realistic estimate of P2P traffic

Our conjecture is that *M3* provides for a realistic estimate of P2P traffic. *M2* and *M4* represent the range of the possible P2P traffic bitrate. However, we accept *M3* as representative of true P2P traffic volume. Our belief is reinforced by observations of our older traces, where fewer P2P protocols supported file transfers in arbitrary port numbers.

Fig. 2 presents P2P bitrates for our B1 traces (D04N, August 2002 and D08N, May 2003) in similar fashion to Fig. 1. The two top plots present the absolute volume of total P2P traffic, whereas the bottom ones present the traffic in terms of link utilization. While comparison is risky due to substantial difference in utilization, the percentage of P2P traffic did increase (approximately 5%) relative to traffic volume [8].

However, the point of the figure is the spacing of the bitrate estimates. In 2002, only eDonkey and FastTrack transfered packets in arbitrary ports in D04N resulting in *M1* and *M2* lines (two bottom lines) being closer. In older P2P clients, use of arbitrary port numbers, if supported, was optional. In contrast, current P2P clients randomize the port number upon installation without requiring user intervention.

More important is the spacing between *M2* and *M3*. In both traces, *M2* and *M3* estimates are literally equal, since in B1 traces there is no MPLS traffic and payload examination is more effective. Thus, we conjecture that *M3* produces a realistic assessment of P2P traffic in B2 traces as well.

[8] Link utilization decreased from 14.6% to 5%, see Table I.

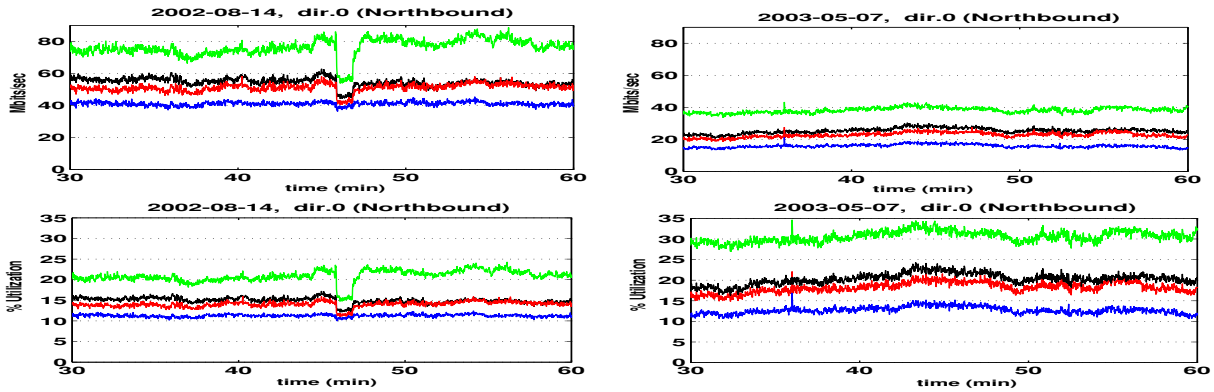

Fig. 2. Absolute volume (top row) and percentage relative to link utilization (bottom row) of P2P traffic in August 2002 and May 2003 for backbone 1. Each line presents a more aggressive estimate (*M1* - *M4*) of P2P traffic staring from the lower line. *M3* estimate closely follows *M2* (payload) since there is no MPLS traffic in the link. Despite the large difference in link utilization between 2003 and 2004, there is an increase of total P2P traffic approximately by 5%.

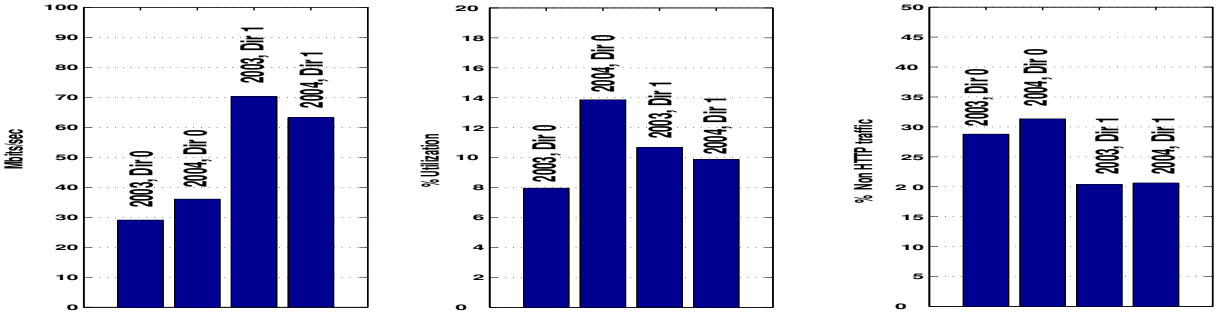

Fig. 3. Volume of P2P traffic from three different points of view. In all cases, our traces show an increase for direction 0. If we consider the percentage of P2P relative to non-HTTP traffic, then P2P traffic increased for direction 1 also.

TABLE III

VOLUMES OF HTTP, SMTP AND P2P TRAFFIC IN OUR B2 TRACES IN MBPS AND PERCENTAGE OF TRAFFIC VOLUME.

	20030507, B2		*20040122, B2*	
	Dir. 0 (D09S)	Dir. 1 (D09N)	Dir. 0 (D10S)	Dir. 1 (D10N)
HTTP	264 (72%)	314 (47.7%)	145 (56%)	333 (52.1%)
SMTP	4.8 (1.3%)	7.7 (1.2%)	8.4 (3.2%)	9.7 (1.5%)
P2P	26 (8%)	70.3 (10.7%)	33.4 (14%)	63.2 (9.9%)

TABLE IV

NUMBER OF SOURCE AND DESTINATION IPS AND ASES OF P2P TRAFFIC.

		20030507, B2		*20040122, B2*	
		Dir. 0	Dir. 1	Dir. 0	Dir. 1
SrcIPs (ASes)	*M1*	14K (245)	77K (1389)	12K (310)	75K (1178)
	M2	18K (256)	108K (1472)	15K (325)	105K (1281)
	M3	24K (256)	139K (1507)	20K (328)	140K (1411)
	M4	31K (267)	139K (1624)	54K (360)	146K (1481)
Dst IPs (ASes)	*M1*	69K (1379)	23K (127)	90K (1074)	32K (246)
	M2	94K (1494)	27K (127)	124K (1153)	39K (249)
	M3	105K (1494)	34K (127)	150K (1242)	57K (269)
	M4	155K (1871)	155K (134)	517K (1585)	343K (271)

C. Statistical confidence of results

In order to examine how valid for comparison our two snapshots of the traffic are, we study P2P traffic in conjunction with other statistical properties of the traces. While effects regarding time of day and utilization variation are difficult to quantify in measurement studies, this comparison is required to increase confidence in our findings. To achieve that, we measure "known" types of traffic to obtain a relative view of P2P versus non-P2P traffic. In addition, we examine the P2P population as identified by number of distinct IPs and Autonomous Systems (ASes). To map IPs to ASes, we use the AS Finder module from Coral Reef suite[16] and BGP tables from Route Views. Finally, we compare common source-destination prefixes and ASes that appear in both years.

P2P relative to "well-known" traffic: To increase our statistical confidence that P2P traffic is comparable across both years, we examined P2P traffic relative to statistics of other traffic types. Table III presents the volumes of HTTP, SMTP and P2P traffic in our traces. Direction 1 appears to have stable characteristics. SMTP and HTTP are comparable between 2003 and 2004, approximately 1.5% and 50% of the total traffic volume. Direction 0 statistics show more variation across the year. HTTP has dropped approximately 15% of the traffic volume and SMTP almost doubled. On the other hand, non-HTTP traffic is comparable for direction 0 between 2003 and 2004 (101 Mbps and 115 Mbps respectively). Thus, difference in direction 0 link utilization was mostly due to different HTTP volumes.

Fig. 3 presents average intensity of P2P traffic from three different perspectives: absolute volumes, percentage relative to total volume and percentage relative to non-HTTP traffic. In all cases, there is an increase in P2P traffic for direction 0. Relative to non-HTTP traffic there is even an increase for direction 1, which was not the case with absolute volume or relative share of P2P. These observations reinforce our conclusions that our traces do not support a significant decrease in P2P traffic. In fact, intensity of P2P traffic appears at least comparable with past years, if not larger. For Fig. 3 we use P2P volumes computed by *M3*, since we believe that it is the realistic estimate of P2P traffic. *M1* and *M2* present similar findings.

TABLE V

CHANGE OF P2P TRAFFIC IN COMMON PREFIX/AS-PAIRS (SRC-DST) BETWEEN 2003 AND 2004, BY M3 AND THE NUMBER OF PAIRS WHERE P2P TRAFFIC INCREASED (ROW 1) AND DECREASED (ROW 2). ALSO, PREFIX/AS PAIRS COMMON IN D09 AND D10, SUCH THAT IN 2003 THEY HAVE NO P2P TRAFFIC AND IN 2004 THEY DO (ROW 3), AND VICE VERSA (ROW 4). THE NUMBERS IN PARENTESES SHOW THE AVERAGE BITRATE INCREASE/DECREASE CAUSED BY THE CORRESPONDING PAIRS.

	Common P2P prefix-pairs		*Common trace prefix-pairs*		*Common P2P AS-pairs*		*Common trace AS-pairs*	
	Dir. 0	Dir.1	Dir. 0	Dir.1	Dir. 0	Dir.1	Dir. 0	Dir.1
# Pairs Increased (Mbps)	5371 (4)	5847 (10.2)	9082 (4.9)	9354 (13.8)	1105 (10.4)	788 (5.8)	2036 (11.1)	1743 (7.8)
# Pairs Decreased (Mbps)	4650 (4.7)	6226 (15.1)	7733 (5.3)	9560 (18.2)	823 (3.8)	777 (9.8)	1701 (5.3)	1724 (17.8)
# New P2P pairs in 2004 (Mbps)	-	-	3711 (0.96)	3507 (3.5)	-	-	931 (0.6)	955 (2)
# P2P pairs in 2003 only (Mbps)	-	-	3084 (0.78)	3334 (3)	-	-	878 (1.5)	947 (7.9)

Monitoring P2P population: We examine how the population of IPs and ASes that participate in P2P flows changes. Our observations confirm that instances are qualitatively comparable, with P2P population slightly larger in 2004. In total, for 2004 traces, there are approximately 60,000 distinct IPs more in P2P flows in 2004 than in 2003. Table IV presents the number of distinct source and destination IPs that participated in P2P flows according to our methodology.

More specifically, there appear to be 4,000 fewer distinct source IPs for 2004, in D10S compared to D09S. However, the number of ASes participating in P2P flows is larger, approximately 70 more in 2004 than in 2003. For direction 1, there are 1,000 more P2P IPS for D10N compared to D09N, considering *M3*.

The population of destination P2P IPs increased for both directions. However, for direction 0 the number of total destination IPs in the traces has increased four-fold in 2004, whereas for direction 1 the population is similar. Thus for direction 0, where the number of P2P IPs is larger (50,000 more considering *M3*), comparison is risky[9] (normalizing these numbers using the total number of destination IPs in the traces yields a decrease of P2P destination IPs from 8.1% to 6.5% of the total IPs). For direction 1 where total destination IPs are comparable, destination P2P IPs and ASes almost doubled in D10N.

Finally, the total P2P population (source and destination IPs) has increased for both directions in absolute numbers. There is an increase of 40,000 and 25,000 distinct IPs, for direction 0 and direction 1 respectively.

Monitoring common address space: In an effort to further corroborate our findings, we attempt to isolate routing effects from P2P traffic variation across the two years. To achieve that, we compare source-destination prefix and AS pairs that appear in both instances. Our results confirm previous findings in the paper.

Specifically, Table V presents population variations within common prefixes and ASes in both traces for each direction as seen by *M3*. We mapped every IP in our traces to a prefix or an AS based on the ISP routing table as seen in Route Views BGP tables for the specific dates of our traces. Then, we examined two categories of common source-destination pairs of prefixes/ASes in our traces: a) Common P2P pairs, i.e., P2P source-destination pairs that are seen both in 2003 and 2004. b) Common trace pairs, i.e. source-destination pairs that exist in both traces, but not necessarily participate in P2P flows.

Table V agrees with general observations in previous sections. Rows 3,4 have blank cells since all common P2P pairs exist in both years. However, common trace pairs do not necessarily appear in both traces in P2P flows. In general, in all cases of southbound direction, the number of pairs where P2P traffic increased since 2003, is larger compared to the number of pairs where P2P traffic decreased. For direction 1, the number of prefixes with source-destination pairs where traffic increased versus prefixes where traffic decreased is comparable (falls within equality tests considering mean + 3*standard deviation). While average bitrates point to a decrease of P2P traffic volume in common prefixes (numbers in parentheses), this difference only represents a minor portion of total utilization in the link, less than 1%.

[9] IP address scans often inflate destination counts [2].

D. Trends of P2P protocols

Examining each protocol separately reveals interesting trends regarding the evolution of P2P networks. Fig. 4 shows the average bitrate of each analyzed P2P protocol in our traces. Similarly, each of the four bars for every protocol represent the four estimates of our methodology. Despite the fact that protocol bitrates might reflect idiosyncrasies of our monitored link, general observations for Fig. 4 are the following:

- BitTorrent bitrate has increased more than 100% in absolute numbers for both directions of the link. BitTorrent has evolved into one of the most popular networks, surpassing Fasttrack traffic.
- Fasttrack portion of P2P traffic has dropped in agreement with media reports. However, the difference between port numbers (*M1*) and payload heuristics (*M2*) bitrate estimates has increased. Thus, Fasttrack traffic appears to be shifting to arbitrary port numbers with time. This assumption is validated by the larger difference between *M1* with *M2* and *M3* in 2004 than in 2003.
- eDonkey, WinMx and Gnutella have comparable portions of total P2P traffic between 2003 and 2004 (Fig. 4 presents only absolute numbers).

These findings exhibit how the existence of increasing network options for P2P users has affected P2P traffic. For example, users might have shifted from Fasttrack to BitTorrent to avoid potential legal issues (the vast majority of RIAA lawsuits targeted Fasttrack users).

VI. CONCLUSIONS - DISCUSSION

This paper emphasizes two main points. First, P2P is here to stay. Our link level measurements show that P2P traffic is at least comparable to last year's levels, if it hasn't increased. An increase in P2P activity over the same period has been observed in at least one study [9], and one user survey [1]. Second, measuring P2P traffic becomes problematic with conventional measurement methodologies resulting in underestimating P2P traffic.

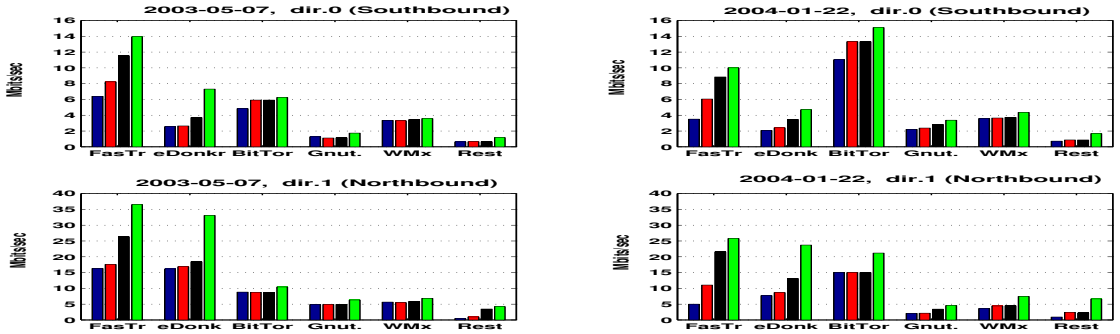

Fig. 4. Average bitrate of P2P protocols as identified by our methodology. Bars present *M1* - *M4* starting from left to right for each protocol. BitTorrent has increased more than 100% while Fasttrack portion of dropped.

The use of non-standard, arbitrary ports is the first level of complication. In addition, packet encryption will eventually make payload heuristics inapplicable.

The significance of these observations is multifaceted. Due to space limitations, we can only highlight the more direct effects.

P2P users vs. the entertainment industry. According to our results, the P2P battle seems to be entering a new phase with the P2P community making its second comeback. The first comeback was the switch from the easy-to-locate Napster, to distributed Gnutella-like protocols. Thus, locating a single responsible entity became impossible. The industry then relied on detecting P2P traffic. Now, the users take a step further by making P2P traffic hard to identify.

Network economics. The increase in P2P traffic is a mixed blessing for end-user ISPs. P2P fuels the demand for home broadband (e.g., DSL) connections; however, the fixed monthly fee paid by home users may not cover the ISP's expenses caused by volume-based charges of upstream providers: flat rate at the network edge is in direct conflict with usage-based charges imposed by carriers. [27].

Another trend that is currently gaining momentum is an intent to directly manipulate P2P applications into desirable traffic patterns [9], e.g. exchanging most of the data inside the ISP's infrastructure. This trend may result in ISPs competing to provide better bitrates for rate-aware applications like BitTorrent [4], in accordance with their economic relations (upstreams pushing more traffic to customers and customers trying to minimize traffic exchanged with upstreams.)

Breaking the asymmetrical bandwidth assumption. If P2P traffic continues to increase and legal complications are overridden, the P2P paradigm will bring dramatic changes in supply and demand in edge and access networks. Bit rates of many access links, in particular for DSL and cable modems, are currently provisioned asymmetrically with significantly lower upstream bandwidth. This provisioning was based on the expectation of users downloading much more data than they send upstream. The relevance of such technologies will be challenged and their market share will dwindle if alternative broadband technologies can offer comparable upstream and downstream performance.

The effect of P2P could propagate from the access points upward the network hierarchy to Tier 2 and even Tier 1 ISPs creating the need for more peering among ISPs. Current practices require balanced bidirectional load among peers[10], a stipulation much easier to achieve with symmetric link utilizations as the norm. There is no doubt that the P2P paradigm will change Internet engineering as we know it today. Given the observed trends, the only remaining question is when, not if.

[10] "The ratio of the aggregate amount of traffic exchanged between the Requester and the WorldCom Internet Network with which it seeks to interconnect shall be roughly balanced and shall not exceed 1.5:1." [18].

REFERENCES

[1] Music group targets file swappers. In *Wall Street Journal Europe*, Jan.26 2004.
[2] A.Broido, Y.Hyun, R.Gao, and kc claffy. Their share: diversity and disparity in ip traffic. In *PAM*, 2004.
[3] R. Bhagwan, S. Savage, and G. Voelker. Understanding Availability. In *IPTPS 03*, 2003.
[4] BitTorrent. http://bitconjurer.org/BitTorrent/.
[5] John Borland. RIAA threat may be slowing file swapping. http://news.com.com/2100-1027-1025684.html.
[6] K. Claffy, H.-W. Braun, and G. Polyzos. A Parametrizable methodology for Internet traffic flow profiling. In *IEEE Journal on Selected Areas in Communications*, 1995.
[7] Direct Connect. http://www.neo-modus.com/.
[8] C.Fraleigh e.a. Packet-Level Traffic Measurements from the Sprint IP Backbone. In *IEEE Network*, 2003.
[9] J.E.Gabeiras e.a. Estrategias para influir en el tráfico p2p. In *Telefonica Investigación y Desarrollo*, 2004.
[10] eDonkey2000. http://www.edonkey2000.com/.
[11] eMule. http://www.emule-project.net/.
[12] Endace. Measurement Systems, 2004. www.endace.com.
[13] A. Gerber, J. Houle, H. Nguyen, M. Roughan, and S. Sen. P2P The Gorilla in the Cable. In *National Cable & Telecommunications Association(NCTA) 2003 National Show*, 2003.
[14] Joltid. http://www.joltid.com.
[15] T. Karagiannis, A. Broido, N. Brownlee, kc claffy, and M. Faloutsos. File-sharing in the Internet: A characterization of P2P traffic in the backbone. Technical report., 2004. http://www.cs.ucr.edu/∼tkarag.
[16] K. Keys, D. Moore, R. Koga, E. Lagache, M. Tesch, and k. claffy. The architecture of the CoralReef: Internet Traffic monitoring software suite. In *PAM*, 2001.
[17] N. Leibowitz, A. Bergman, Roy Ben-Shaul, and Aviv Shavit. Are File Swapping Networks Cacheable? Characterizing P2P Traffic. In *7th IWCW*, 2002.
[18] MCI. Worldcom policy for settlement-free interconnection with internet networks, 2003. http://global.mci.com/uunet/peering/.
[19] J. Micheel, S. Donnelly, and I. Graham. Precision timestamping of network packets. In *ACM Sigcomm (IMW)*, 2001.
[20] MP2P. http://www.slyck.com/mp2p.php.
[21] Pew Internet & American Life Project. Sharp decline in music file swappers: Data memo from PIP and comScore Media Metrix, January, 2004. http://www.pewinternet.org/reports/.
[22] S. Saroiu, P. K. Gummadi, and S. D. Gribble. A Measurement Study of Peer-to-Peer File Sharing Systems. In *MMCN*, 2002.
[23] S. Sen and J. Wang. Analyzing Peer-to-Peer Traffic Across Large Networks. In *ACM SIGCOMM IMW*, 2002.
[24] Soulseek. http://www.slsknet.org/.
[25] Sprint. Packet Trace Analysis. http://ipmon.sprintlabs.com/.
[26] tcpdump. http://www.tcpdump.org/.
[27] J.Roberts (France Telecom), 2004. Private communication.
[28] WinMx. http://www.winmx.com/.