On Blind Mice and the Elephant*

Understanding the Network Impact of a Large Distributed System

John S. Otto† Mario A. Sánchez† David R. Choffnes†
Fabián E. Bustamante† Georgos Siganos‡
† Northwestern University ‡ Telefónica Research

ABSTRACT

A thorough understanding of the network impact of emerging large-scale distributed systems – where traffic flows and what it costs – must encompass users' behavior, the traffic they generate and the topology over which that traffic flows. In the case of BitTorrent, however, previous studies have been limited by narrow perspectives that restrict such analysis.

This paper presents a comprehensive view of BitTorrent, using data from a representative set of 500,000 users sampled over a two year period, located in 169 countries and 3,150 networks. This unique perspective captures unseen trends and reveals several unexpected features of the largest peer-to-peer system. For instance, over the past year total BitTorrent traffic has *increased by 12%*, driven by 25% increases in per-peer hourly download volume despite a 10% decrease in the average number of online peers. We also observe stronger diurnal usage patterns and, surprisingly given the bandwidth-intensive nature of the application, a close alignment between these patterns and overall traffic. Considering the aggregated traffic across access links, this has potential implications on BitTorrent-associated costs for Internet Service Providers (ISPs). Using data from a transit ISP, we find a disproportionately large impact under a commonly used burstable (95th-percentile) billing model. Last, when examining BitTorrent traffic's paths, we find that for over half its users, most network traffic never reaches large transit networks, but is instead carried by small transit ISPs. This raises questions on the effectiveness of most in-network monitoring systems to capture trends on peer-to-peer traffic and further motivates our approach. [1]

Categories and Subject Descriptors

C.2.4 [**Communication Networks**]: Distributed Systems—*Distributed applications*; C.2.5 [**Communication Networks**]: Local and Wide-Area Networks—*Internet*; C.4 [**Performance of Systems**]: Measurement techniques

General Terms

Experimentation, Performance, Measurement

Keywords

Internet-scale Systems, Peer-to-Peer, Evaluation

*A variation on the Indian fable of the seven blind men and the elephant.

SIGCOMM'11, August 15–19, 2011, Toronto, Ontario, Canada.

1. INTRODUCTION

The network impact of popular, widely distributed services has implications for capacity planning, traffic engineering, and interdomain business relationships. Accurately characterizing and understanding this impact requires a view of the service in question that includes not only the traffic it generates and the networks it traverses, but also the underlying user behaviors that drive it. Such a comprehensive view is typically impossible to capture from any single network or in-network monitoring system.

It comes as no surprise, then, that the overall impact of BitTorrent, arguably the most widely distributed peer-to-peer system, remains unknown.

In this paper, we present the first comprehensive study of BitTorrent based on a longitudinal, representative view from the network edge, including two years of application traces from over 500,000 user IPs located in 3,150 ASes and 169 countries.

Our study reveals BitTorrent usage trends and traffic patterns that have been previously hidden or obscured by limited perspectives. After demonstrating the representativeness of our dataset as a sample of the overall BitTorrent system (Sec. 3), we discuss trends regarding how users interact with the system. We find that while the number of concurrent active users has decreased since 2008, the overall volume of traffic that BitTorrent generates has grown by 12%, likely due in part to increased bandwidth capacities. While session times have also decreased over this period (by 21%), the temporal patterns behind these sessions are increasingly aligned with those of rest of Internet traffic, despite BitTorrent's known high-bandwidth demands. This shift in usage patterns suggests an increasing role for BitTorrent on ISPs' infrastructure and costs.

After describing key trends regarding *how* BitTorrent is being used, we then focus on *where* the corresponding traffic is flowing. We leverage hundreds of millions of traceroute measurements between peers to map the vast majority (89%) of BitTorrent traffic to the networks it traverses. Our analysis reveals that the traffic surprisingly exhibits significant locality across geography (32% of BitTorrent traffic stays in the country of origin) and networks (49% of traffic is intradomain or crosses a single peering or sibling AS link). Using a recent network classification scheme, we unexpectedly find that most traffic does not reach the core of the network. Among other points, this raises questions on the effectiveness of in-network monitoring approaches to capture trends on BitTorrent traffic and further motivates our approach.

Using information about where BitTorrent traffic is flowing and when it is generated, we model its contribution to ISPs' costs. We observe that, under the 95th-percentile billing model typically used between transit ISPs and their customers, the *time* at which traffic occurs can be as important as the volume of traffic. Incorporating traffic volumes from a large, global ISP, using this cost model, we find that current time-of-day patterns of BitTorrent often result in significantly higher cost, byte-for-byte, when compared to other traffic on the network.

In sum, our results highlight how limited perspectives for analyzing Internet-wide systems do not generalize, demonstrating the need for comprehensive views when analyzing global features of such widely distributed systems. One application of our analyses is enabling ISPs to better understand the impact of such systems and reason about the effects of alternative traffic management policies.

2. BACKGROUND AND RELATED WORK

P2P systems have received much attention from operators and the research community due in part to their widespread popularity and their potential network impact. Among P2P systems, BitTorrent is the most popular one, potentially accounting for between 20% and 57% of P2P file-sharing traffic [17, 22]. A number of studies provide detailed summaries of the BitTorrent protocol, conventions and dynamics [11, 12, 19, 20]. In this paper, we focus on data connections between peers, the flows they generate, the network paths they traverse and their temporal characteristics.

Numerous studies have analyzed P2P usage trends and attempted to characterize the overall network impact from various perspectives based on either simulations or limited perspectives [14, 15, 17, 21, 22, 25]. Conclusions vary considerably among studies, due in part to variations in P2P usage in each ISP and the challenges with identifying P2P traffic from network flow summaries (e.g., due to randomized ports or use of connection encryption). Our study is the first to examine the network impact of the BitTorrent P2P system, based on the perspective of a set of users distributed over several thousand networks worldwide. Since these traces are gathered from within the application, they are not subject to classification errors.

Given the potential impact of P2P-associated cross-ISP traffic on network operational costs, several studies have investigated approaches to evaluate and improve P2P traffic's locality [5, 6, 13, 16, 18, 30]. Xie et al. [30] base their results on testbed evaluations in a small number of ISPs, Piatek et al. [18] use a single vantage point outside of classical research platforms, and Cuevas et al. [6] simulate peer interactions based on information derived from tracker scrape results. As in some of our previous work [5], we rely instead on a global view and actual BitTorrent connections to evaluate locality aspects of this system. Here we move beyond coarse-grained locality analysis in an attempt to understand the cost associated with BitTorrent traffic using a detailed Internet map that combines public BGP feeds with peer-based traceroute data [3]. As we demonstrate in Section 5.1, network paths collected from end-users are indispensable to determine the path that BitTorrent traffic takes through the network.

Understanding how P2P-associated traffic affects an ISP's transit charges is important for determining subscriber charges and informing traffic engineering policies. Following the approach used by Stanojevic et al. [23] (which examined the cost impact of individual ISP subscribers' traffic), we are the first to apply the game-theoretic Shapley analysis to examine the relative cost of interdomain BitTorrent traffic under the common 95th-percentile charging model.

3. DATASETS

We now describe the traces we use in the rest of this study. We posit that this dataset comprises the first comprehensive and representative view of BitTorrent. The following paragraphs demonstrate each of these properties in turn.

3.1 A Comprehensive View of BitTorrent

Our study is based on the largest collection of detailed end-user traces from a P2P system. Specifically, we use data gathered through users of the AquaLab's ongoing Ono [5] and NEWS [4] projects, our *vantage points (VP)*, collectively representing more than 1,260,000 installations. Our data collection software, implemented as extensions to the Vuze BitTorrent client [28], periodically report application and network statistics, excluding any information that can identify the downloaded content.[2] This dataset is comprehensive in that it is longitudinal across time and covers a broad range of networks and geographic regions.

To inform BitTorrent usage trends during the past year (Sec. 4), we use data from the second week of November 2008 and every two months from November 2009 through November 2010 (about 1 TB of trace data). For our detailed study of BitTorrent traffic, we use continuous data from March through May 2010 (Secs. 5-6). Altogether, our dataset includes traces from more than 500,000 IPs located in 3,150 ASes and 169 countries.

This dataset includes per-connection transfer data, such as source and destination (our vantage points and the peers they connect to), current transfer rates at 30-second intervals, and the cumulative volume of data transferred in each direction for each connection. It also allows us to compute user session time, i.e. the length of time that a user runs BitTorrent.

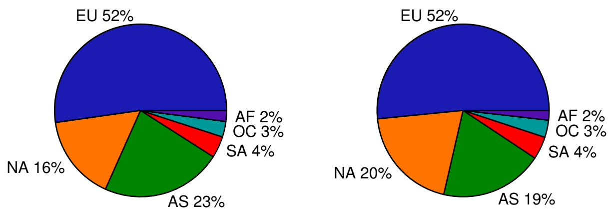
In addition to passively gathered data, the dataset contains traceroutes to a subset of peers connected to each vantage point. Targets for the probes are selected at random from connected peers, and at most one traceroute is performed at a time (to limit probing overhead). Each measurement is performed using the host's built-in traceroute command. From March through May 2010, our dataset comprises 202 million traceroute measurements. In Sec. 5.1, we discuss how we use these measurements to map per-connection flows to the AS paths they traverse.

3.2 Representativeness

We now analyze the representativeness of our dataset as a sample of activity in the Internet-wide BitTorrent system. While the vantage points are limited by the set of users who voluntarily install our extensions, we do not expect to find any strong platform or language-specific bias that could impact our results. The Vuze BitTorrent client, as well as our two instrumented plugins, run on all major platforms and are translated into nearly every language. There are several other potential sources of bias such as extension-specific behavior and the distribution of vantage points in terms of geography and networks, the peers they connect to, and the BitTorrent clients those peers use. We address these in the following paragraphs.

We first account for bias introduced by the extensions that our VPs run. While NEWS uses BitTorrent traffic to detect service-level network events without affecting the application, Ono biases peer selection using CDN redirections to reduce cross-ISP traffic.

[2] Anonymized traces are available to researchers through the EdgeScope project. [10]



(a) Vantage points (b) Connected peers

Figure 1: Distribution of BitTorrent users by continent for our vantage points (left) and the set of all peers connected to our vantage points (right) in November 2010. Both distributions match closely but for small differences in North America and Asia.

Network type	Vantage points	Remote peers
Large transit	4.07 %	5.18 %
Small transit	53.7 %	51.6 %
Access providers	42.3 %	43.2 %

Table 1: Portion of VPs and remote peers located in each type of network, suggesting that the VP population is representative of the general BitTorrent population.

We avoid this bias by filtering out all data for connections that Ono selected for preferred peering.

Given the default random peer selection strategy in BitTorrent, we expect that the remaining peers that the VPs connect to are representative of the system as a whole. We evaluate this by testing for similarities in geography, network topology, and BitTorrent client use.

We compare the VPs' geographic distribution to that of the peers they connect to, using geolocation information obtained from a popular IP-to-ASN mapping service [24]. Figure 1 shows, side by side, the distribution of VPs and their connected peers per continent. These distributions match closely, having equal portions in Europe (52%), with small differences of 16/20% in North America, and 23/19% in Asia.

For reference, we also compare the observed distributions with those reported in previous studies and find strong similarities. Zhang et al. [31] crawl tracker sites from 2008 to 2009 and report the number of BitTorrent users for the top-20 countries sorted by peer population. These data show that 49% of users are in Europe, 28% in North America and 18% in Asia. Note that these statistics under-represent the fractions of European and Asian peers since many countries in these regions did not appear in the top-20 list. This explains why North American peers are a larger portion of the peers (28%) relative to our distributions (16/20%). The reported fractions of European and Asian peers match closely those found in our dataset.

We also compare the distribution of VPs and remote peers in terms of network topology, by contrasting the number of IP addresses per network class, following the scheme recently proposed by Dhamdhere and Dovrolis [7]. Table 1 shows that these distributions align – the portion of peers in each network class differs by at most 2%.

Last, we determine whether there is any significant bias based on the types of BitTorrent clients that our VPs connect to. This could result, for example, from Vuze peers preferentially connecting to other Vuze peers. We evaluate this by comparing the distribution of BitTorrent clients connected to our VPs with that of an independent source (Table 2). We use VP data from November 2009 and a client distribution collected in August 2009 that is derived from crawls of 400 swarms [27]. As the table shows, there is a strong correspondence in both client rank and market share between the two sets.

Client	Our Data (Nov 2009)	Aug 2009 [27]
μTorrent	50.59 %	56.81 %
Azureus/Vuze	22.48 %	18.13 %
Mainline	9.28 %	11.79 %
BitComet	5.29 %	4.71 %
Transmission	2.68 %	2.95 %

Table 2: Comparison between connected client distribution in our dataset in November 2009 and results from a swarm crawl conducted in August 2009.

Overall, we find no strong evidence of significant bias in our dataset using any of these metrics. In the following sections, we use this dataset to analyze the network impact of BitTorrent, starting with a description of observed usage patterns and trends.

4. BITTORRENT USAGE TRENDS

In this section, we use our BitTorrent traces to analyze several key usage trends that affect the system's network impact. In particular, we find that despite reports of declining usage [17] the absolute volume of BitTorrent traffic continues to rise. Further, we find that BitTorrent's temporal usage patterns are increasingly aligned with diurnal traffic patterns, which has implications for its contribution to ISPs' costs.

4.1 Sampling Methodology

Obtaining representative, longitudinal snapshots of BitTorrent traffic and user behavior is challenging, given the high degree of churn in the system. To enable comparisons across multiple time scales, we aggregate user statistics at one-hour granularity, then use random sampling of our dataset to obtain a constant number of users (1000) during each hour for inclusion in our analysis. We repeat this random sampling 5 times to derive a statistically significant average for each hour. In general, the standard deviations of our samples are relatively small, and we include corresponding error bars in our figures.

The following analysis focuses on comparisons using the second week of every 2 months over the year of study. In some cases, we use data collected in November 2008 to analyze trends over two years.

4.2 Key Trends

In the following paragraphs, we examine trends in overall BitTorrent traffic in terms of number of connected peers and per-peer download volumes.

We begin by examining the volume of traffic generated by individual peers. Figure 2 depicts average per-peer hourly download volumes over one year, between November 2009 and 2010. As the figure shows, BitTorrent's network impact in terms of hourly download volumes per-peer have grown consistently over this period, increasing by 25% on average.

Beyond the download volumes generated by users, we find two important and related trends that refer to (1) the number of BitTorrent users and (2) the temporal patterns behind their

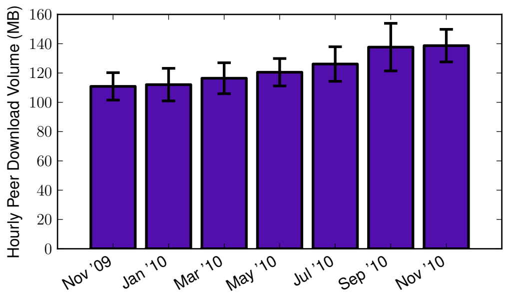


Figure 2: Average per-user hourly download volume between November 2009 and 2010, showing that download volumes have increased by 25%.

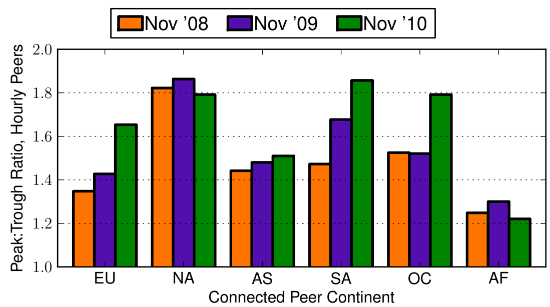


Figure 4: Average daily peak-to-trough ratio of hourly peers seen by continent, from 2008 to 2010. Peers in Europe, Oceania and South America appear to have increasingly strong diurnal patterns. The diurnal patterns of North American peers remain consistently strong.

activities. Variations in the number of connected peers may hint at changes in the size of the overall BitTorrent population. An understanding of temporal patterns behind user activities, on the other hand, is necessary to understand the potential contributions of BitTorrent to congestion and near-peak transit charging rates.

We begin by plotting weekly timelines of the number of peers connected to each vantage point in 2008 and 2010 (Fig. 3). We group peers by location to highlight variations in usage over time and across regions. Focusing first on Europe, the largest contributor to connected peers, we see increasingly defined diurnal patterns with peak usage in the late evening, and relatively larger peaks and troughs in 2010 compared to 2008. This is surprising given the logical, common belief that BitTorrent is used out-of-phase with other applications because of the high load it imposes on users connections [17].

To better illustrate changes in diurnal patterns for the number of connected peers over time, we plot the average peak-to-trough ratio of the hourly connected peers by continent (Fig. 4). Larger ratio values indicate that greater portions of peers in each region use BitTorrent at the same time. The figure shows that the ratio of connected peers in North America has remained consistently high over the last two years, with 80% more peers online during peak usage. Meanwhile, diurnal patterns in Europe have grown more pronounced during the same period. While the exact causes for this behavior are beyond the scope of this paper, we speculate that variations in copyright law and enforcement (which can affect how long users leave BitTorrent running) across time and regions may contribute to this effect.

	Metric	**2009**	**2010**	**Δ**
A	Peer download rate	110.9	138.7	+25.0%
B	Unique peers per hour	276.6	248.0	-10.3%
C	Concurrent flows	32.7	28.9	-11.6%
D (A/C)	Per-flow download rate	3.39	4.80	+41.5%
E (B*C)	Total flows	9040	7170	-20.7%
F (D*E)	Total download rate	30700	34400	**+12.1%**

Table 3: Summary of calculations to determine overall BitTorrent traffic, based on product of number of flows and per-flow download rate. All download rates are in MB/hr. We find that total BitTorrent traffic has increased 12.1% from 2009 to 2010.

Figure 3 also indicates that the average number of connections to each vantage point per hour has decreased during the observation period by 10% (compare the "All" curves in Figs. 3a and 3b). This could be explained in part by a drop in the system popularity and/or shorter session times. While it is difficult to quantify the former, we can use our dataset to directly evaluate the latter. Figure 5 shows a typical distribution of session times, and also plots median session time for vantage points in each continent from November 2008 to 2010. We find that, from 2009 to 2010, median session times have decreased for each of the three main continents. Since 2008, session times in Europe, North America and Asia decreased by 13%, 23%, and 20%, respectively.

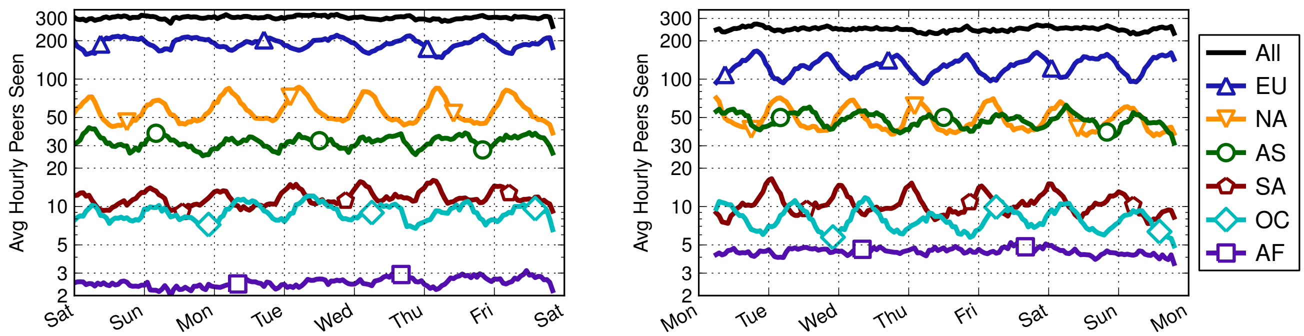
Finally, we note that changes in the average number of connected peers are dominated by a 30% drop in the number of European peers. This is aligned with previous reports that European users are increasingly using direct download sites in lieu of P2P [17]. In contrast, usage in Africa and Asia *increased* by 75% and 45% from 2008 levels, indicating a growing presence in developing regions.

These results reveal conflicting trends that make it difficult to estimate changes in aggregate BitTorrent traffic. With this in mind, we compute a measure of the total BitTorrent traffic as the product between the number of BitTorrent flows in the network and the average per-flow hourly traffic volume. For this, we determine the average number of concurrent connections (i.e. flows) maintained by each peer and use it to compute the average per-flow download rate and an estimate of the total flows in the system. Table 3 provides a summary of these metrics and calculations. We use the *Peer download rate (A)* and the number of *Concurrent flows (C)* to estimate the *Per-flow download rate (D)* and compute an estimate of the number of *Total flows (E)* as the product of the number of *Unique peers per hour (B)* and the number of *Concurrent flows (C)*. The *Total download rate (F)* is then computed as the product of the *Per-flow download rate (D)* and the number of *Total flows (E)*. The reduction in the number of flows per peer results in an increase in the per-flow download rate of 41.5%. Thus, while the total number of flows in the system has shrunk by 20.7%, the BitTorrent traffic has had a net increase of 12.1% between 2009 and 2010.

To summarize, we find that BitTorrent traffic volume is growing, and its traffic is increasingly generated from shorter sessions that tend to occur during peak hours. As we show in Sec. 6, these temporal trends have a significant impact on transit charges. The next sections build on the identified trends to determine which parts of the Internet are most affected by the corresponding traffic and the impact of this traffic in terms of costs and revenue.

5. WHERE BITTORRENT FLOWS

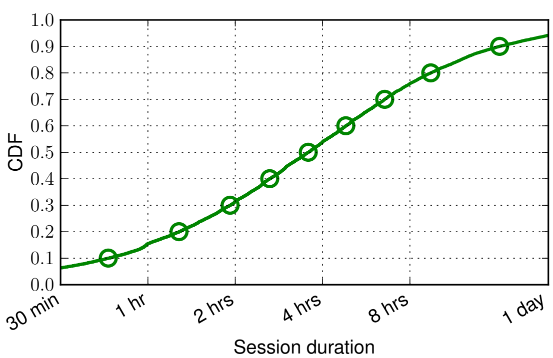
We now discuss *where* BitTorrent traffic flows through the network. We begin with a discussion of how to map traffic flows to the network paths they traverse. We use these mappings to study the



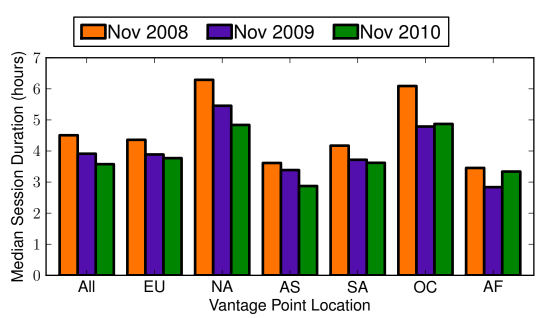
(a) November 2008

(b) November 2010

Figure 3: Distribution of connected peers per hour, grouped by continent, in 2008 and 2010 (semi-log scale). Vertical grid lines correspond with midnight, UTC. Note the increasingly defined diurnal patterns.



(a) All Session Times, Nov 2010



(b) Changes in Median Session Time

Figure 5: CDF of average session time per vantage point for all peers in Nov 2010 (top). Overall, session times have decreased from Nov 2008 to Nov 2010. For peers in Asia, Europe, and North America, median session times dropped by 13% to 23% over this time interval (bottom).

geographic and topological characteristics of the traffic exchanged between users from March through the end of May 2010.

5.1 Mapping BitTorrent Flows

In the following paragraphs, we address the problem of mapping BitTorrent flows to the paths they traverse. In particular, we show that publicly available path information such as BGP feeds are insufficient for mapping the vast majority of BitTorrent traffic, and address this by supplementing the public view with traceroutes collected between our vantage points and their connected peers. While the limitations of the public view of Internet topology are well known [3], we focus on what this implies for estimating P2P traffic locality and costs.

To infer traceroute-based AS path information, we combine over 202 M traceroutes between peers in our dataset with data gathered from public BGP feeds [26] using heuristics from Chen et al. [3]. Altogether, our dataset consists of 13.1 M distinct AS paths.

We then determine the portion of BitTorrent flows that, for the same time period, can be mapped to an AS path. First we map each flow's endpoint IP addresses to a source/destination AS pair [24]. For each of the resulting 2.1 million AS pairs, we determine whether an AS-level path exists, using either the paths in the BGP public view alone or using a combination of the public view with traceroute-derived AS paths. We say that such a path exists for a pair if both the source and destination of the pair appear in any path.

Figure 6 plots the cumulative distribution function of the portion of BitTorrent traffic per vantage point that can be mapped to an AS path using either BGP only paths (curve labeled "BGP") or the combined set of BGP and traceroute-derived paths ("BGP + Traceroute"). The figure shows that paths available in the BGP public view are not sufficient to account for the majority of flows in our traces – over 80% of vantage points cannot even map half of their traffic, while the median vantage point is able to map less than 14% of its traffic.

After adding traceroute-derived AS paths to this analysis, we can map nearly all BitTorrent traffic to AS paths. In particular, despite not having complete all-to-all traceroutes, for 90% of VPs we can map at least half of their traffic and are able to map over 96% of traffic for the majority (>50%) of peers.

These results show that when evaluating the Internet-wide impact of a globally distributed system, it is necessary and sufficient to supplement public views of Internet topology with topological information gathered from the edge of the network. The remainder of this paper uses this information to understand where BitTorrent flows and its impact on ISP costs and revenue.

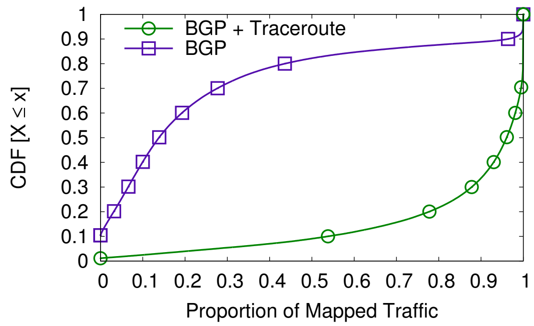


Figure 6: CDF of the portion of each vantage point's traffic that can be mapped to a path, using only BGP paths, or BGP and traceroute-derived paths. Paths in the public view cannot map most BitTorrent traffic, but adding traceroute paths results in nearly complete coverage.

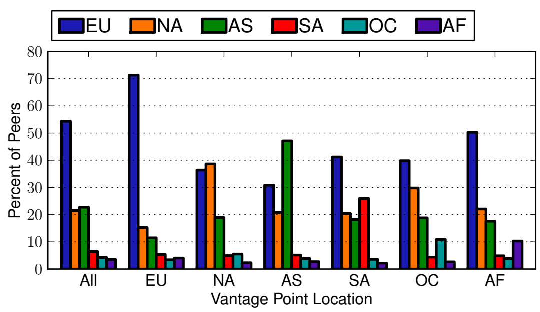


Figure 8: Distribution of the location of connected peers, according to the location of the vantage point, for November 2010. The VP's locale is always more strongly represented than in the "All" distribution.

5.2 Geographic Locality

While it is well known that BitTorrent is used in nearly every region and country worldwide, it is unclear how much of its traffic stays local. In this section, we show that traffic typically crosses few country boundaries, and the average distance it travels for a VP is strongly dependent on the VP location.

We first discuss the issue of locality of traffic. To represent this graphically, at the continent granularity, we determine the portion of each vantage point's traffic that flows to or from each continent. Figure 7 plots this as CDFs, where a point (x, y) for a given continent indicates that for a fraction y of the peers, the portion of their traffic flowing to endpoints in that continent is less than or equal to x. Curves closer to the lower right indicate continents receiving the largest share of peer traffic.

The figure includes these data for vantage points in each of the top three continents (by number of BitTorrent users). We observe that on average a VP exchanges more traffic with peers in the same continent than in any other. The effect is strongest in Europe (75% of traffic from European VPs stays within Europe), which contains the largest portion of BitTorrent users. Both North America and Asia exchange much larger portions of intracontinental traffic than their user populations would indicate.

Some of the reasons for the observed locality include content interest (e.g. based on language) as well as temporal trends – peers in a continent tend to use the system at the same time (as shown in Sec. 4). To test whether we find locality trends in traffic patterns, we plot the geographic distribution of connected peers, grouped by the continent for each VP (Fig. 8). The graph indeed shows that the distributions of connections per VP continent is similar to those for traffic.

To obtain a finer-grained view of how far aggregate BitTorrent traffic travels, we plot a CDF of each vantage point's traffic that passes through up to C countries (Figure 9). For 80% of vantage points, the majority of their traffic travels at most to one other country. Of the total traffic, we find that 32% stays within the same country, and an additional 41% travels to only one other country.

These results show that while BitTorrent's flows are geographically diverse, the location of a user (and the popularity of BitTorrent in that region) has a strong influence on the location of connected endpoints. In aggregate, BitTorrent traffic exhibits surprisingly high geographic locality – often traveling to at most one additional country. In the next section, we evaluate whether this locality holds when viewed in terms of the network topology.

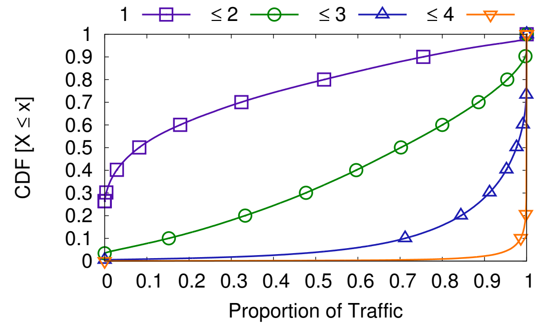


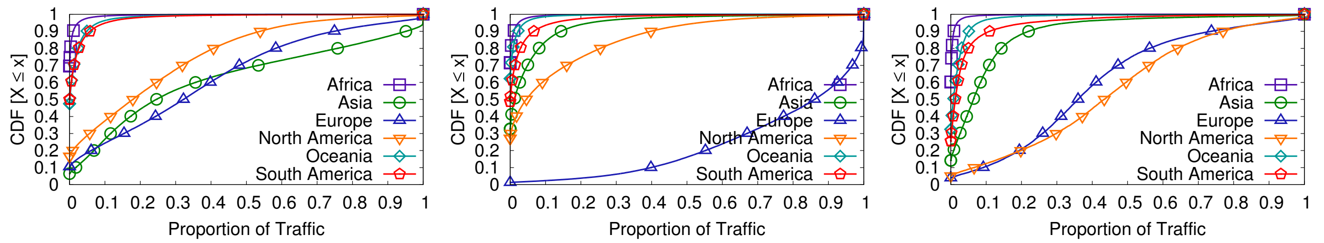
Figure 9: For each peer, the proportion of traffic that passes through up to C different countries. 73% of all traffic at most travels one country from its origin.

Tier	Category [7]	AS Count
1	–	10
2	Large Transit Providers	20
3	Small Transit Providers	2012
4	Content/Access/Hosting Providers Enterprise Customers	40993

Table 4: Description of each network tier, as well as the number of networks in each tier. We define Tier 1 to consist of ten well-known transit-free networks, a subset of ASes that are classified as "Large Transit Providers" by Dhamdhere and Dovrolis [7].

5.3 Topological Locality

We now examine the topological properties of BitTorrent traffic to determine which types of networks it traverses. We note that while these results may be affected by ISP-imposed throttles on interdomain traffic, our goal is to understand the impact of the system in its current environment. For this analysis, we map traffic to Internet tiers based on the classifications by Dhamdhere and Dovrolis [7], last updated in January 2010. This work classifies ASes into the tiers shown in Table 4, based on inferred business relationships. We apply this to categorize peers and routers in our dataset and find that, as one would expect, most BitTorrent users are located in lower network tiers (Table 5).



(a) Asia (b) Europe (c) North America

Figure 7: For each vantage point in a given continent, the proportion of its traffic according to the continent of destination. The curves show strong locality at the continent-level; this is particularly the case for Europe where most users are located.

	Tier-2	Tier-3	Tier-4
Vantage Points	13,838	181,981	143,368
VP ASes	17	611	2,524
Remote Peers	6,226,321	61,999,202	51,976,554
Remote ASes	18	1,363	14,562
Total ASes	18	1,364	14,573

Table 5: Distribution of vantage point and remote peer IPs and ASes, by tier. As expected, most of our VPs and remote peers are located in tier-3 and tier-4 networks.

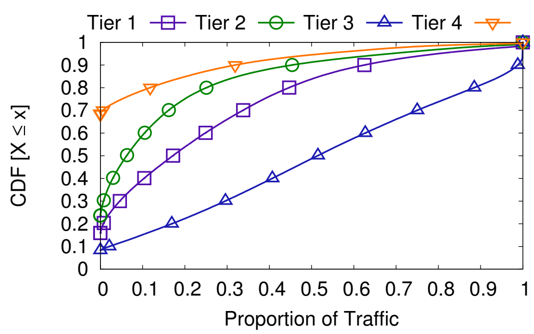


Figure 10: For each peer, the proportion of its traffic reaching Tier T. The vast majority of traffic only reaches tier 3, with significantly less traffic going to tier 1 or tier 2.

An interesting question related to the network impact of BitTorrent traffic is how deep into the "core" of the network it flows. We want to understand whether the traffic more frequently enters large transit providers or stays at lower tiers of the topology. To evaluate this, we determine the portion of each vantage point's traffic that reaches tier T. For example, if a flow traverses from tier 4 to tier 2 and back to tier 4, the flow is counted as reaching tier 2. Figure 10 plots the result as a CDF for each originating tier. Curves near the bottom right indicate tiers receiving the largest portion of traffic.

We find that tier-3 networks handle more BitTorrent traffic than any other tier. While over 50% of the median peer's traffic stays in tier 3, less than 10% (20%) goes up to tier 2 (tier 1).

To understand the role of the endpoint locations on the spread of BitTorrent traffic, we separate individual traffic flows by their starting and ending Internet tiers, and determine the portion of that traffic flowing to each of the other tiers. Figure 11 plots, for traffic between Tier T and Tier U, the proportion of that traffic reaching Tier V (such that $V \leq T$) as CDFs. As an example, Figs. 11a–11c show that, for traffic with at least one endpoint in tier 2, the majority of traffic stays in a tier-2 AS without passing through a tier-1 network.

Overall, the figures show that BitTorrent traffic most often stays in the same tier from which it originated. For instance, the trends for tiers 3 and 4 – where the vast majority of BitTorrent users are located – show that most traffic *does not go above tier-3* (Figs. 11d–11f). Further, for traffic between two tier-4 ASes, we see that the largest component of traffic unexpectedly *stays in tier 4*. When combined with results from geographic locality, this indicates that much of BitTorrent traffic remains in the same region and can be handed off among regional ISPs instead of using large transit providers.

This section showed that BitTorrent traffic exhibits strong locality, both geographically and in terms of network topology. In the next section, we evaluate the economic impact on ISPs as a result of these patterns.

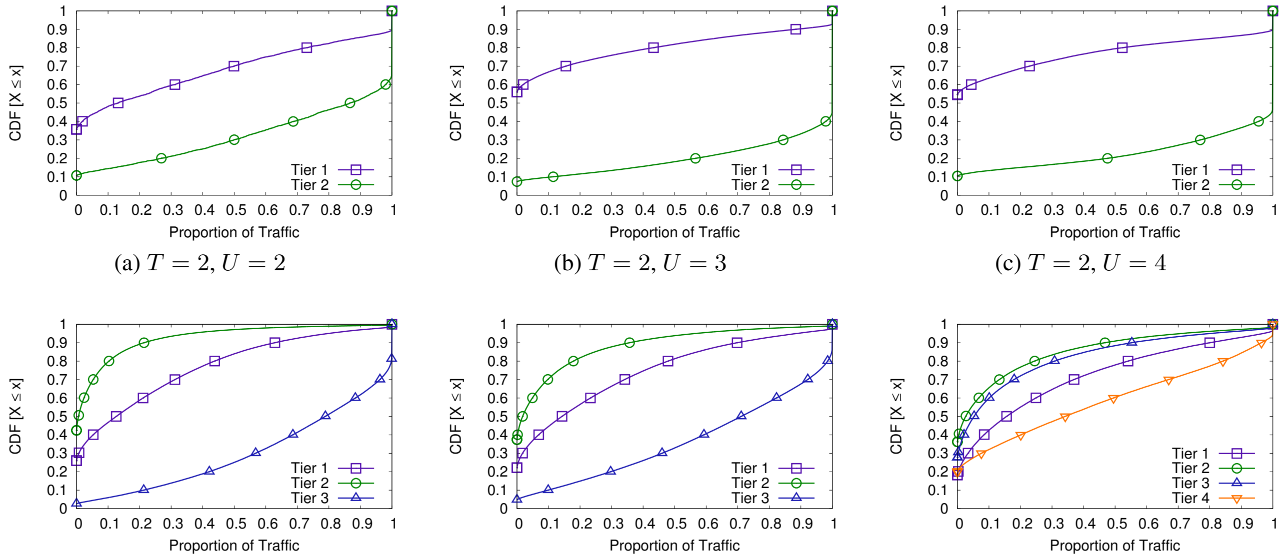
6. ECONOMIC ASPECTS OF NETWORK IMPACT

In this section, we address one of the key question driving P2P research and ISP policies: how does the network impact of P2P translate to costs and revenue for ISPs? The following paragraphs present a detailed analysis of the potential impact of BitTorrent traffic on the variable costs/revenues of ISPs. For our analysis, we use detailed traces of BitTorrent traffic, comprehensive AS topologies annotated with business relationships, and additional information on interdomain traffic volumes from a large ISP.

6.1 Overview

Interdomain traffic is an important component of ISPs' operational costs. A number of research efforts have focused on reducing interdomain traffic generated by P2P systems [1, 2, 5, 30]. Earlier studies have assumed that *all* interdomain traffic incurs charges. In practice, however, charges are a function of both the total traffic flowing over each interdomain link and the business relationships between ISPs.

Thus, in the first step of our analysis, we map BitTorrent flows to inter-AS links annotated with actual business relationships – customer-provider, provider-customer, peer or sibling. To this end, we use the algorithm proposed by Xia et al. [29], which leverages the valley-free and selective export policies of BGP routing to infer the relationship between connected ASes. This allows us to infer relationships for 98.3% of the 222,675 AS links in our dataset. We assume that transit charges occur only between customers and providers, resulting in costs for the customer and revenue for the



(a) $T = 2, U = 2$ (b) $T = 2, U = 3$ (c) $T = 2, U = 4$

(d) $T = 3, U = 3$ (e) $T = 3, U = 4$ (f) $T = 4, U = 4$

Figure 11: For each peer with data flowing between Tier T and U, the portion of that traffic that reaches up to Tier V.

provider. We further assume peering and sibling relationships to be settlement free, with no side paying the other to carry traffic [8].

In the next step of our analysis (Sec. 6.2), we model the net impact of those flows on ISP costs. A significant challenge is the diverse and commonly confidential charging models for transit agreements between ISPs. Absent this information, we first use a basic cost model that focuses on variable costs, assigning transit charges in proportion to the volume of traffic traversing a link. This allows us to understand trends in the balance between revenue- and cost-generating flows for the ASes in our dataset (Sec. 6.3).

Finally, we conduct case studies of the economic impact of BitTorrent on several ISPs using the 95th-percentile charging model (Sec. 6.4), in which the *temporal pattern* of traffic – not just the overall volume – plays a significant role in determining cost [23]. This burstable billing model is generally considered the most popular model used between small, access networks and their providers [9]. We have shown (Sec. 5) that a significant fraction of BitTorrent traffic is handled by small transit providers near the edge of the network and that this traffic is increasingly exhibiting strong diurnal usage patterns (Sec. 4). While we cannot assign dollar values to 95th-percentile traffic, we can determine *whether BitTorrent is relatively more expensive* than the rest of the traffic traversing each link.

6.2 Portion of Charging Traffic

In this section, we analyze BitTorrent traffic in terms of the types of links that it traverses. To begin, we find that 8% of all traffic in our dataset stayed in the same AS. Though this may seem to be a small number, one should consider that peers in our dataset are distributed across nearly 16,000 ASes. For these flows, we assume that there are no transit charges and we exclude them from the remainder of our analysis.

We focus then on interdomain traffic and compute the portion of each AS's total BitTorrent traffic that crosses links to its customers, to its providers, and to its peers. This allows us to understand the portion of BitTorrent traffic that traverses charging links and thus contributes to ISPs' costs.

We begin by describing summary results for tier-1 traffic (not shown). Not surprisingly, none of this traffic flows to a provider (by definition), but interestingly the tier-1 ASes experience significantly more peering traffic relative to customer traffic. The implication is that even when traversing tier-1 networks, BitTorrent flows are relatively unlikely to incur variable charges.

For traffic in tiers 2, 3 and 4, Fig. 12 plots a CDF of the proportion of per-AS interdomain traffic grouped by business relationship. In tier-2 networks (Fig. 12a), the vast majority of traffic crosses no-cost peering links, while a small portion of the traffic crosses charging links. In the median case, over 95% of tier-2 traffic crosses no-cost links. We also note that, on average, more of their non-peering traffic traverses customer links than provider links.

For tier-3 ASes (Fig. 12b), we again find that significantly more traffic crosses peering links than provider or customer links; 25% of these ASes send the majority of BitTorrent traffic to provider links. Unlike with tier 2, provider traffic is much larger than customer traffic for tier 3, indicating that these ISPs on average are paying for rather than profiting from transit charges due to BitTorrent traffic.

Last, we analyze traffic distributions for tier-4 networks (see Fig. 12c). As expected, only a small fraction of these ASes have any customer traffic, so BitTorrent does not generate substantial revenue here. We also see that most tier-4 networks are connected either over peering or provider links. For half of tier-4 networks, the majority of BitTorrent traffic is handled by provider links, suggesting that BitTorrent is incurring significant transit charges for these networks.

6.3 Traffic Ratios

While the previous graphs indicate the portion of traffic along links for different business relationships, they do not allow straightforward calculations of the relative amounts of customer and provider traffic for each AS (and thus which direction of charging traffic dominates). Figure 13 plots CDFs of these ratios for each tier, except for tier 1 where the denominator would be zero. Values greater than one indicate cases where an AS receives more

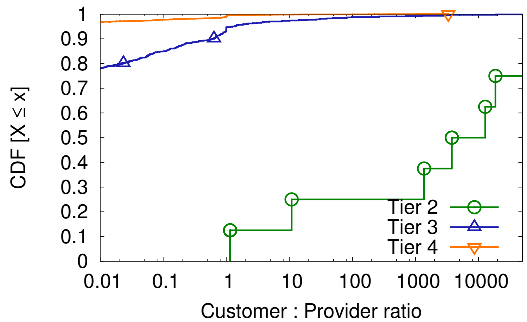


Figure 13: For each AS in Tier T, the ratio between traffic on customer and provider AS links. More customer traffic than provider traffic (a ratio >1) indicates a net revenue (if provider and customer traffic have the same cost). This shows the proportion of ASes in each tier that have a net revenue.

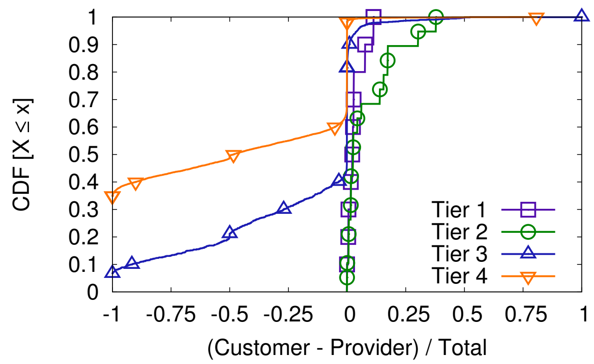


Figure 14: Average revenue per byte of BitTorrent traffic (i.e., the difference between customer traffic and provider traffic, divided by total traffic) for each AS, grouped by tier.

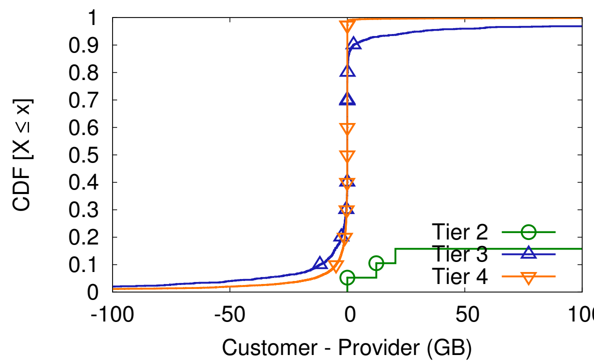


Figure 15: For each AS in Tier T, the difference between customer traffic and provider traffic. In contrast to the data in Figures 13 and 14, this perspective allows us to compare the scale of the revenue or expenses of ASes by tier.

customer traffic than it sends to providers (presumably generating a net revenue). Overall, this is always the case for tier-2 ASes. For lower tiers we find that a significant fraction of them do not have any customer traffic, resulting in a ratio of 0. The ratio is less than one most of the time, indicating that BitTorrent traffic is costing these networks – only 17% of tier-3 and 15% of tier-4 ASes with customer traffic have a ratio > 1. While this may seem to indicate that BitTorrent is harmful to lower-tier ASes, it is difficult to determine the relative cost of BitTorrent without understanding the volumes of non-BitTorrent traffic over the same links (an issue we address in Sec. 6.4).

The ratios above indicate when there is a net imbalance in charged traffic volumes but do not show their relative size compared to all BitTorrent traffic, including those flowing over no-cost links. We now address this by computing the average revenue (or cost) per byte for each AS (Fig. 14). This is defined as the balance of charging traffic (customer bytes minus provider bytes) divided by the total number of bytes flowing through the AS. When peering traffic accounts for a large proportion of AS traffic, the revenue of each byte of P2P traffic will be close to zero. However, when most AS traffic is from providers (or customers), it will have a more significant cost (or revenue) per byte for flows that travel through its network.

In Figure 14, all tier-1 ASes have a net revenue (values > 0) because, by definition, they do not have any provider links. In addition, all tier-2 ASes have a net revenue as well, reflecting the fact that the majority of their traffic is on peering, customer, or sibling links. The ASes in tiers 3 and 4 have incrementally larger average costs per byte overall, corresponding to the larger proportions of traffic traversing their provider links.

While most tier-4 ASes do not generate revenue from BitTorrent traffic, there are a few exceptions. This is explained by the fact that the tier classification algorithm is not strictly hierarchical, so a tier-4 AS can be a provider for another AS. In this case, large portions of traffic can traverse this revenue-generating link, resulting in a net profit per byte in the graph.

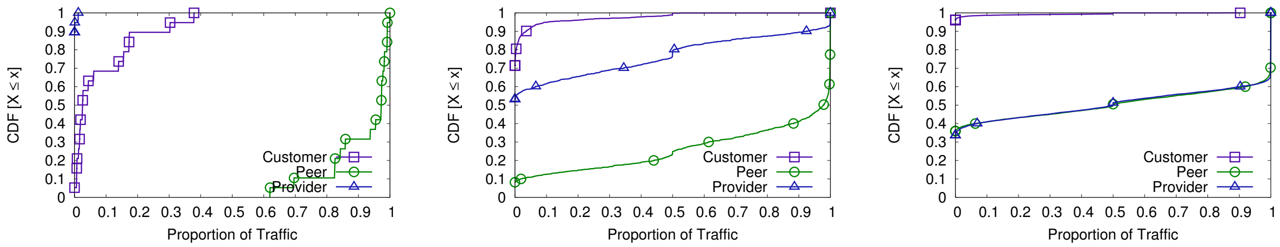
Finally, we attempt to quantify the relative scales of these costs and/or revenues by calculating a basic "balance sheet" for each AS in our study. In Fig. 15, we report customer minus provider traffic for each AS in tiers 2–4. Since tier-1 ASes do not have any providers, they have large net balances, several orders of magnitude larger than the net balances shown here. The balances of tier-2 networks range from 12 GB to 13 TB. By comparison, we see that the net differences from tier-3 and tier-4 ISPs are relatively small. Although tier-4 ASes had the largest average cost per byte of P2P traffic, we see that they have relatively small net balances of traffic compared to ASes in all other tiers.

6.4 Impact on 95th-Percentile Transit Costs

We now examine the cost of BitTorrent traffic under a 95th-percentile charging model, with a goal of understanding the impact of temporal trends in BitTorrent traffic on ISPs' costs. This is important because BitTorrent and network traffic are not uniform across time (e.g. due to diurnal trends), and costs computed under a 95th-percentile charging model are essentially set by usage in the busiest hours for network usage, typically in the evening. Intuitively, if BitTorrent traffic is more prevalent during these peak hours than off-peak hours, then we say it is *relatively more expensive* for the ISP, in comparison to the rest of the traffic. Appendix A provides a detailed description of the 95th-percentile charging model and the Shapley analysis that we use to determine the relative cost of BitTorrent traffic.

For this analysis, we obtain traces of total link volume between a major transit ISP "T" and several of its providers (A, B) and customers (C-G).[3] In addition, we compute for each pair of ASes that

[3]The identities of these ISPs are protected by nondisclosure agreements.



(a) Tier 2 (b) Tier 3 (c) Tier 4

Figure 12: For each AS located in Tier T, the CDF of the proportion of its traffic traversing provider, customer, or peering/sibling links.

the time series of BitTorrent traffic seen in our dataset. For both data sets, we use 5-minute intervals, a resolution commonly used in determining 95th-percentile charges. Both sets of traces are from the same 1-week period, January 6-14, 2011, and thereby are comparable and capture both weekday and weekend traffic patterns.

Following the typical use of 95th-percentile billing, we focus on the link direction with the greater 95th-percentile traffic volume. For ISP T's providers, this is the inbound direction (toward ISP T); for its customers, the dominant direction is outbound (from ISP T).

6.4.1 *"Relative Cost" Metric and Scaling*

For each type of traffic that we study – BitTorrent and "all the rest" – we compute that traffic's "relative cost". This is defined as the ratio between its Shapley value (how much it costs) and its overall fraction of traffic on the link. For example, if our BitTorrent trace accounts for 10% of all traffic seen over a link, but the Shapley value is 20% (e.g. if traffic occurs during peak hours), then BitTorrent traffic's "relative cost" is 2. Therefore, BitTorrent is contributing *more* to defining the 95th-percentile transit costs than the rest of the traffic. Using this metric, we can evaluate the *relative* contribution of any subset of network traffic over a link.

Though we have detailed BitTorrent traces for the networks we study in this section, it is important to note that there is no ground truth information to determine the relative volume of BitTorrent traffic compared to "all the rest" of traffic for the links that we study. To address this issue, we assume that our BitTorrent sample is representative of the larger population of all BitTorrent traffic on the network (following from our analysis in Sec. 3.2). This allows us to assess the relative cost of any BitTorrent traffic ratio by scaling our time-series of BitTorrent traffic to the corresponding fraction of overall link traffic.

Using the BitTorrent traffic ratio as a free variable, we scale each value in our time-series of BitTorrent traffic by a factor such that the sum of BitTorrent bandwidth matches a given percentage of the total aggregate link volume over the week. Then, we examine the impact of different fractions of BitTorrent traffic over a link on BitTorrent's role in setting the 95th-percentile costs.

6.4.2 *Relative Impact of BitTorrent on 95th-Percentile Costs*

We evaluate now the relative cost of BitTorrent traffic over several links between a large transit provider and several of its customers and providers.

First, we examine the trends in the relative cost of BitTorrent traffic as we vary the percentage of BitTorrent traffic ($X\%$) out of the total traffic on the link. To compute the relative cost for each X, we subtract the BitTorrent trace from the total trace to obtain the time-series of "all the rest" of the traffic and run the analysis described in Appendix A. A relative cost of 1 means that the Shapley value is the same as the fraction of traffic – BitTorrent traffic costs the same as other traffic, in terms of setting the ISP's 95th-percentile costs. Relative costs greater than 1 mean that BitTorrent is contributing *more* to setting the 95th-percentile costs than all the rest of the traffic crossing the link.

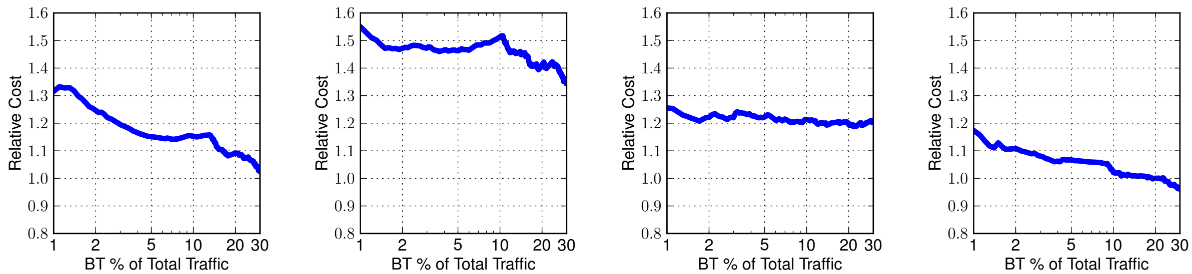
ISP	Cost	X-Corr	C.V.
Customer D	1.03	-0.4	109%
Provider A	1.15	-7.1	130%
Customer C	1.21	0.8	160%
Customer G	1.43	-0.2	186%
Provider B	1.50	3.2	188%
Customer E	1.52	1.6	158%
Customer F	1.83	7.4	325%

Table 6: For each link we study, we compute the cross-correlation offset ("X-Corr") that resulted in the best overlap between BitTorrent and total traffic, and the coefficient of variation ("C.V.") of the time series of BitTorrent traffic. We sort the links by increasing relative cost ("Cost", when BitTorrent is scaled to 10% of total traffic). Increased variation in the BitTorrent traffic curve is strongly correlated with increased relative cost of that traffic.

Figure 16 shows the results of this analysis for two providers and two customers of ISP T. Among these results, we find significant diversity in terms of the relative cost of BitTorrent, ranging from 0.95 (relatively less expensive) to over 1.5 (relatively more expensive). In general, BitTorrent tends to be relatively more expensive than the rest of traffic on the link. Note that as we increase the percent of BitTorrent traffic, the relative cost metric by definition approaches 1, which explains the downward trends in the figures.

To explain the diversity of the relative cost of BitTorrent traffic over different links, we characterize each of our BitTorrent traces by its variations over time, as well as how it aligns with the overall traffic on the link. To represent how much BitTorrent traffic varies over time for each link, we use the coefficient of variation (i.e., the normalized dispersion of values in a distribution). To capture the temporal alignment between BitTorrent traffic and total traffic, we conduct a cross-correlation analysis between normalized time-series of traffic data and report the time offset at which we found the peak overlap between BitTorrent and total traffic. Negative offsets occur if the peak in BitTorrent traffic appears *later* than the total traffic.

Table 6 shows the results of these analyses for all of ISP T's links that we study, sorted by increasing relative cost (when BitTorrent is scaled to 10% of total traffic). The relatively small values (e.g.,



(a) Provider A to ISP T (b) Provider B to ISP T (c) ISP T to Customer C (d) ISP T to Customer D

Figure 16: Trends in the relative cost of BitTorrent traffic for links between ISP T and several of its providers and customers, as we scale BitTorrent traffic to various percentages of the total link traffic.

≤ 2) in the third column indicate that the peak traffic for BitTorrent indeed coincides with peak volumes for the rest of traffic for most of these links. Moreover, we find that larger variations in BitTorrent traffic (i.e., burstiness captured by the coefficient of variation metric) correlate strongly with higher relative cost. Customer E is the only exception to this trend, which can be explained by the fact that Customer E's BitTorrent traffic is more closely aligned with total traffic (according to the cross correlation metric) than either Provider B or Customer F.

6.5 Summary

In summary, this section showed that capturing the impact of BitTorrent on network costs requires a global view of where traffic flows as well as an understanding of the business relationships and charging billing models used. We found that large portions of BitTorrent traffic flow over settlement-free links, and the portion of interdomain traffic generating charges varies significantly among ISPs. This analysis also shows that the recent trend toward more diurnal usage patterns leads to traffic that is closely aligned with peak usage for other Internet traffic, which has a significant impact on the common 95th-percentile charging model. While the relative cost of BitTorrent traffic is variable, we found it to be generally more expensive than other traffic for links we evaluated.

These results can inform ISPs' decisions about the impact and efficacy of specific traffic management policies. Our BitTorrent-focused analysis is an example of how applications can have varied impacts on ISP costs, based on when the traffic occurs.

7. CONCLUSION

In this paper, we demonstrated the importance of a comprehensive view when evaluating the network impact of a globally distributed system like BitTorrent. By incorporating application traces gathered from hundreds of thousands of IPs, we answered key questions regarding trends in BitTorrent's traffic volumes and user behavior, where BitTorrent traffic flows, and whether such traffic incurs net costs or produces revenue for ISPs.

This unique view allowed us to reveal several properties of the system previously hidden from studies based on limited perspectives. First, we used a longitudinal view of user behavior to show that the BitTorrent system continues to evolve, both in terms of when and where the application is being used. We then evaluated the impact of this behavior on where BitTorrent traffic flows geographically and topologically, an analysis that requires extensive traceroutes between P2P users. We found that despite its global reach, BitTorrent is able to remain local for large portions of its traffic. Further, our results show that most traffic generated by BitTorrent users stays at or below tier 3. Some of our results call into question the ability of in-network monitoring approaches to capture salient features of widely distributed systems, particularly when deployed in higher Internet tiers.

Last, we evaluated the economic impact of BitTorrent traffic on ISPs' variable costs. Using inferred business relationships between ISPs, we showed that most BitTorrent traffic flows over cost-free paths and that it generates substantial revenue potential for many higher tier ISPs. We also highlighted the importance of the temporal pattern behind the generated traffic under the common 95th-percentile charging model. By combining traces from operational interdomain links with our corresponding BitTorrent traces, we determined that the relative cost of BitTorrent is variable and tended to be higher than other traffic for several ISPs.

In summary, we emphasize that no single aspect of this study alone – application traces, topology information and in-network traces – were sufficient to develop a complete picture of a widely distributed system such as BitTorrent. Only in combination did these perspectives allow us to view the system as a whole and its impact on the network.

8. ACKNOWLEDGEMENTS

We would like to thank our shepherd, Augustin Chaintreau, and the anonymous reviewers for their detailed and helpful feedback. We are always grateful to Paul Gardner for his assistance with Vuze and the users of our software for their invaluable data. This work was supported in part by NSF Awards CNS 0644062, CNS 0917233 and CNS 0855253.

9. REFERENCES

[1] V. Aggarwal, A. Feldmann, and C. Scheideler. Can ISPs and P2P users cooperate for improved performance? *SIGCOMM Comput. Commun. Rev.*, 37(3):29–40, 2007.

[2] R. Bindal, P. Cao, W. Chan, J. Medved, G. Suwala, T. Bates, and A. Zhang. Improving traffic locality in BitTorrent via biased neighbor selection. In *Proc. of ICDCS*, 2006.

[3] K. Chen, D. Choffnes, R. Potharaju, Y. Chen, F. Bustamante, and Y. Zhao. Where the sidewalk ends: Extending the Internet AS graph using traceroutes from P2P users. In *Proc. of ACM CoNEXT*, 2009.

[4] D. Choffnes, F. Bustamante, and Z. Ge. Using the crowd to monitor the cloud: Network event detection from edge systems. In *Proc. of ACM SIGCOMM*, 2010.

[5] D. R. Choffnes and F. E. Bustamante. Taming the torrent: A practical approach to reducing cross-ISP traffic in peer-to-peer systems. In *Proc. of ACM SIGCOMM*, 2008.

[6] R. Cuevas, N. Laoutaris, X. Yang, G. Siganos, and P. Rodriguez. Deep diving into BitTorrent locality. In *Proc. of IEEE INFOCOM*, 2011.

[7] A. Dhamdhere and C. Dovrolis. Ten years in the evolution of the Internet ecosystem. In *Proc. of IMC*, 2008.

[8] A. Dhamdhere and C. Dovrolis. The Internet is Flat: Modeling the transition from a transit hierarchy to a peering mesh. In *Proc. of ACM CoNEXT*, 2010.

[9] X. Dimitropoulos, P. Hurley, A. Kind, and M. P. Stoecklin. On the 95-percentile billing method. In *Proc. of PAM*, 2009.

[10] EdgeScope – sharing the view from a distributed Internet telescope. `http://www.aqualab.cs.northwestern.edu/projects/EdgeScope.html`.

[11] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, analysis, and modeling of BitTorrent-like systems. In *Proc. of IMC*, 2005.

[12] M. Izal, G. Urvoy-Keller, E. Biersack, P. Felber, A. Hamra, and L. Garces-Erice. Dissecting BitTorrent: Five months in a torrent's lifetime. In *Proc. of PAM*, 2004.

[13] T. Karagiannis, P. Rodriguez, and K. Papagiannaki. Should Internet service providers fear peer-assisted content distribution? In *Proc. of IMC*, 2005.

[14] H. Kim, k. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee. Internet traffic classification demystified: myths, caveats, and the best practices. In *Proc. of ACM CoNEXT*, 2008.

[15] C. Labovitz, S. Iekel-Johnson, J. Oberheide, and F. Jahanian. Internet inter-domain traffic. In *Proc. of ACM SIGCOMM*, New Delhi, India, August 2010.

[16] S. Le Blond, A. Legout, and W. Dabbous. Pushing BitTorrent Locality to the Limit. *Computer Networks*, 55(3):541–557, February 2011.

[17] G. Maier, A. Feldmann, V. Paxson, and M. Allman. On dominant characteristics of residential broadband Internet traffic. In *Proc. of IMC*, 2009.

[18] M. Piatek, H. V. Madhyastha, J. P. John, A. Krishnamurthy, and T. Anderson. Pitfalls for ISP-friendly P2P design. In *Proc. of HotNets*, 2009.

[19] J. A. Pouwelse, P. Garbacki, D. H. J. Epema, and H. J. Sips. The BitTorrent P2P file-sharing system: Measurements and analysis. In *Proc. of IPTPS*, 2005.

[20] D. Qiu and R. Srikant. Modeling and performance analysis of BitTorrent-like peer-to-peer networks. In *Proc. of ACM SIGCOMM*, 2004.

[21] H. Schulze and K. Mochalski. ipoque: Internet study 2007, Nov. 2007. `http://www.ipoque.com/media/internet_studies`.

[22] H. Schulze and K. Mochalski. ipoque: Internet study 2008/2009, Nov. 2009. `http://www.ipoque.com/media/internet_studies`.

[23] R. Stanojevic, N. Laoutaris, and P. Rodriguez. On economic heavy hitters: Shapley value analysis of 95th-percentile pricing. In *Proc. of IMC*, 2010.

[24] Team Cymru. The Team Cymru IP to ASN lookup page. `http://www.cymru.com/BGP/asnlookup.html`.

[25] R. D. Torres, M. Y. Hajjat, S. G. Rao, M. Mellia, and M. M. Munafo. Inferring undesirable behavior from P2P traffic analysis. In *Proc. of ACM SIGMETRICS*, 2009.

[26] University of Oregon Route Views project. `http://www.routeviews.org/`.

[27] uTorrent still on top, BitComet's market share plummets. `http://torrentfreak.com/utorrent-still-on-top-bitcomets-market-share-plummets-090814/`.

[28] Vuze, Inc. Vuze. `http://www.vuze.com`.

[29] J. Xia and L. Gao. On the evaluation of AS relationship inferences. In *In Proc. of IEEE GLOBECOM*, 2004.

[30] H. Xie, R. Yang, A. Krishnamurthy, Y. Liu, and A. Silberschatz. P4P: Provider portal for P2P applications. In *Proc. of ACM SIGCOMM*, 2008.

[31] C. Zhang, P. Dhungel, D. Wu, and K. W. Ross. Unraveling the BitTorrent ecosystem. *IEEE Transaction on Parallel and Distributed Systems*, July 2011. To appear.

APPENDIX

A. 95TH-PERCENTILE AND SHAPLEY

95th-percentile billing is one of the most common models used by providers to charge for traffic over interdomain links [9]. Under this billing model, costs are determined by near-peak usage, calculated by the 95th-percentile value at fixed intervals (e.g. 5 minute bins) over each billing cycle (usually 1 month). As a result, the effective cost of each byte of traffic varies depending on the particular time of day – bytes sent during times of high usage are more expensive than off-peak bytes.

Following the approach in [23], we use the game-theoretic concept of Shapley value to compute the *average marginal cost contribution* of individual classes of traffic under the 95th-percentile billing model.[4] This is obtained by averaging the marginal increase in 95th-percentile cost over all possible "arrival orders" for the classes of traffic being examined.

In our study of "BT" and "Other" traffic, we have two arrival orders: [BT, Other] and [Other, BT]. Given a function v_{95th} that returns the 95th-percentile value of a series of bandwidth measurements, we compute the Shapley value of BitTorrent traffic by averaging the marginal contributions:

$$\begin{aligned} m_1 &= v_{95th}(BT) \\ m_2 &= v_{95th}(Other + BT) - v_{95th}(Other) \\ SV_{BT} &= (m_1 + m_2)/2 \end{aligned}$$

Since the Shapley value is *efficient* (i.e. the sum of the average marginal costs equals the total cost), we can compute the proportion of the total cost attributable to BitTorrent traffic:

$$p_{BT} = SV_{BT}/(SV_{BT} + SV_{Other})$$

Finally, we compare the proportion of total cost attributable to BitTorrent traffic relative to the overall fraction of BitTorrent traffic f_{BT} carried over the network:

$$\text{Relative Cost}_{BT} = p_{BT}/f_{BT}$$

When the relative cost of BitTorrent is > 1, that means that increases in BitTorrent traffic are temporally aligned with increases in the total traffic (typically during waking hours) and BitTorrent traffic is *comparatively more expensive* than other traffic. Likewise, a relative cost < 1 means that BitTorrent traffic is *comparatively less expensive* than other traffic, and tends to occur during off-peak (e.g,. overnight) periods.

[4] This analysis is general to any such cost model.