Technical report:

An Estimate of Infringing Use of the Internet

January 2011

Version 1.8
Envisional Ltd,
Betjeman House,
104 Hills Road,
Cambridge,
CB2 1LQ

Telephone: +44 1223 372 400
www.envisional.com
piracy.intelligence@envisional.com

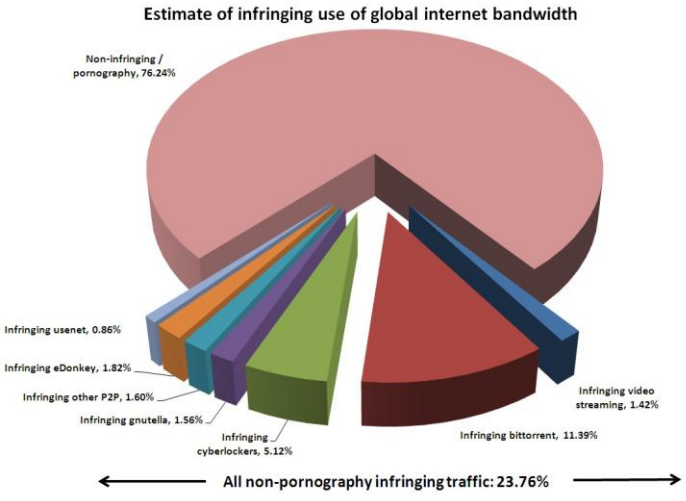
1 Introduction

Envisional was commissioned by NBC Universal to analyse bandwidth usage across the internet with the specific aim of assessing how much of that usage infringed upon copyright. This report provides the results of that analysis and is in three main parts.

- **Part A** examines the internet arenas most often used for online piracy – peer-to-peer networks (with a specific focus on bittorrent), cyberlockers (file hosting sites such as Rapidshare), and other web-based piracy venues (such as streaming video) – and estimates the proportion of infringing content found on each.
- **Part B is** a critical analysis of recent studies from four network equipment and monitoring companies. These companies measured network traffic at multiple (and different) sites worldwide to characterize overall internet usage.
- **Part C** combines the data and analysis from Part A and Part B in an attempt to show what proportion of internet traffic represents unauthorised distribution of copyrighted material.

1.1 Executive Summary

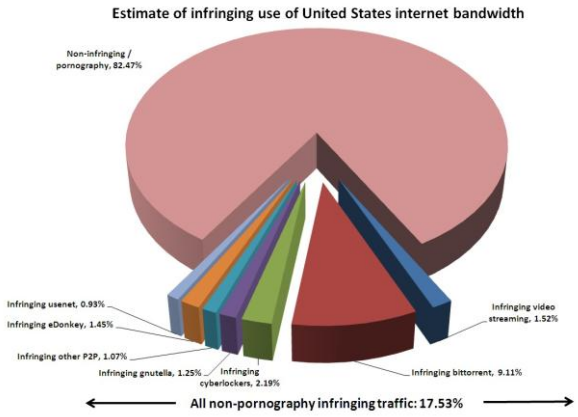
- Across all areas of the global internet, **23.76% of traffic was estimated to be infringing**. This excludes all pornography, the infringing status of which can be difficult to discern.
- The level of infringing traffic varied between internet venues and was highest in those areas of the internet commonly used for the distribution of pirated material.



- **BitTorrent traffic** is estimated to account for 17.9% of all internet traffic. Nearly two-thirds of this traffic is estimated to be non-pornographic copyrighted content shared illegitimately such as films, television episodes, music, and computer games and software (63.7% of all bittorrent traffic or 11.4% of all internet traffic).
- **Cyberlocker traffic** – downloads from sites such as MegaUpload, Rapidshare, or HotFile – is estimated to be 7% of all internet traffic. 73.2% of non-pornographic cyberlocker site traffic is copyrighted content being downloaded illegitimately (5.1% of all internet traffic).

piracy.intelligence@envisional.com

- **Video streaming traffic** is the fastest growing area of the internet and is currently believed to account for more than one quarter of all internet traffic. Analysis estimates that while the vast majority of video streaming is legitimate, 5.3% is copyrighted content and streamed illegitimately[1], 1.4% of all internet traffic.
- Other **peer to peer networks and file sharing arenas** were also estimated to contain a significant proportion of infringing content. An examination of eDonkey, Gnutella, Usenet and other similar venues for content distribution found that on average, 86.4% of content was infringing and non-pornographic, making up 5.8% of all internet traffic.
- In the **United States**, 17.53% of Internet traffic was estimated to be infringing. This excludes all pornography. A breakdown of internet usage yields the following results:
  - Peer to peer networks were **20.0% of all internet traffic** with bittorrent responsible for 14.3%. The transfer of infringing content located on these networks comprised 13.8% of all internet traffic.
  - **Video streaming made up between 27% and 30% of traffic**, though only a small percentage of this was believed to be infringing (1.52%)
  - **Cyberlocker traffic was estimated at 3%** of all network traffic and infringing use was estimated at 2.2% of all internet traffic.



Given the enormous, ever-growing, and constantly-changing size, shape, and consistency of the internet and the use that is made of it means that methodological issues abound when attempting to produce measurements of traffic and content. Yet even given the limitations of the data available, Envisional believes that the estimates produced in this report are more accurate than any that have been published before. This report draws together the data in a way that allows, for the first time, the organisations which can help shape the ways in which users interact and obtain content to understand how much of the internet is devoted to the distribution and consumption of infringing material.

Piracy Intelligence

Envisional Ltd

[1] Mostly from hosts commonly used for pirated content such as MegaVideo and Novamov rather than sites more often used for legitimate user generated content such as YouTube and DailyMotion, for instance.

piracy.intelligence@envisional.com

2 Part A: Internet Usage Assessment

2.1 Introduction

Part A of this report examines the major arenas of the internet known to be used – either primarily or as one of a number of uses – to distribute pirated content. Included in our analysis are:

- BitTorrent
- Cyberlockers
- Video streaming sites
- eDonkey and Gnutella
- Usenet

For each, we estimate the percentage of available content likely to be infringing. Then, in Part C, we translate these individual percentages into estimates of Internet traffic – to do this we rely upon data from studies into network traffic that were conducted by a range of vendors last year and which are discussed in detail in Part B. These individual estimates of infringing traffic are used to yield an estimate of the overall percentage of global internet traffic that results from their use (and which is infringing).

2.2 Executive Summary

Our major findings for each of the four major areas of our investigation follow.

BitTorrent

- BitTorrent is the most used file sharing protocol worldwide with over 8m simultaneous users and 100m regular users worldwide.
- Over 2.72m torrents managed by the largest bittorrent tracker were examined for this report. Our analysis suggests nearly two-thirds of all content shared on bittorrent is copyrighted and shared illegitimately. [2]
- An in-depth analysis of the most popular 10,000 pieces of content managed by PublicBT found:
  - **63.7% of content managed by PublicBT was non-pornographic content that was copyrighted and shared illegitimately**
  - 35.2% was **film** content – all of which was copyrighted and shared illegitimately

[2] PublicBT (publicbt.com) is the largest and most popular bittorrent "tracker" worldwide. A recent Envisional survey found that all of the most popular content listed on two popular portals referenced PublicBT trackers. With 2.72 million torrent files available in December 2010, PublicBT is believed to have comprehensive coverage of most files transferred using bittorrent and is therefore a suitable proxy for anyone seeking to assess the percentage of those transfers that infringe copyrights.

- 14.5% was **television** content – all of which was copyrighted and shared illegitimately. Of this, 1.5% of content was Japanese anime and 0.3% was sports content.
- 6.7% was **PC or console games** - all of which was copyrighted and shared illegitimately
- 2.9% was **music** content – all of which was copyrighted and shared illegitimately
- 4.2% was **software** – all of which was copyrighted and shared illegitimately[3]
- 0.2% was **book** (text or audio) or **comic** content – all of which was copyrighted and shared illegitimately
- 35.8% was **pornography**, the largest single category. The copyright status of this was more difficult to discern but the majority is believed to be copyrighted and most likely shared illegitimately[4]
- 0.48% (just 48 files out of 10,000) could not be identified

- Of all 10,000 files comprising the most popular content held on the PublicBT tracker, **only one was identified as non-copyrighted** (a file containing a list of IP addresses used to help users guard against spam and peer to peer monitoring). There is no evidence to support the idea that the transfer of non-copyrighted content such as Linux distributions makes up a significant amount of bittorrent traffic.[5]
- Analysis strongly indicates that private bittorrent sites (which would not usually make use of PublicBT) are overwhelmingly used for the purposes of illegitimately sharing copyrighted data.

eDonkey and Gnutella

- Analysis of known copyrighted and non-copyrighted material on the eDonkey network suggests that the vast majority of content held and transferred on the network is likely copyrighted (98.8%).
- Similar analysis using search queries on Gnutella found that most users on the network appeared to be looking for copyrighted content: 94.2% of non-pornographic search queries which could be identified were apparently for copyrighted material.

Cyberlockers

- An examination of 2,000 random links pointing to content held on cyberlockers found that 91.5% of links pointing to non-pornographic material were linking to copyrighted material, or 73.15% of all links.

[3] A very small proportion (0.13% of the top 10,000 or 13 individual files) was cracks aimed at removing the copy protection from copyrighted software such as Windows 7 or Microsoft Office.

[4] For the purposes of this report, the copyright status of any pornography identified is ignored, though the piracy of such content is obviously of interest to the adult video industry (reflected in the many legal suits filed against downloaders during 2010).

[5] Similar analysis conducted by Envisional in December 2009 found only a single Linux distribution as the only piece of non-copyrighted content in the top 10,000 torrents shared by OpenBitTorrent, then the largest bittorrent tracker online.

Video streaming sites

- A comparison of video streaming site usage estimated that 4.7% of video streaming data traffic is copyrighted content illegitimately streamed from video hosting sites.

Usenet

- Analysis of content posted to a number of Usenet newsgroups found that at least 93.4% of posts contained copyrighted material.

piracy.intelligence@envisional.com

2.3 Discussion: BitTorrent

All available data strongly suggests that bittorrent is the most used file sharing protocol worldwide. Part B of this report contains data conservatively estimating that bittorrent usage makes up 14.6% of *all* internet bandwidth worldwide. Envisional consistently measure over eight million users simultaneously connected to the bittorrent network and the distributor of two of the most-used bittorrent clients, uTorrent and BitTorrent Mainline, claims that the clients have over 100 million unique users worldwide and 20 million daily users[6].

This section of the report aims to establish what proportion of the data transferred through bittorrent is legitimate and approved by the content owner and what proportion is illegitimate and copyrighted. This is a complicated task. The estimate provided here is produced from a number of data points but primarily from a major investigation into the activities of the largest public bittorrent tracker, PublicBT.

2.3.1 Tracker Analysis

Much of the communication on bittorrent takes place with the aid of a central server called a *tracker*. A tracker helps users on bittorrent find those who are already downloading or uploading the file or files in which they are interested. The tracker records the IP addresses of those actively involved in obtaining or distributing a particular file and then shares them with other bittorrent users when requested.[7]

Trackers also record data on each **torrent or file** which they track: this data includes the 'hash' of that file (a unique code that identifies that file alone) as well as the number of **seeds** (users holding an entire copy of the file), **leechers** (users in the act of downloading), and (in most cases) total completed **downloads**. Trackers do not tend to record file names.

The largest tracker worldwide is the **PublicBT tracker**. At the point that this analysis was conducted, it held information on over 2.7m individual torrents[8]. Launched in 2009, the tracker became the most-used tracker for bittorrent swarms during 2010. PublicBT is simple to use, open to any bittorrent user, and free. It has also proved very reliable during its life to date. PublicBT does not cover *every* file available on bittorrent: bittorrent users are free to create torrents using any trackers of their choice and some niche content – such as sport broadcasts or technical ebooks – may be more often found at private trackers which require

[6] http://www.businesswire.com/news/home/20110103005337/en/BitTorrent-Grows-100-Million-Active-Monthly-Users

[7] Trackers are not the only way to obtain IP addresses: bittorrent clients can also communicate through a decentralised network overlay. Additionally, some clients will swap IP addresses of known downloaders or uploaders of a specific file in a transaction known as 'peer exchange', though they must have already managed to locate the other client in the first place. However, trackers are used as the first port of call in almost all torrent downloads and are likely to be the source of a significant proportion of the IP addresses gathered by a client.

[8] http://publicbt.com/

registration. However, analysis of the most popular 100 torrents on two popular portals (ThePirateBay, the most used portal worldwide and Torrentz[9]) found that every single torrent listed could be found on the PublicBT tracker, indicating that PublicBT can be assumed to have close to comprehensive coverage of the content that is most downloaded on bittorrent. The sheer size of the tracker also means that such coverage will be deep and broad.

Envisional was able to gather data on **every file tracked by PublicBT** on a specific day. This data was then used in an attempt to estimate the amount of legitimate against illegitimate and copyrighted content carried by the tracker. On the day of analysis (a weekday in mid-December 2010), PublicBT held information on **2.72m individual torrent swarms** and managed connections from just over **19.5m peers**.[10]

The analysis below examines the characteristics of all the 2.72m torrent swarms found on PublicBT. A detailed study was also made of the 10,000 torrents managed by PublicBT that had the most active downloaders, in order to better understand the make-up of the most sought-after content on bittorrent. An analysis of these swarms found that pornography, film, and television were the most popular content types. Further, with pornography excluded, **only one identified swarm in the top 10,000 offered legitimate content** (a file holding a list of IP addresses used to guard users against spam and peer to peer monitoring).

2.3.2 Summary analysis

On the day chosen for analysis of PublicBT , **2,721,440 torrents** were being managed by the tracker. These are unique files but the figure does not mean 2.72m different films or television episodes or pieces of music. There may be many different copies of a specific film title available through PublicBT – for instance, at different file sizes or in different formats or different qualities (as an example, seventy-one different versions of the film *Inception*, one of the most popular titles at the time of analysis, were located in the top 10,000 torrents).

Each file available on bittorrent is identified by a unique 'hash' – a unique code that identifies that file and no other.[11] PublicBT thus held information on the active downloaders and uploaders of just over 2.7m unique hashes.

---

[9] www.thepiratebay.org and www.torrentz.me

[10] This does not mean 19.5m individual users: a peer connected to two torrents will be counted twice in that total of peers due to the nature of bittorrent. It is not possible to know the average number of swarms to which an average user is connected at any one time. However, even assuming that each user is connected to nineteen torrents tracked by PublicBT (a very high estimate judging on anecdotal evidence) would still mean that 1m individual users were connected to PublicBT, around one-eighth of the total simultaneously connected bittorrent population of 8m. A more likely possibility is that most users connect to far fewer swarms and that PublicBT activity reflects a large proportion of public bittorrent transfers.

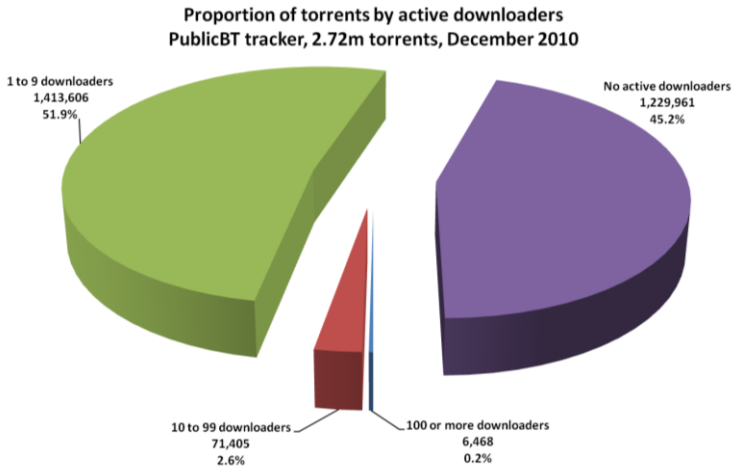
[11] A "hash" is a unique alpha-numeric sequence used to identify files (movies, music, documents, etc) on bittorrent. On the bittorrent network, the hash is generated by the SHA1 algorithm which creates a small identifier from a large file (such as a movie). Even trivial modifications to the original file results in a completely different hash.

Content analysis

On the day of analysis, most upload and download activity was concentrated amongst **a small number** of those 2.7m torrents with 34.9% of all peers involved in the top 10,000 (just 0.37% of all torrents). There was **an enormous long-tail of content** which had only a few or no seeds or a few or no leechers.

The chart shows the breakdown of all 2.72m swarms according to the number of downloaders (commonly called leechers) attached to each swarm[12]. Clearly, most of the swarms had only a small number of active downloaders or no active downloaders at all.

- 0.2% of torrents (6,468) had 100 or more downloaders
- 2.6% of torrents (71,405) had from ten to 99 downloaders
- 51.9% of torrents (1,413,606) had from one to nine downloaders
- 45.2% of torrents (1,229,961) had no active downloads



A similar spread was evident for seeders (users holding a complete copy of the file). For almost **half of all torrents** (1.32m or 48.5%), no seed was connected.

On the other hand, a very small overall proportion of content attracted large numbers of downloaders, representing a large proportion of all connected users. As stated above, torrent swarms with 100 or more downloaders represented just 0.24% of the available 2.72m torrents, but more than one in three – 30.4% - of all peers connected to PublicBT. Torrents with ten or more downloaders represented 2.6% of the 2.72m available torrents but over half – 53.9% - of all peers.

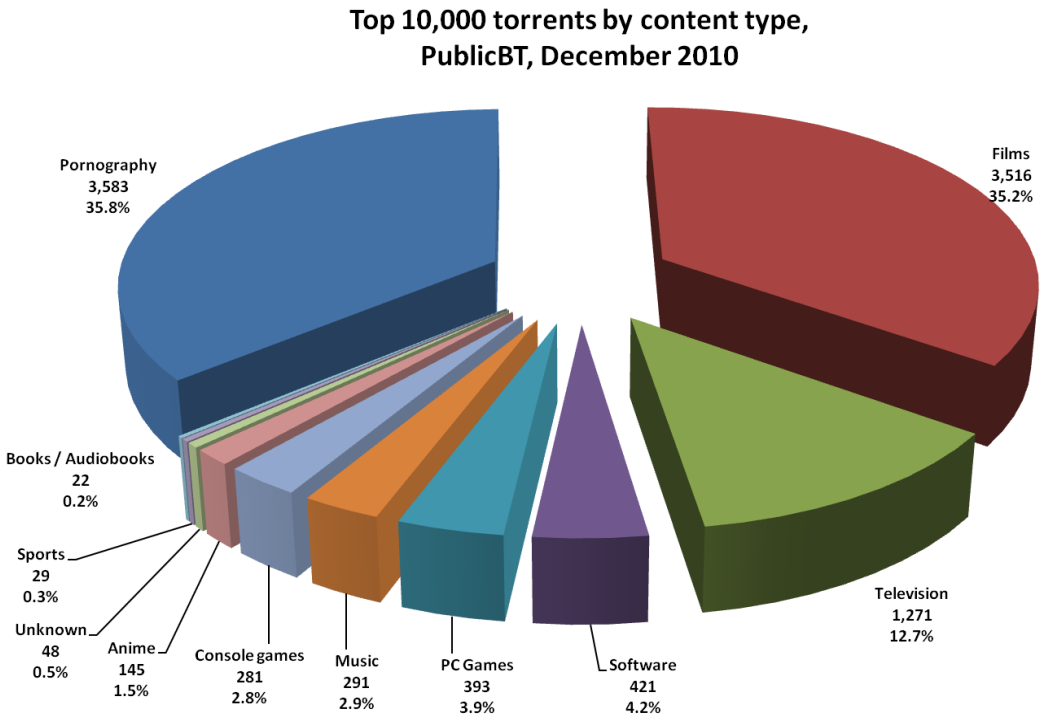
[12] This report uses the term 'swarm' even where no participants were actively sharing content (for instance, where there were no downloaders or no seeds). Technically perhaps, a torrent for which there is a tracker and a seed but no downloader should be known as a 'potential swarm' or similar but the term 'swarm' is retained for the sake of simplicity and understanding.

Analysis of the top 10,000 torrent swarms

To determine the percentage of infringing content associated with PublicBT, Envisional made a throrough analysis of **the top 10,000 swarms** (as determined by the number of downloaders). This is a small sample of the overall number of torrents (0.37%) but represents **34.9% of all peers** connected to PublicBT. To put it another way, more than one-third of all connections to PublicBT were interested in just 0.37% of the swarms managed by the tracker, showing a strong interest in a very small proportion of content. The seeds connected to these most popular 10,000 swarms were 35.5% of all seeds while the downloaders were 33.8% of all leechers.

The content being shared by each swarm in the top 10,000 was verified in almost every case using various methods[13]. Overall, **9,952 of the top 10,000 swarms were identified and confirmed** (99.52%) with only 48 swarms containing unknown content.[14]

The chart shows the distribution of swarms by content type with video dominating overall. Pornography video was the largest single type at 35.8% of all of the top 10,000 torrents. Film was the second largest type at 35.2%, followed by television episodes at 12.7%. Japanese anime episodes added a further 1.5% and sports broadcasts another 0.3%. These results mean that **85.5% of all of the top 10,000 torrents were video content of some kind**.

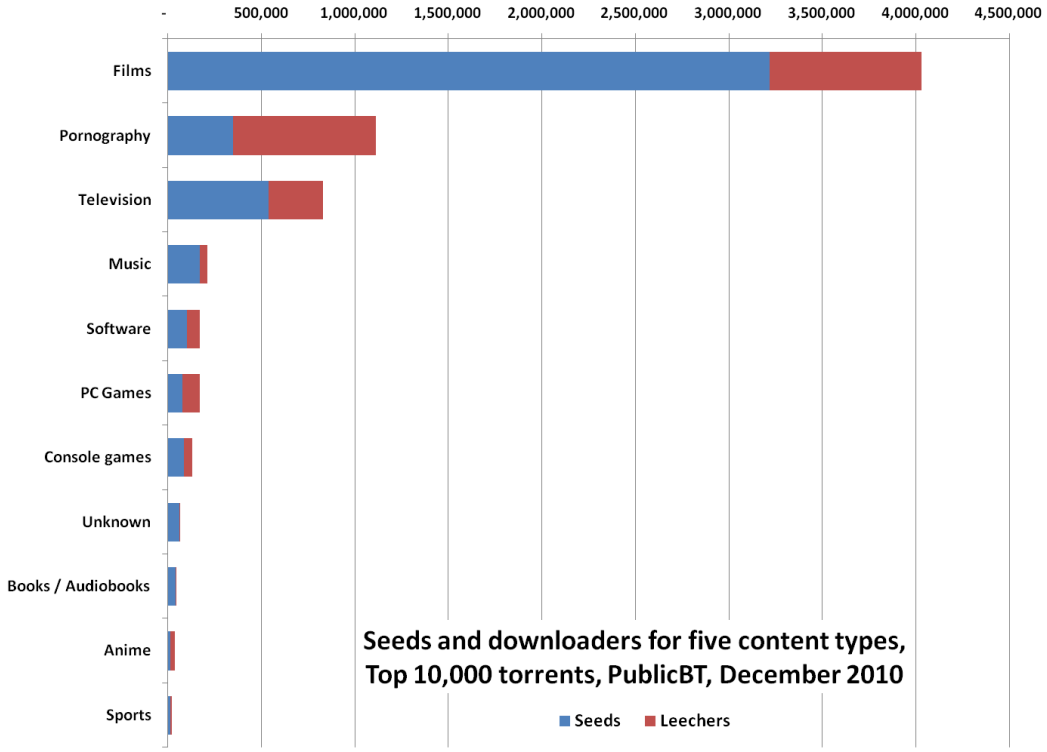


[13] In most cases, the hashes for each torrent were checked against a range of torrent portals for verification. For many video files, a section of the file was downloaded and viewed.

[14] Note that the analysis of the top 10,000 swarms contained here does not include 139 files which contained enough leechers to merit inclusion within the top 10,000 but were found to be fake. Fake files are often uploaded to bittorrent by interdiction companies hoping to confuse downloaders or by virus and malware distributors. The top 10,000 is therefore **the top 10,000 non-fake files** – or to put it another way, the top 10,139 files with the fake files removed.

Software comprised 4.2% of all of the top 10,000 torrents with computer games adding 6.7% (PC games were the largest proportion at 3.9% and console games contributed 2.8%). Music was 2.9% of the total with books (including comics) and audiobooks adding 0.2%. The remaining 0.5% of torrents could not be identified.[15]

The chart below looks at the number of seeds and downloaders for each content type within the top 10,000 torrents: again, video content – particularly film – gathered the largest number of seeds and downloaders (indicating strong demand and strong supply)[16]. In total, just over **4.0m peers were seeding or downloading a piece of film content** located in the top 10,000 torrent swarms on PublicBT at the point that this sample was taken. This is 59.2% of all peers connected to the top 10,000 swarms.



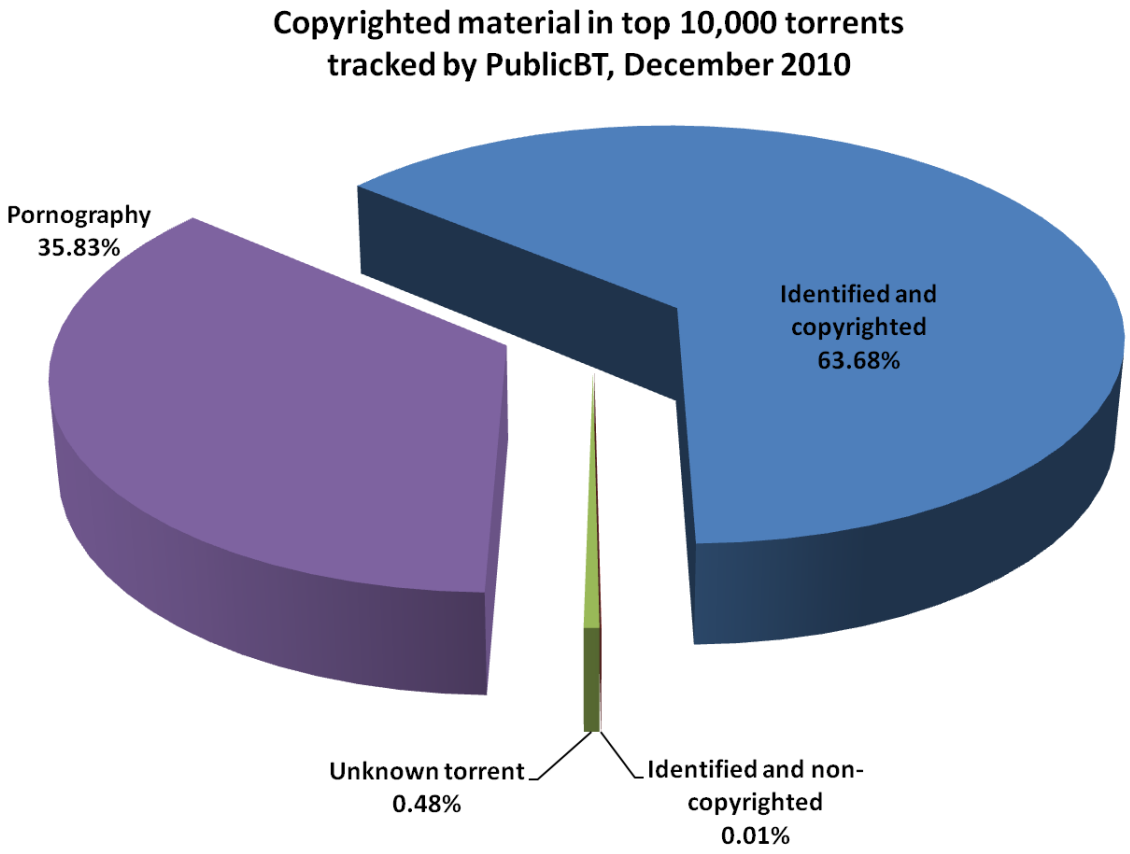
While pornography was the largest single type by numbers of torrents, there were many fewer total peers, principally because there were many fewer seeds than for film content. 828,000 peers were seeding or downloading television content and there were much lower numbers for the remaining content types in the top 10,000 torrents. Across all categories, peers connected to swarms for video content (films, television, anime, sports, and pornography) made up 88.4% of all peers in the swarms for the top 10,000 torrents.

[15] Overall, this analysis is similar to that conducted by Envisional in December 2009 on the OpenBitTorrent tracker, though the current effort successfully identified significantly more torrents. The earlier analysis could not identify 25.0% of the top 10,000 torrents though most of these unidentified torrents were believed to be pornography. The more recent analysis reported here suggests that this belief was correct.

[16] Numbers for seeders and downloaders were taken from PublicBT during the period of analysis.

Proportion of copyrighted material

As noted, the contents of 9,952 swarms were identified and verified. Excluding the swarms containing pornography (3,583 swarms or 35.83%) provides 6,369 pieces of verified content. Of these identified swarms, **only one was found to contain non-copyrighted content**. This was a torrent containing a list of IP addresses used to help peer to peer users block spam results and fake content.[17]



With the pornography content discarded, this means that at a minimum, **99.24% of the top 10,000 files managed by the PublicBT tracker** were copyrighted material with the rest of the content unknown (0.75%) or non-copyrighted (0.01%).

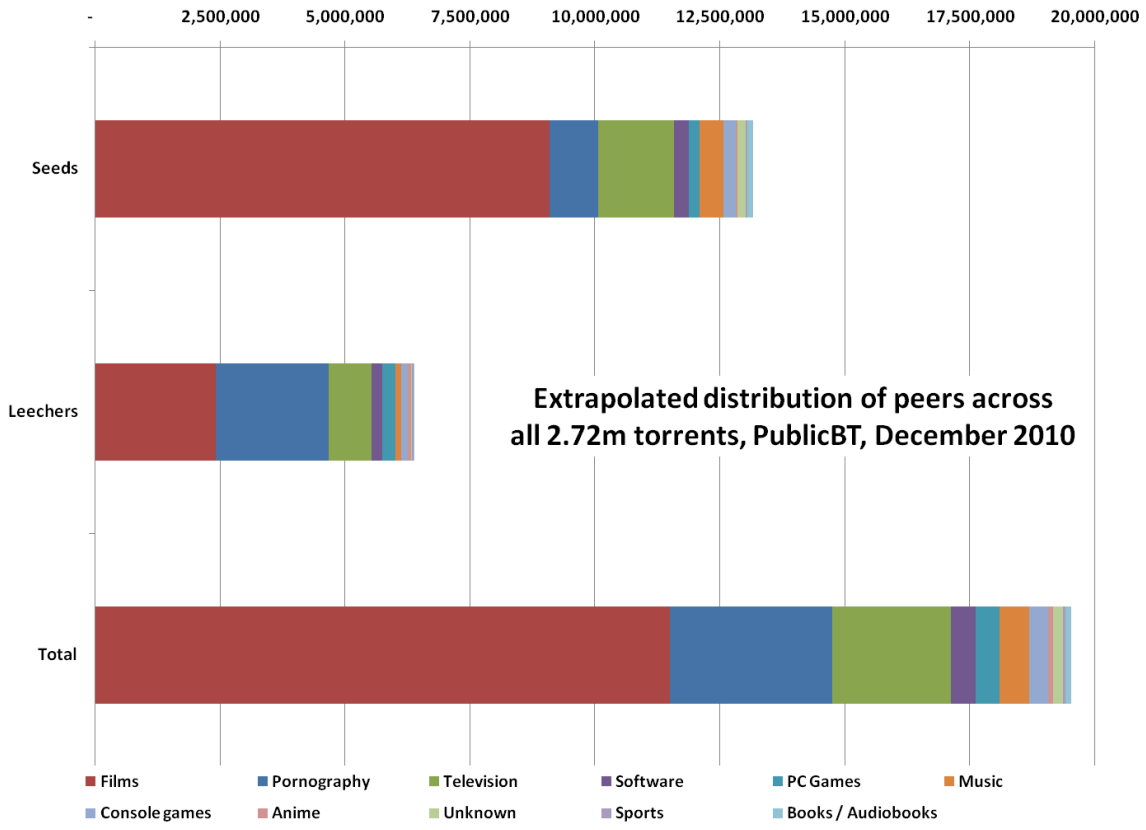
Analysis of content from outside the top 10,000 torrents found a similar dominance of copyrighted material. Five samples, each of 100 torrents, were taken from various points in the long tail of PublicBT content. Discarding

[17] The file was named "hostiles.txt". The torrent hash was a55603e3b98fb51fd05fb2ed3fbc2b2c6d254c6e. The results mirror the Illinois State University study conducted by Jon Peha and Alex Mateus (Carnegie Mellon University) in which it is noted: "…there is no evidence to support the hypothesis that the transfer of Linux distributions is a driver for the use of P2P, even among users that do not use P2P for copyrighted material." See *Dimensions of P2P and digital piracy in a university campus*: http://www.ece.cmu.edu/~peha/dimensions_of_piracy.pdf

pornography, no non-copyrighted content was located in these samples though there was a slightly higher spread of unknown material (as might be expected from less popular content).[18]

Extending the results

If the figures underlying the chart above for the top 10,000 torrents are extrapolated to all of the content present on PublicBT, it would mean that on the day of analysis, **11.5m peers were seeding or downloading film content** through the PublicBT tracker, **2.4m peers were seeding or downloading television content**, 3.2m pornography, 593,000 seeding or downloading music, and 862,000 games.[19] The chart shows the result of this calculation and the table over provides further details.



[18] This result accords with past analysis which have indicated that the majority of content offered on torrent portals is infringing. For instance, Judge Steven Wilson noted in his Isohunt decision that "In a study of the Isohunt website, [Dr. Richard] Waterman [of the University of Pennsylvania] found that approximately 90% of files available and 94% of dot-torrent files downloaded from the site are copyrighted or highly likely copyrighted."
http://www.wired.com/images_blogs/threatlevel/2009/12/fungruling.pdf

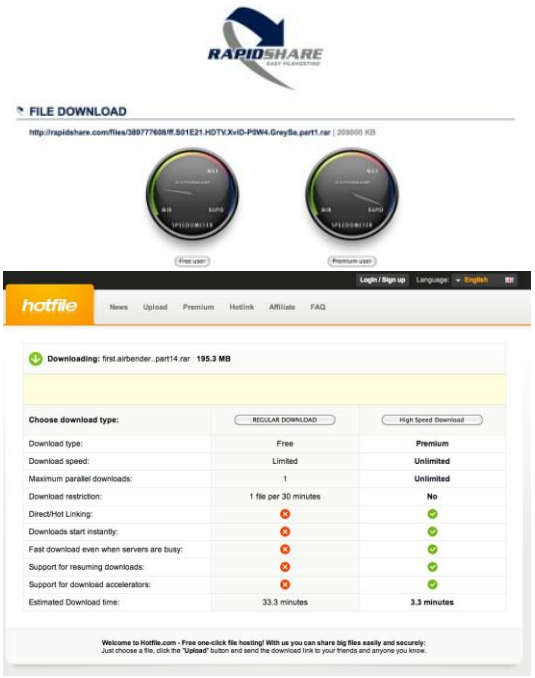
[19] For instance, 69.05% of all seeds for the top 10,000 swarms were involved in swarms for film content (3,220,293 seeds). Assuming that 69.05% of seeds across *all* swarms were involved in swarms for film content provides an extrapolated figure of 9,084,608 seeds.

	Seeds			Downloaders (leechers)			Total
Content type	**Seeds in top 10,000 swarms**	**Percent of all seeds in top 10,000**	**Estimated seeds across all swarms**	**Downloaders in top 10,000 swarms**	**Percent of all downloaders in top 10,000**	**Estimated downloaders across all swarms**	**Total peers (seeds plus downloaders)**
Films	3,220,293	69.05%	9,084,608	812,648	37.73%	2,404,271	**11,488,879**
Pornography	347,618	7.45%	980,648	766,157	35.57%	2,266,725	**3,247,372**
Television	538,607	11.55%	1,519,437	289,426	13.44%	856,285	**2,375,723**
Music	170,989	3.67%	482,369	37,399	1.74%	110,647	**593,016**
Software	99,645	2.14%	281,104	71,259	3.31%	210,824	**491,928**
PC Games	78,543	1.68%	221,574	91,059	4.23%	269,404	**490,978**
Console games	85,118	1.83%	240,122	44,148	2.05%	130,615	**370,737**
Unknown	58,687	1.26%	165,559	6,630	0.31%	19,615	**185,174**
Books (incl. audiobooks)	41,621	0.89%	117,415	2,777	0.13%	8,216	**125,631**
Anime	12,536	0.27%	35,365	24,211	1.12%	71,630	**106,994**
Sports	10,337	0.22%	29,161	8,046	0.37%	23,805	**52,966**

piracy.intelligence@envisional.com

2.4 Discussion: Cyberlockers / File hosting sites

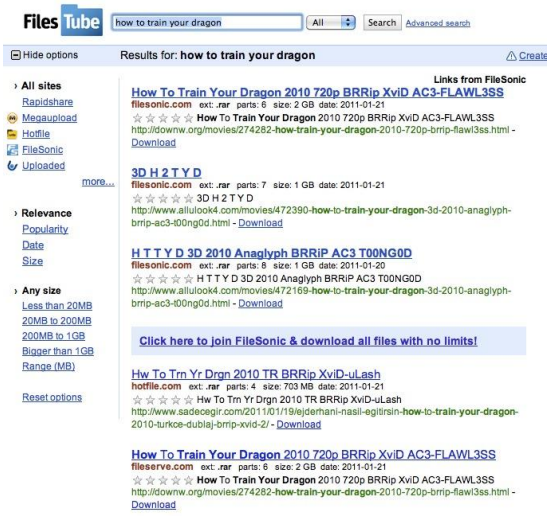
Over the last two years, various technological factors such as the decline in the cost of data storage combined with the increasing use of the web as the most important and central part of the internet for most users have led to the appearance and increasing use of what have become widely known as 'cyberlockers': centralised file storage services to which individuals can upload material for access by themselves or others. There are a number of widely used cyberlockers such as **MegaUpload**, **4Shared**, **Rapidshare**, and **Hotfile**. Envisional monitor over one hundred different cyberlockers.



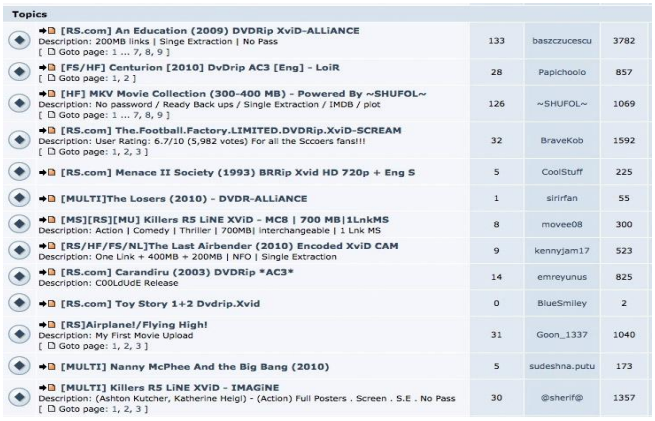
To store or access content on a cyberlocker, users need only a web browser – unlike P2P programs like bittorrent and eDonkey which require a dedicated client application. Also, direct downloading from a cyberlocker can be quicker than P2P on high bandwidth connections, more anonymous than P2P, and is often (at least at present) less prone to malware, viruses, and spoofing.

Users can freely upload any material to such sites and are then provided with a link with which anyone can then access that content. For non-paying users, content remains on the service for a limited period, can only be downloaded a certain number of times, and can only be downloaded after a waiting period of a minute or so while the potential downloader is presented with various advertisements. Premium memberships (typically costing around USD $13 / €10 a month) allow content to be stored for longer and – more importantly for downloaders – grant those prepared to pay with instant and high speed downloads of any content (not just their own) stored on the service.

Significantly, the vast majority of cyberlockers do not allow the content they hold to be searched in the same manner as a torrent portal: there is no way to query Rapidshare or MegaUpload for every file they hold that matches the phrase 'Lost' or 'Spiderman', for instance. This would seem to limit the attraction of these sites for piracy purposes but, as with many pieces of web-based technology, they were quickly co-opted for the purposes of containing and distributing pirated material. Hundreds of third-party **cyberlocker indexing sites** (such as FilesTube, right) and **link sites** (such as Warez-BB, shown in the screenshot below) have appeared in the last



couple of years which collate and make available links to pirated content held on cyberlockers. A user of such a site uploads a file to Rapidshare or another cyberlocker and then posts the link to that file on one of the many bulletin boards, forums, or indexing sites that cater to cyberlocker users. Any user can then click to obtain the material. As noted above, downloads are free, though users must sit through a wait time before the download can start and speeds are limited *unless* a premium account is purchased – this brings downloads that begin instantly at speeds which are usually as fast as the user's broadband capacity.



Screenshot from Warez-BB link site

The practice is not as large as bittorrent (and the need to pay for a premium account before the full benefits can be realised is one of the reasons why), though it has grown significantly over the last two years. The largest cyberlockers are among the most popular web sites in the world: for instance, ComScore estimates that **4Shared and MegaUpload have around 78m unique users** each month (more than twice as many as ThePirateBay, the largest bittorrent portal); RapidShare 60m unique users; and Hotfile 53m unique users. Alexa ranks 4Shared.com as the 66th most popular site in the world and MegaUpload as the 67th most popular. The usage studies in Part B estimate traffic to web-based cyberlockers and centralised file hosts at around 7% of all internet usage, though this varies significantly from country to country and may be as low as 2.5% for North America and the United States. Sandvine estimates overall usage of Rapidshare and MegaUpload together as 5.1% of all internet traffic.

Methodology

Envisional's Discovery Engine technology (an automated search, identification, and classification system for internet content) was employed to crawl the internet to locate links to content stored on ten large cyberlockers like Rapidshare and MegaUpload. The intention was to locate as many links as possible and then to analyse those

links to see what type of content had been uploaded to the cyberlocker (e.g., a film, television episode, ebook, photograph) and to determine whether that content was likely copyrighted or not.[20] A random sample[21] of 2,000 links gathered by the Discovery Engine was taken and analysed and the content type noted[22]. The results are below together with the proportion of each found to be copyrighted.

	Links found		Copyrighted	
Content type	*#*	*%*	*#*	*%*
Films	715	35.8%	709	99.2%
Television	169	8.5%	162	95.9%
Pornography	401	20.1%	345	86.0%
Music	201	10.1%	189	94.0%
Games	187	9.4%	155	82.9%
Software	199	10.0%	180	90.5%
Books / Audio books	52	2.6%	38	73.1%
Other / unknown	76	3.8%	30	39.5%
Total	**2,000**	**100.0%**	**1,808**	**90.4%**
Excluding pornography	**1,599**	**79.95%**	**1,463**	**91.5%**

As with bittorrent, much of the analysed content – over 90% – appeared to be copyrighted. The vast majority of films, television episodes, music, software, and games were copyrighted and available on cyberlockers illegitimately.

[20] An obvious shortcoming of this approach is the difficulty of finding links to non-copyrighted files legitimately stored on cyberlockers as such use does not generally involve publicizing a link onto the wider internet (personal photos, for instance, would likely be shared with family and friends via an email link). Still, it is reasonable to assume that while cyberlockers such as Rapidshare may host a non-trivial amount of non-copyrighted content, the *popularity* of that content – and hence the number of downloads and amount of bandwidth utilised – is likely limited.

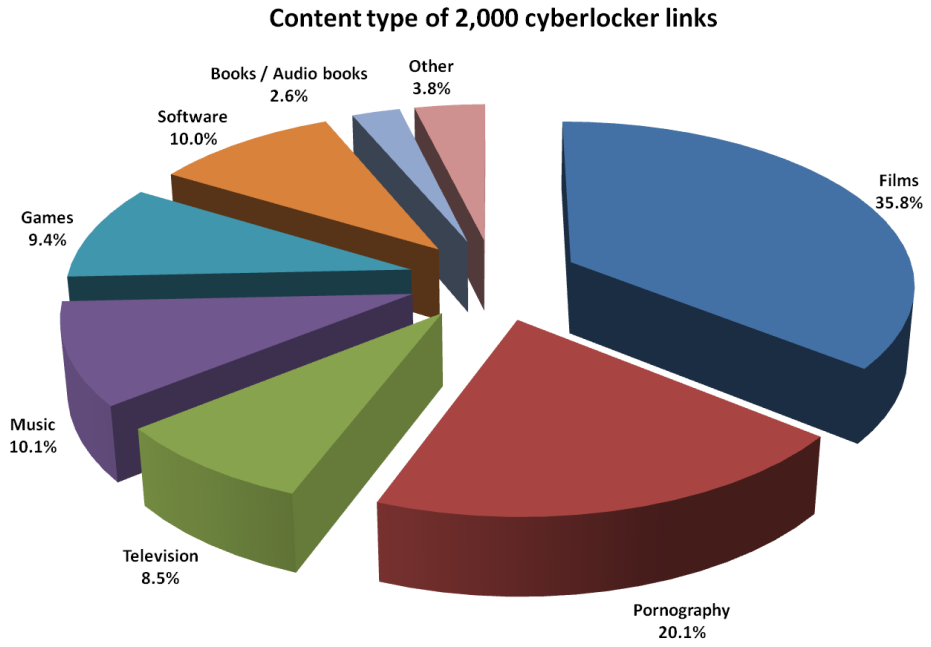
For example, Rapidshare announced a bandwidth upgrade to 600 Gbps (75 GBps) in March 2010 (http://en.wikipedia.org/wiki/RapidShare). This enabled a theoretical maximum of 194.4 PetaBytes/month to be transferred. Applying an 80% utilization factor results in an estimate of 155 PetaBytes of content transferred each month. With 50 million unique monthly users of Rapidshare (a figure taken from Google Trends), this amount of content equates to each user of the service downloading 4.15 movies per month. If films were replaced by collections of non-copyrighted photographs, those 50m unique users would need to download 307 collections of photos each month (assuming that each batch of photos comprised forty photos at 250Kb each = 10MB) were Rapidshare's bandwidth to be used entirely by this type of content.

The focus in this example is on downloading for, as Sandvine noted in its 2009 report: "Rapidshare is used primarily for data acquisition (*there is relatively little upstream traffic*) [emphasis added] and is generally not popular with average broadband subscribers." See: http://bit.ly/sandvine

The basic fact is that experienced internet analysts and researchers can find very little evidence that the bandwidth consumed by cyberlockers is used in the distribution of non-copyrighted content to any substantial extent.

[21] The sample was selected using a random number generator.

[22] Many cyberlockers only allow files of a particular size to be uploaded. This means that files greater than this size must be uploaded in parts. The common way to do this is to break the larger file into smaller 'Rar' files generated by the Rar archiving tool. The files will typically be named 'Filename.rar' and 'Filename.ra1' or 'Filename.part01.rar' and 'Filename.part02.rar'. When the Rar files are unarchived, the resulting file is re-created. For the purposes of this analysis, a file with multiple parts was treated as being a single file.

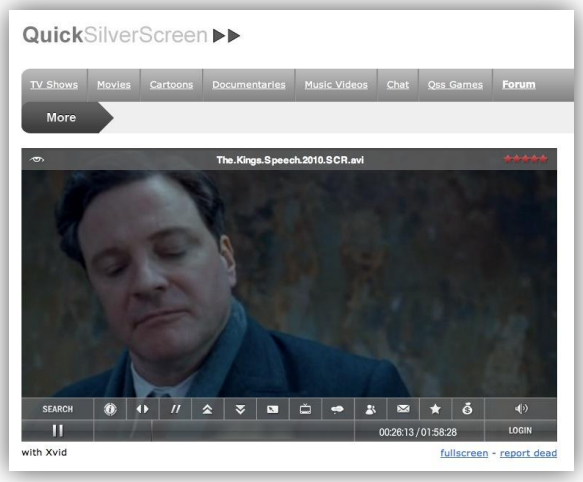


There is a larger proportion of smaller files such as eBooks and music on cyberlockers than on bittorrent. This accords with Envisional's experience of how each file sharing method is used. For example, with a cyberlocker, uploading is a simple one-click process that lasts only for the time necessary to upload the full file. There is no long-term uploading relationship and the upload occurs once at the decision of the uploader. Bittorrent, on the other hand, relies on a group of individuals exchanging small parts of a large file and the initial file creation process and upload process takes time and some knowledge. Seeding files is an ongoing process which can require long-term usage of a bittorrent client and an internet connection. Finally, files are uploaded only when and if another individual decides to download the file on offer – an element of uncertainty not present with cyberlockers. All in all, these differences provide cyberlockers with an ease-of-use advantage over P2P and users may respond by uploading a greater number of smaller files such as music and books.

2.5 Discussion: Video streaming

Every recent report which examines the recent past and immediate future of internet usage (see Part B) identifies streaming video as the fastest growing segment of bandwidth consumption worldwide. Led by YouTube, determined by most research to consume at least 5% of all internet bandwidth alone, the use of streamed video has become widespread across the entire internet. Sandvine believe that 'real-time entertainment' (streamed content consumed as it downloads) comprises 26.6% of all internet usage; Cisco state that 'streaming' traffic is 27.8%; and Arbor Networks estimate that 25% of traffic is streamed video or audio of some kind. All studies also cite the significant rise in this segment of internet usage and all predict further growth in this area.

Unlike bittorrent, eDonkey, and cyberlocker usage, experience indicates that most usage of video streaming is benign and poses no threat to copyright: Facebook videos of parties, news reports, YouTube rants, and so on. The rise in video streaming has gone hand-in-hand with the increase in user generated content pushed onto the internet and it is obvious to anyone with a passing familiarity with sites like YouTube that the majority of content currently uploaded onto such sites is produced by users and is not copyrighted or is uploaded legitimately by content owners (for instance, of the top ten 'most viewed' videos on YouTube, six are legitimately-uploaded music videos totalling 850m views).

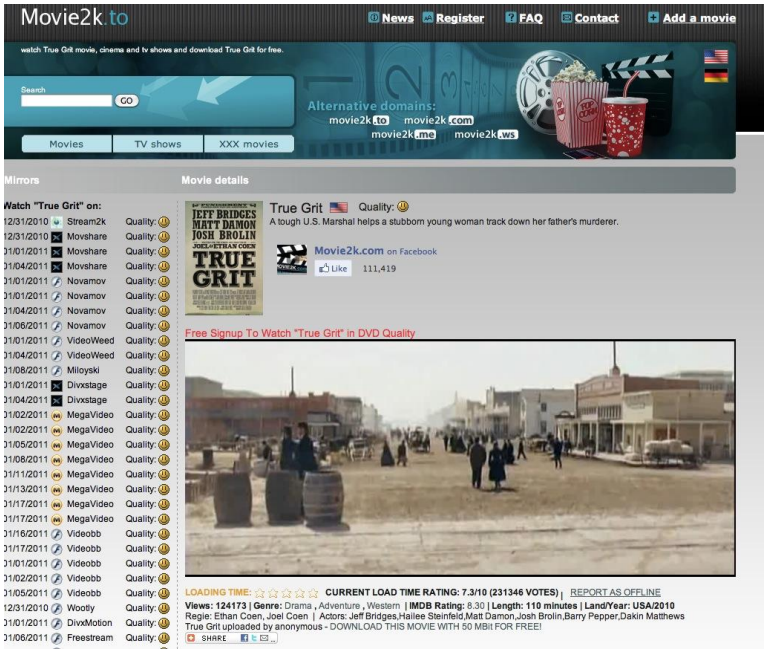


However, there can also be no question that there is a significant amount of pirated content available which has been uploaded to video hosting sites across the world. There is an obvious appeal to internet users of films and television episodes which begin seconds after a user clicks play rather than requiring a wait for the download to complete before consumption. Browser-based and easy-to-use, video streaming web sites are a major concern of content owners and it is not difficult to find pirated versions of any major film or television series with a few minutes of persistence.

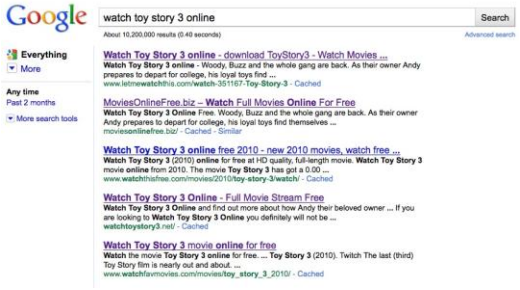
YouTube itself prevents most users from uploading content longer than fifteen minutes in length and has added tools such as digital fingerprinting to ensure that copyrighted material is identified and banned but the site has been host to a broad section of unauthorised copyrighted material in the past. Other video hosts are often much

less willing to implement proactive barriers to pirated content, allowing longer-duration uploads while enabling high quality streaming and refusing to implement filtering for copyrighted material.

In a similar fashion to the way that cyberlocker link sites have co-opted cyberlockers for piracy purposes, so video link sites have done the same for video hosts. Sites such as **LetMeWatchThis** and **Movie2k** index pirated content held on video hosts to present users with numerous choices for the latest film or television show. For instance, LetMeWatchThis currently offers forty-three separate working links to view *Inception* on different video hosting sites. Video link sites either embed Flash-based video players which stream content hosted on sites like MegaVideo or directly link viewers to the hosts that contain the streaming video.



Streaming videos of pirated content can also be found using a normal search engine. For example, querying Google for terms such as '*watch toy story 3 online*' reveals a plethora of linking sites and blogs in the top ten results which offer links to streams of unauthorised pirated versions of the film.



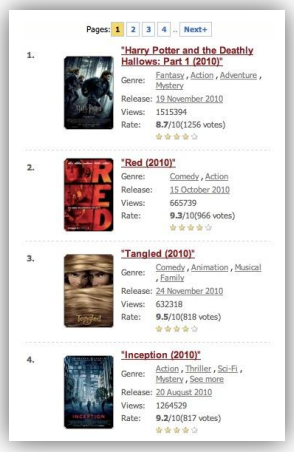
The most popular piracy video link sites gather millions of visitors each month. ComScore estimate LetMeWatchThis to have 6.5m unique users each month and Movie2K to have 5.0m unique users, for example.

Estimating pirated usage of video streaming

Estimating the amount of total video streaming bandwidth that may be unauthorised copyrighted material is difficult. Unlike bittorrent, where the PublicBT tracker manages millions of separate swarms, there is no major repository of video which can be taken to provide a good overall indicator of total video use: YouTube is certainly dominant in this space but as mentioned, there are a number of factors which ensure that YouTube is currently minimally used for new pirated content. The widespread nature of video use across the web means that a link analysis as performed for cyberlockers would be unlikely to gather accurate data.

After reviewing a number of possible methodologies, the best approach to this difficult area was deemed to be one which compared the popularity of index sites used to locate streaming pirated content with index sites used to locate pirated material available via bittorrent.

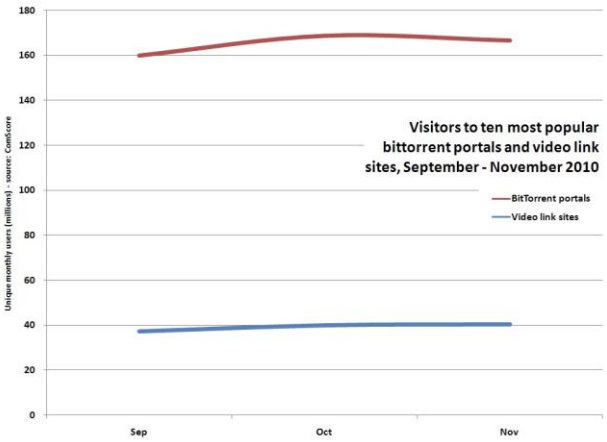
Web metric providers such as ComScore and Alexa offer statistics on the number of daily or monthly visitors to bittorrent portals such as ThePirateBay, IsoHunt, and Torrentz, the main sites from which the vast majority of bittorrent users find links to the pirated content that they ultimately download using the bittorrent protocol – and which then results in the large amount of bittorrent traffic seen in the usage studies. In the same way, users of video streaming sites use portals such as LetMeWatchThis, ZMovie (right) and Movie2K to locate links to pirated content they wish to see, clicking through to the video hosts where the content is hosted. By comparing the known audience for bittorrent portals with the known audience for video link sites, a rough estimate of pirated usage may be possible.



ZMovie

Both types of sites – bittorrent portals and video streaming link sites – are almost entirely devoted to pirated content: scans of the content available on bittorrent sites like ThePirateBay and IsoHunt and video link sites such as LetMeWatchThis and TVShack find close to no content which is not copyrighted (and that this content is unpopular when and if it does exist). It can then be broadly assumed that visitors to video streaming link sites will be consuming pirated material.

The chart shows data from ComScore for monthly unique users to the top ten bittorrent portals and the top ten video link sites worldwide from September to November 2010. Clearly, bittorrent is a much more popular activity on this measure: on average across these three months, the top ten video link sites had an audience just under one-quarter (23.71%) that of the top bittorrent portals – or to put it another way, the



bittorrent portals had slightly over four times as many visitors (4.22x).

Assuming that the end result of a visit to a bittorrent portal is the same as a visit to a video streaming link portal – that a user locates and downloads or streams the content in which they are interested – then the total data which is then transferred must be considered. The amount of data required to consume a file via a video streaming site is usually significantly less than when downloading a film or television episode from bittorrent. The file size is usually much smaller (and hence the final quality of what the user views is often poorer – which may be one reason why bittorrent is more popular as it provides higher quality content).

For example, each link for the ten most recent films posted to a popular video linking site was analysed and the streaming file to which it pointed on a video host was measured in terms of file size. On average, the streamed content comprised 384.2MB. Data taken from the analysis of PublicBT earlier in this report found that the average file size for downloaded films was 937.7MB. On this estimate, it means that each film downloaded via bittorrent results in almost 2.5 times (2.44x) as much data for the same content as via video streaming (or, stated another way, consuming a film via video streaming results in less than half the network traffic (40.97%) as downloading it via bittorrent).

$$\frac{\text{Visitors to Video Link Sites}}{\text{Visitors to Bittorrent Portals}} \times \frac{\text{Streaming File Size}}{\text{Bittorrent File Size}} = \text{Ratio of streaming traffic to bittorrent traffic}$$

As such, video link site traffic may generate the amount of data equivalent to **9.71% of all bittorrent traffic** (video link site visitors as a proportion of bittorrent portal visitors divided by the difference in average file size consumed). The detailed calculation is shown below which, assuming that Sandvine's estimate of bittorrent traffic is correct (14.56%), finds that the traffic which comes from video link sites that link to pirated material is equivalent to **1.42% of all internet traffic**.

piracy.intelligence@envisional.com

A. Amount of all internet traffic measured as bittorrent (Sandvine) [23]	**14.56%**
B. Amount of all internet traffic measured as video streaming of any kind (average estimate from Sandvine, Arbor, and Cisco – see Part B of this report)	**26.5%**
C. Video link site visitors as a percentage of bittorrent portal visitors	**23.71%**
D. Average streamed file size from video link sites (384.2MB) as a percentage of average film file size downloaded via bittorrent (937.7MB)	**40.97%**
E. Estimated pirated data usage of video link sites as a percentage of all bittorrent internet traffic **(C * D)**	**9.71%**
F. Estimated pirated data usage of video link sites as a percentage of *all* internet traffic **(A * E)**	**1.42%**
G. Estimated pirated data as a percentage of all streaming traffic **(F / B)**	**5.34%**

Given the difficulty of gathering data in this area, these figures should be taken as a cautious estimate.

[23] Sandvine estimates bittorrent traffic to be 14.56% of total internet usage and is the only company to provide a figure specifically for bittorrent based on a large amount of data – Ipoque did estimate bittorrent usage but its estimate is based on a small amount of total data from a low number of monitoring sites. Other companies talk of "peer-to-peer" usage and not "bittorrent usage".

Also, Sandvine measured peer-to-peer usage as a lower proportion of all internet usage than some other providers (particularly Cisco) leaving open the possibility that bittorrent usage may be higher. As Sandvine are the only company to provide data for bittorrent alone, their estimate will be used but should likely be taken as a minimum.

piracy.intelligence@envisional.com

2.6 Discussion: Other file sharing arenas

Analysis was also made of three other file-sharing arenas where copyrighted content is generally distributed: eDonkey, Gnutella, and Usenet.

2.6.1 eDonkey

The eDonkey peer to peer network is one of the oldest peer-to-peer networks still in existence. It is heavily used in mainland Europe (particularly in Spain, Italy, and France). Envisional measure between 2.5m to 3m users simultaneously connected to the network or a decentralised network overlay for the network called Kad. Sandvine estimates eDonkey traffic at 1.5% of all internet usage globally.

The most accurate way to calculate the proportion of pirated material available on eDonkey would be through analysis of one or more eDonkey servers and the content which is indexed and downloaded. However, such servers are high priority targets for anti piracy organisations and would be unlikely to cooperate with a request for oversight of the content which they have indexed. While it is possible for anyone to establish a server, doing so helps facilitate the distribution of content between users connected to that server and with much content felt to be pirated, this was not deemed to be a suitable way to research this area.

Instead, searches were made using the eMule client and Envisional's own peer-to-peer monitoring technology for one hundred pieces of content for which results would likely be pirated (new films and television episodes, for instance) and one hundred pieces of content for which results would not be pirated (content legitimately allowed to be distributed such as live concerts from some artists and books licensed under Creative Commons).[24] In each case, the most popular instances of each content type were chosen. The number of complete sources for each piece of named content were counted.

The amount of legitimate content available amounted to **1.2%** of all the content located on the network. This is a tiny proportion and while the research is not methodologically perfect, it does indicate that the majority of material held and transferred on eDonkey (in this analysis, **98.8%**)[25] is likely copyrighted.

[24] For example, copyrighted film content such as *The Dark Knight* and *Avatar* and television episodes from series such as *Lost*, *Heroes*, and *Doctor Who* and non-copyrighted material such as live concerts from Pearl Jam, books licensed under Creative Commons such as Cory Doctorow's *Makers*, and films like *Steal This Film*.

[25] Though this figure excludes pornographic content for which searches were not made.

2.6.2 Gnutella

The Gnutella network is widely used for the distribution of music as well as other content. Envisional's own Gnutella crawler estimates the network to have around 2.0m users at any one time since the closure of the company behind the LimeWire client at the end of 2010. Sandvine estimates Gnutella usage at 1.9% globally and the network is particularly popular in North America.

Envisional analysed the searches made by users on the network[26]. A sample of 3,500 search queries were examined for the content type to which they most likely referred and as to whether the content sought was copyrighted or not[27]. The table below shows the results. The 'copyrighted' column only includes those queries for which the copyright status could be clarified.

	Search queries		Copyrighted	
Content type	**#**	**%**	**#**	**%**
Film	144	4.12%	144	100.00%
Television	254	7.26%	254	100.00%
Pornography	453	12.95%	Unknown	Unknown
Games	59	1.69%	53	89.90%
Music	1,920	54.87%	1,786	93.00%
Other	108	3.11%	105	96.70%
Unknown	560	16.00%	Unknown	Unknown
Total	**3,500**	**100.0%**	**2,342**	**66.9%**
Excluding pornography and unknown	**2,487**	**71.06%**	**2,342**	**94.2%**

It was not possible to determine the copyright status of the pornography for which users searched. A large section of 'unknown' queries included many queries in Japanese (around one-fifth of all unknown queries) which could not be accurately translated. However, a majority of such Japanese queries for which translation *was* possible indicated that the search was likely for a pornographic video of some kind.

While it seems clear that music content is the most popular on the network – a finding supported by other research into Gnutella – there are some obvious methodological issues with using this process to calculate copyrighted content. For instance, search queries do not necessarily translate into downloads, particularly if the query cannot be matched exactly. Nonetheless, it is telling that 94% of the non-pornographic searches that could be identified were for copyrighted material. A similar study by Professor Richard Waterman of the University of

[26] Clients which act as 'supernodes' receive search queries from other peers on the network and other supernodes.

[27] For instance, a search for 'Lady Gaga telephone' was assumed to be a search for the audio version of this song. A search for 'Lady Gaga telephone video' or 'gaga video' was assumed to be looking for a music video. A search for 'telephone' could not be classified as any particular content type and was thus categorised as 'unknown'.

Pennsylvania which used a sample of 1,800 files found that 98.8% of files requested on Gnutella were either copyrighted or highly likely to be copyrighted.[28]

2.6.3 Usenet

Usenet is one of the oldest communications arena on the internet – and as with many areas of the internet, the system was quickly co-opted by those wishing to spread pirated content after its initial appearance. A few years ago, a small web site (recently shut down after legal action in the UK[29]) created the 'NZB' system for quickly retrieving large files from Usenet. NZB files opened up Usenet to a much larger potential audience and offered third-party services an opportunity to create businesses centred around facilitating access to Usenet. Some of these businesses, such as Usenext in Germany, are now multi-million Euro operations (Usenext had revenue of €30m in 2007). Significantly, almost all committed Usenet users pay for access: Usenext charge between €10 and €25 Euros per month and similar services do the same. The necessity to pay for access to Usenet has certainly limited the spread of the system as a way to obtain pirated content but Envisional believes that up to half a million users connect regularly to Usenet to obtain pirated content[30]. The usage studies cited in Part B that look explicitly at Usenet estimate overall traffic devoted to the arena at between 0.5 – 1% of overall internet usage.

Usenet began as a text-based medium meant for sending simple text messages. This remains the only real use for the system outside of transmitting files and it is unlikely that this aspect of the service takes up more than a tiny percentage of overall Usenet usage. In order to determine usage of Usenet for the transmission of copyrighted material, a random selection of 100 newsgroups from the many thousands available through the Giganews Usenet provider[31] were sampled and the last 100 complete files or messages posted to each newsgroup analysed. The copyright status of each post was checked. Text messages made up 3.2% of all posts; **93.4% of all posts** (all of which were files) contained copyrighted content; 2.3% were likely copyrighted; and for 1.1% of posts (all files), the copyrighted status could not be identified.

Thus at least 93.4% of sampled posts made to Usenet contain copyrighted content. However, given the size of these files (for instance, a typical film posted to Usenet will be at least 700MB in size), each post containing copyrighted content will dwarf the size of any text posts made. In terms of the actual amount of data transferred over the network, copyrighted material likely makes up more than the 93.4% of individual posts.

---

[28] See http://www.scribd.com/doc/31284309/Arista-et-al-v-Lime-Wire-et-al-summary-judgment.

[29] http://newzbin.com/

[30] An estimate made by reference to the amount of traffic received by major Usenet providers and NZB sites as well as through analysis of the published accounts of a large Usenet access provider in Europe.

[31] http://giganews.com/

3 Part B: Internet Usage Assessment

3.1 Introduction

This part of this research report critically evaluates recent research produced by a number of companies that offer different pictures of overall internet usage. Four main studies of bandwidth usage were examined. Each study was released during the second half of 2009 and were conducted by four network monitoring companies, mostly using data gathered during 2009:

- Sandvine Incorporated
- Arbor Networks
- Cisco
- iPoque

Each of the studies had the same broad aim: to illustrate the protocols and applications which are used across the internet and to show how much of the internet's bandwidth is used by each. For instance, each study analysed the amount of internet traffic taken up by peer to peer technologies or by streaming video as well as more traditional pursuits such as normal web browsing and email. However, direct comparison between each was problematic.

Each study:

- used different monitoring techniques
- was based on varying periods of time, examined different amounts of data and looked at different areas of the world
- used different categorisations for types of traffic

The categorisation issue is one of the largest problems with comparing the four studies. For instance, all four studies identify streamed video as a growing portion of internet traffic. However, each study uses a slightly different method of identifying this traffic and sometimes include the content in a different broad category which also comprises other items. For instance, Arbor Networks uses the simple term 'Video' to mean progressive video downloads; Sandvine speaks of 'Real-time entertainment' to denote video and other content such as audio which is consumed as it is downloaded or streamed; Cisco classifies 'Internet video to PC' as video or television on demand viewed on a computer; while iPoque uses the category 'Streaming' to refer to any kind of streamed audio and video. Some categories appear to be fairly consistent across all four studies: for example, all use 'P2P' as a broad identifier for known peer to peer networks. However, it was not always possible to determine the range of peer to peer networks detected by each monitoring company (though the largest known networks such as bittorrent, eDonkey, and Gnutella seemed to be always included), nor to know their rate of successful detection.

None of the four studies can be accepted without reservation, though some offered more confidence than others. The following sections discuss each of the four studies in detail, outlining the main points, the basis of the findings, and the methodological issues which are attached to each of them.

3.2 Sandvine: 2009 Global Broadband Phenomena

Monitoring period: September 1st – 22nd 2009

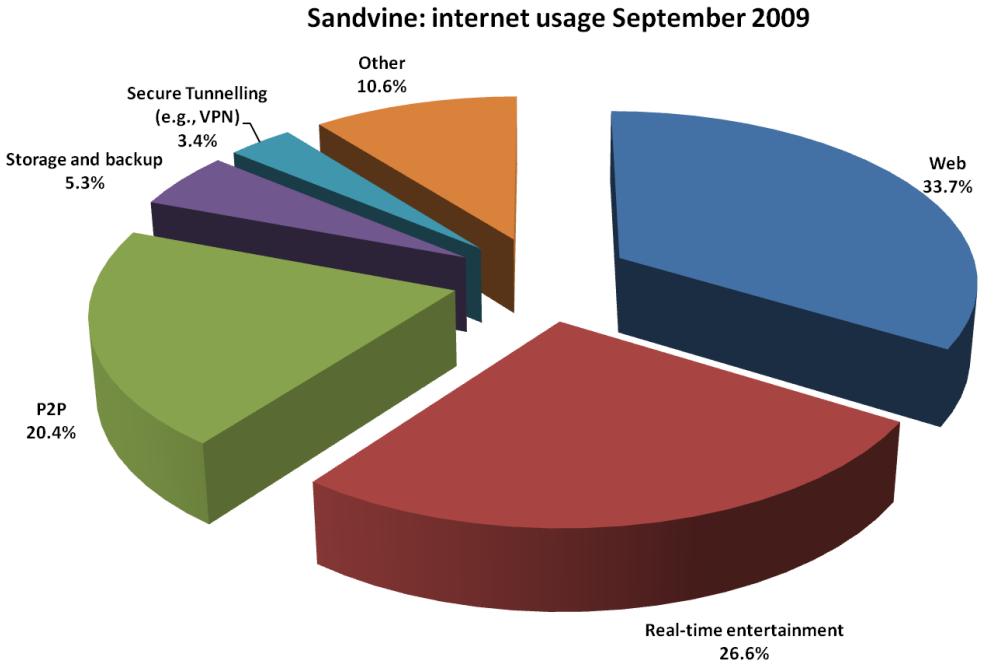
Monitoring locations: 22 ISPs in five regions: nine from North America, five from Europe, four in the Middle East and Africa, two in the Caribbean and Latin America, and two in Asia-Pacific

Number of subscribers: 24 million.

Amount of traffic monitored: Unknown

P2P traffic: 20.4%

Streaming video traffic: 26.6% (categorised as 'real-time entertainment' – content consumed as it is downloaded)



Other points:

- 'Storage and backup' services (which include cyberlockers and web-based backup services) consume 5.3% of internet traffic
- P2P proportion is 18.5% in North America
- Streaming video proportion is 26.7% in North America
- 'Real-time entertainment' category (streamed or buffered video or audio) more than doubled from 12.6% in 2008 to 26.6% in 2009.
- Significant variation between regions

3.2.1 Methodology

Sandvine is a Canadian-based network monitoring provider. The company's *2009 Global Broadband Phenomena* report repeated analysis completed in 2008. The study contained a detailed categorisation of content and thorough analysis of current trends based on 24 million subscribers from twenty ISPs in five regions, including nine ISPs located in the United States. Their data is based on internet traffic flowing through Sandvine's monitoring equipment and captures application usage from the subscriber's perspective. The company is also able to detect visitors to some popular web sites (such as YouTube and Rapidshare). Analysis looks at the average subscriber in a number of regions across the world and also uses a weighted global average of data to provide overall figures.

The main finding of the Sandvine study is the identification of a "*dramatic shift' from bulk download 'experience later' behaviour towards real-time 'experience now' application'*". Sandvine uses a category termed 'Real time entertainment' to denote streamed video or audio which is consumed as it is downloaded. In 2009, this category accounted for 26.6% of total traffic, an increase from 12.6% in 2008. The increasing consumption of video content by internet users is a common theme within most of the studies.

Sandvine issued a *2010 Global Broadband Phenomena* update as this Envisional report was being finalised. The 2010 report[32] did not provide data for worldwide traffic but found that 'real-time entertainment' continued to grow, accounting for up to 43% of peak time traffic in North America (with Netflix measured at 20% of peak time downstream traffic alone). Peer to peer traffic remained very important: bittorrent was found to comprise nearly 17% of downstream traffic during peak periods in North America and 37% in Latin America.[33]

3.2.2 Discussion

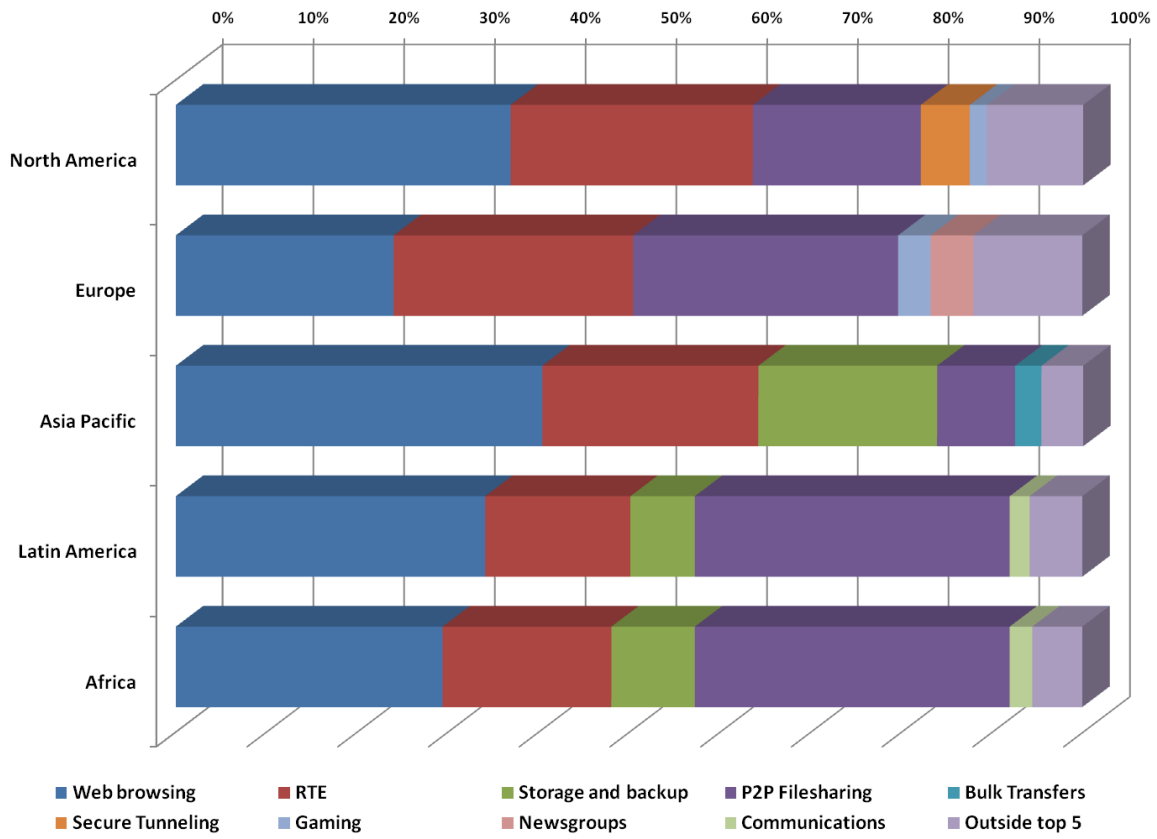
The chart on the page above illustrates the top five categories in terms of traffic detected worldwide by Sandvine. Web surfing contributes just over one-third (33.7%) of all traffic with the 'Real-time entertainment' (RTE) category responsible for more than one-quarter (26.6%, more than doubling in size since 2008). While much of this activity takes place through the web or browser it is separately categorised by Sandvine. Peer to peer filesharing then adds a further one-fifth (20.4%) of all traffic. More than 80% of internet traffic is thus taken up by these three categories alone. A 'storage and backup' category refers to cyberlocker sites such as Rapidshare which allow centralised file hosting and retrieval via the web (and which are often used for piracy purposes).

[32] http://www.sandvine.com/news/global_broadband_trends.asp

[33] There are some areas in which the 2010 report raises questions - for instance, in highlighting Zshare as the most popular cyberlocker in Europe. All other information gathered by Envisional from our own and other data sources cite Rapidshare, Hotfile, and MegaUpload as the three most-used cyberlockers with Zshare a second- or even third-tier site. For instance, data from ComScore place Zshare as the eighth most popular cyberlocker with one-tenth of the number of visitors of the most popular site.

Sandvine's report also makes clear that internet usage varies greatly across the world, a theme that is repeated in the reports from Cisco and iPoque. The chart below shows the top five categories of traffic in the different monitoring regions used by Sandvine. Some of the main differences are as follow:

- Web browsing as a portion of internet traffic ranges from 24% in Europe to 40% in Latin America
- P2P usage ranges from 8.6% in the Asia Pacific region to 34.7% in Africa
- Storage and backup services (online file hosts) are under 1.9% of internet usage in North America but 19.7% in Asia Pacific (influenced by the heavy use of centralised 'web-hard' services like PDBox in Korea, a location where Sandvine have monitoring equipment installed)
- Gaming traffic is (just) one of the five largest categories in North America and Europe but nowhere else.
- Newsgroups provide 4.7% of traffic in Europe but do not appear in the top five in any other region.
- Real-time communications traffic appears in the top five categories for Latin America and Africa but not elsewhere.

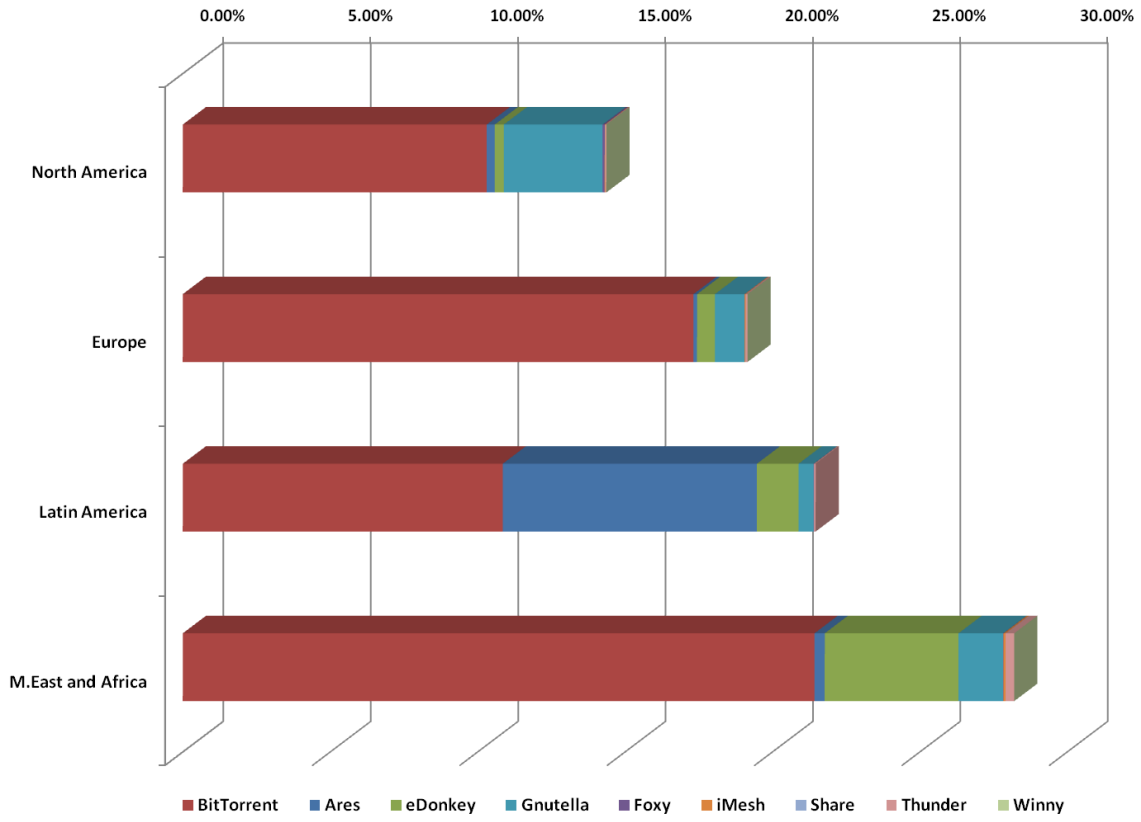


3.2.3 Additional detail

Sandvine provided Envisional with further detailed information on traffic from individual P2P applications as well as a small number of central web sites[34]. This additional data was broken down by four regions.

Sandvine tracked nine P2P applications: BitTorrent, eDonkey, Ares, Gnutella, iMesh (a client that connects to a legitimate music network), and four clients predominantly used in Asia: Foxy (a variant of Gnutella), Share and Winny (two popular Japanese networks), and Thunder (a download manager / P2P application popular in China where it is usually known as 'Xunlei'). Absent are some well-known protocols such as Shareaza and DirectConnect. It is unknown whether Kad, the decentralised sister network to eDonkey, was classified under the eDonkey header. (Peer to peer television (P2P TV) clients such as PPLive and PPStream are classified as 'Real time entertainment'.)

The chart below shows the percentage use of these P2P networks in each of the regions examined by Sandvine. Again, usage differed from region to region. The data is the average downstream usage of each application.

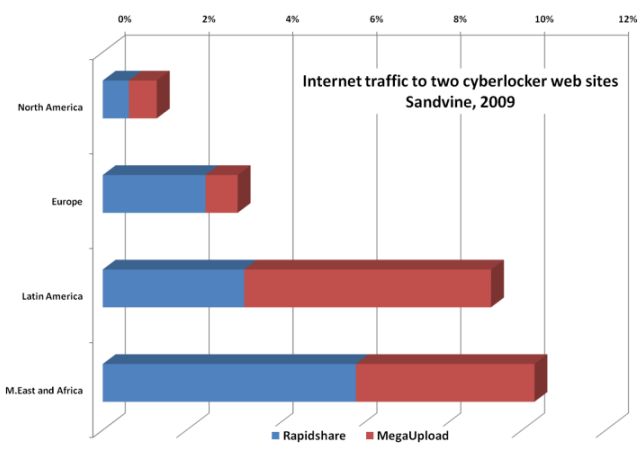


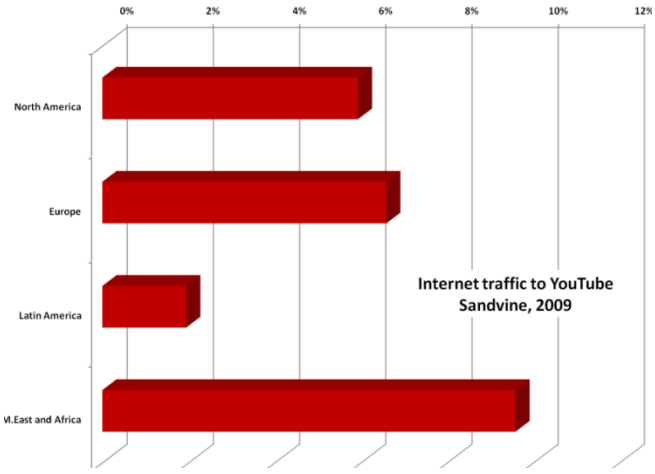
[34] Envisional is grateful to the author of the Sandvine study for supplying this additional data.

It is clear that BitTorrent dominates the peer to peer world in the locations monitored by Sandvine: the network makes up more than half of all peer to peer usage detected by Sandvine in each of these four regions and almost all in Europe. There is little eDonkey usage apart from in the Middle East and Africa. This finding likely reflects the countries in which Sandvine has monitoring locations in Europe: eDonkey is believed to be well used in many European countries such as France, Spain, and Italy and it would be difficult to believe that the network is responsible for just 0.3% of internet traffic in these such countries. Ares contributes over 8.6% of traffic in Latin America[35] while the four Asian clients comprise no more than 0.3% of all internet traffic in any of the four regions (unsurprising, as data was not supplied for the Asia-Pacific region).

- In **North America**, bittorrent (10.3%) and Gnutella (3.4%) make up almost all of the P2P proportion of overall internet traffic of 14.4% (the lowest of the four regions).
- 90% of P2P use in **Europe** is through bittorrent with the network making up 17.3% of all internet traffic in the region and Gnutella contributing a further 1% of all traffic. As noted above, eDonkey usage is believed to be higher in Europe than shown by Sandvine: other estimates place it at 3-5% of internet traffic.
- **Latin America** also sees more bittorrent usage than any other peer to peer application but Ares comprises 8.6% of internet traffic and 40% of all P2P traffic.
- BitTorrent contributes more to overall internet traffic (21.4%) in the **Middle East and Africa** than anywhere else while there is more P2P use (28.2% of all traffic) in this region than any of the other locations monitored by Sandvine, with eDonkey contributing 4.5% to all internet traffic.

Sandvine also supplied data to Envisional on visitors to the two most popular **cyberlocker web sites**: Rapidshare and MegaUpload. In the North America locations, 1.3% of downstream internet traffic was visits to one or another of these sites (MegaUpload was slightly more popular); in Europe, the figure was 3.2% of total downstream traffic (with Rapidshare much more popular); in Latin America, 9.3% of traffic went to these two cyberlockers (that is more traffic than used by the Ares application and almost as much as bittorrent); while in the Middle East and Africa, these two sites alone were responsible for 10.3% of all downstream internet traffic (with Rapidshare contributing 6% of all internet traffic alone).





[35] Including the Caribbean.

Across all four regions, these **two cyberlocker sites alone comprise 5.1% of all downstream internet traffic**. To put this into perspective, it is only a little less than the 6.2% of internet traffic consumed by YouTube worldwide, recognised by Sandvine (and Arbor) as the largest single domain contributor to overall internet traffic. As the second chart above shows, traffic to YouTube also varied from region to region, ranging from 1.9% in Latin America to 9.6% in the Middle East and Africa.

3.2.4 Summary

Sandvine's study shows a good level of detail and accompanying analysis. The company's willingness to discuss their approach and provide additional data upon request demonstrates their confidence in the methodology and figures.

However, it is important to remember the relatively small number of monitoring locations from which the data is drawn for some regions (only two locations for Latin America, Asia-Pacific, and the Middle East and Africa) as well as the fact that an overall figure for the amount of data analysed in the study could not be obtained. Further, analysis took place in September, a month when there are few major film releases and the Fall television season in the United States (which tends to produce an increase in the use of P2P networks to download content) is yet to properly begin.

3.3 Arbor Networks: ATLAS Observatory 2009 Annual Report

Monitoring period: July 2007 – July 2009

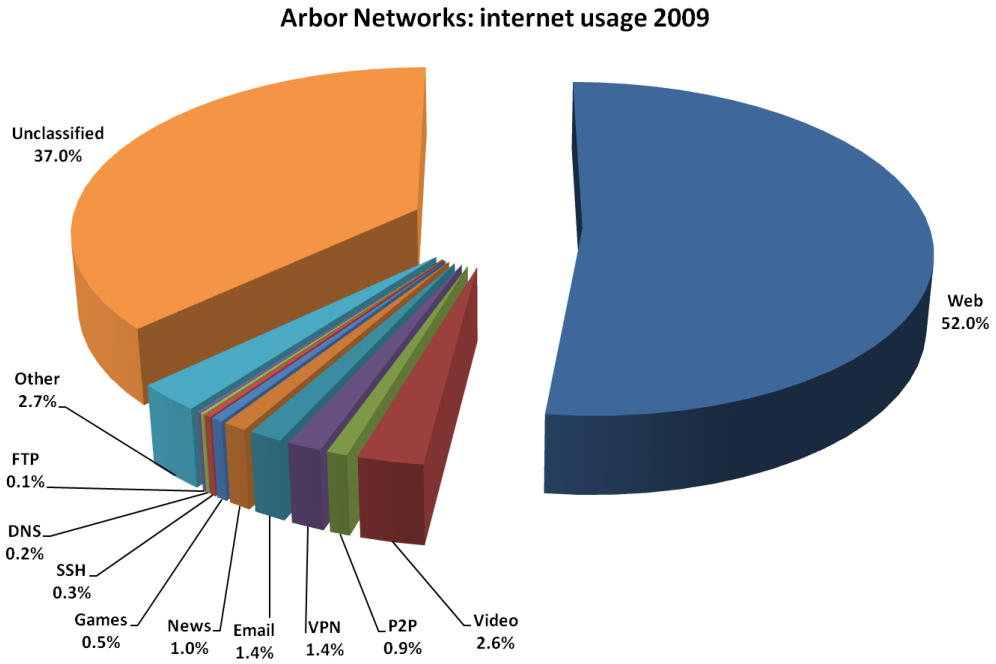
Monitoring locations: 110 deployments across ISPs and Content Providers worldwide (with an emphasis on North America, Europe, and East Asia (including Japan)).

Number of subscribers: unknown.

Amount of traffic monitored: 264 Exabytes of data at a peak rate of 14 Terabytes per second. On average, 9 Exabytes per month. This is by far the largest of all the studies[36].

P2P traffic: 0.85% (inspected by port number); 18% (payload inspection of a smaller dataset from 5 ISPs)

Streaming video traffic: 2.64% (estimate of 25% on payload inspection of same smaller dataset)



Other points:

- Streaming video is the fastest growing internet traffic category.
- Google (including YouTube) accounted for 5.5% of all internet traffic in May 2009.
- MegaUpload (a large 'cyberlocker' file host) accounted for at least 0.5% of all internet traffic in May 2009.
- Game console traffic accounted for 0.6% of all internet traffic in May 2009.
- Annual internet traffic growth of 44%.

[36] 264 Exabytes = 276m Terabytes = 283bn gigabytes = 64 billion DVDs.

3.3.1 Methodology

Arbor is an established network monitoring and security company. The company's monitoring study is produced in collaboration with authors at the University of Michigan and uses a number of monitoring locations worldwide that employ Arbor's network equipment. These servers sit on the edge of an ISP's network and categorise traffic as it passes with an 'anonymous XML file' containing data reports then sent to central analysis servers.

The Arbor study examines an extremely large amount of content data over a two year period – by far the most substantial data base of any of the four studies. The 264 Exabytes of data is equivalent to 283,500,000,000 Gigabytes – around 64 *billion* full-sized DVDs. The data is taken from a wider spread of monitoring points than others (110, compared to 20 for both the Sandvine and Cisco analyses and just 11 for the iPoque study). A precise breakdown of traffic by region is not outlined but monitoring appears to mainly use locations in North America, Europe, and East Asia (including Japan).

3.3.2 Discussion

The chart above shows the dominance of web-based communication: over half of all internet traffic identified by Arbor took place through the web. Against that, no other identified category was responsible for more than 3% of internet traffic. The video and P2P categories amounted to 3.5% in total.

However, the study is hampered –as the large orange 'Unclassified' segment on the chart makes clear – by issues with detection. In 2009, **37% of the 264 Exabytes of traffic could not be classified** by Arbor. This represents an enormous amount of traffic which could not be identified by the routine monitoring techniques employed by the company. According to subsequent analysis by Arbor, the majority of this unclassified proportion is believed to be either peer to peer traffic or video streaming and downloads, a belief based on analysis of a second and smaller dataset of traffic subjected to more detailed probing.

This second, smaller, dataset was taken from **five consumer ISPs** based in the United States, Canada, Europe, and Asia, though the precise locations and number of subscribers represented are not supplied, and nor is the actual amount of data analysed. This dataset was subjected by Arbor to Deep Packet Inspection (DPI) techniques in an attempt to detect traffic based on the payload of the data. Arbor are confident that their DPI detection is accurate, but detection of peer to peer protocols is not their core business and as such, they may not be catching as much of this traffic as some other companies – certainly, it might be expected that they under-measure P2P than over-measure. However, Arbor were clear in conversation that observations show that there is broad correlation between the overall trends from the smaller DPI-based analysis and the larger, main dataset though without detailed analysis of the smaller dataset this is not possible to confirm.

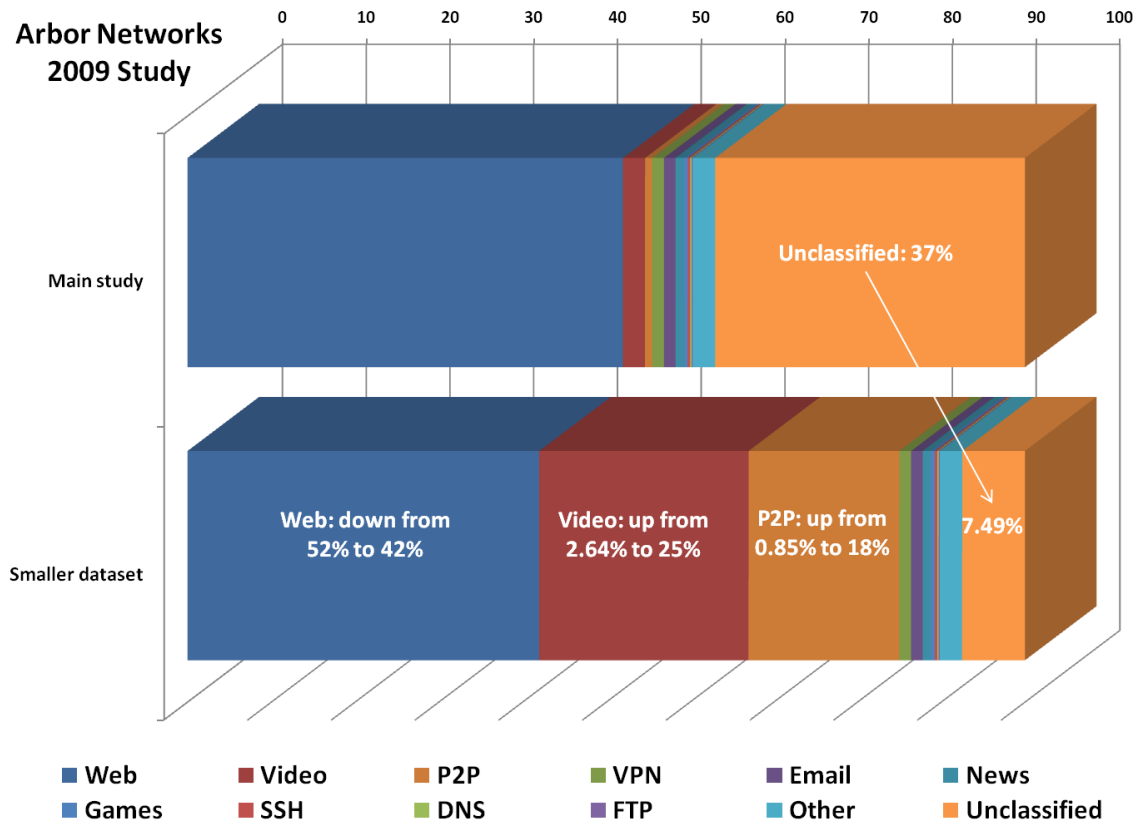
The deeper analysis of the second and smaller dataset via DPI led Arbor to conclude that "**P2P is likely closer to 18%**". This wording is imprecise and there is no attempt to break down P2P usage by protocol or by region, as

Sandvine manage, for instance[37]. The dataset was taken from the midpoint of the monitoring (assumed to be during 2008) and no further information is provided on additional changes to P2P traffic after that point.

Similarly, the larger main study appeared to base its analysis of video traffic on older protocols and did not account well for the enormous growth of other transmission methods. A second estimate is made by Arbor through similar DPI analysis of the smaller dataset which estimated video traffic at "**25%+ of all traffic (including 10% of HTTP)**". Again, the wording is vague and slightly confusing, drawing part of the HTTP traffic to make up the total video proportion.

3.3.3 Accounting for the unidentified data

As noted, **37% of traffic** from the main study was unidentified in 2009. The smaller dataset placed **P2P** usage at 18% rather than 0.85% in the main study, which might account for 17.15% of that unidentified block – leaving 19.85% of unidentified traffic. Some of that may also be accounted for by the **video** data identified by the smaller dataset.



[37] The authors do state that P2P varies by region and type of network but this is not elaborated upon.

The smaller dataset identified 25% of data to be video traffic rather than 2.64% in the main study, a difference of 22.36%. However, that figure of 25% for video includes 10% of previously identified HTTP traffic, leaving 12.36% of traffic which can be taken from the unidentified block of traffic.

So if the assumption is made that the smaller dataset portrays similar overall usage patterns to the larger study (and there must obviously be reservations about doing this, given the smaller amount of data and lower regional coverage), calculations then leave **7.49% of traffic unidentified** (37%: 17.15% identified as P2P – 12.36% identified as video).

The chart above shows how the overall usage pattern from the main study significantly changes if the classification of video and P2P usage by the smaller dataset is accepted as correct. While the smaller categories of use (such as email and FTP) remain the same, the three major categories of identified use from the smaller dataset (web, video, and P2P) show large differences.

It is possibly only to speculate what the remaining 'unidentified' amount of traffic might be: given Arbor's primary focus on network monitoring and security, it is possible that some of this data may be peer to peer or other file sharing traffic. Arbor do not mention protocols like those behind 'P2PTV' applications such as PPLive and Sopcast that are often used for video distribution in Asia in their reporting and these may also make up some of the unidentified proportion.

3.3.4 Summary

In summary, the Arbor study, while clearly based on a vast treasure trove of data, is affected by the large amount of that treasure which could not initially be classified. Additional DPI inspection of a smaller dataset provided some additional insight but it is only rational to accept the figures provided for P2P and video consumption in particular as a broad estimate of data usage online rather than a more exact representation.

The issues involved in estimating P2P and video traffic must also affect confidence in figures for the other categories of traffic – although as many of these categories will have changed little over time (for instance, web-based transmissions, email, FTP, and VPN traffic are well established), detection and categorisation should be easier.

3.4 Cisco: 2009 Visual Networking Index Usage Study

Monitoring period: Third quarter of calendar year 2009

Monitoring locations: Over 20 service providers, mostly consumer-based ISPs.

Number of subscribers: 1m

Amount of traffic monitored: Unknown.

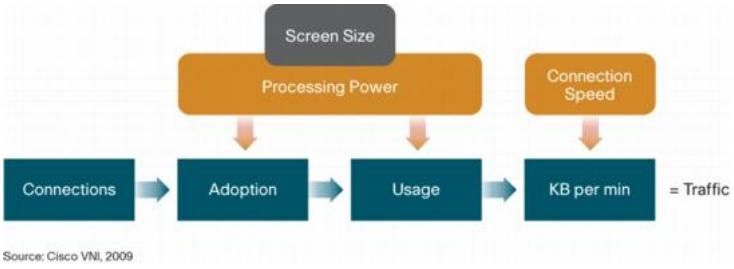
P2P traffic: 38% (worldwide)

Streaming video traffic: 27.7% (worldwide); 30.7% (United States)

Cisco regularly publish data on internet traffic and bandwidth usage within an ongoing research initiative known as the **Visual Networking Index**. The majority of published work within this initiative is based on the interpretation of analyst predictions about the future of internet usage. For these studies, Cisco state:

> *The core methodology relies on analyst projections for Internet users, broadband connections, video subscribers, mobile connections, and Internet application adoption. Analyst forecasts come from SNL Kagan, Ovum, Informa Telecoms & Media, Infonetics, IDC, Frost & Sullivan, Gartner, ABI, AMI, Screendigest, Parks Associates, Yankee Group, Dell'Oro, and Synergy.*

Cisco produces data on the overall use of the internet for the VNI by combining these analyst predictions with an analysis of what are termed 'fundamental enablers' of internet usage such as broadband speed, computing power, and screen size, with the company positing a 'supply-side' aspect to internet usage as well as an end-user demand aspect.



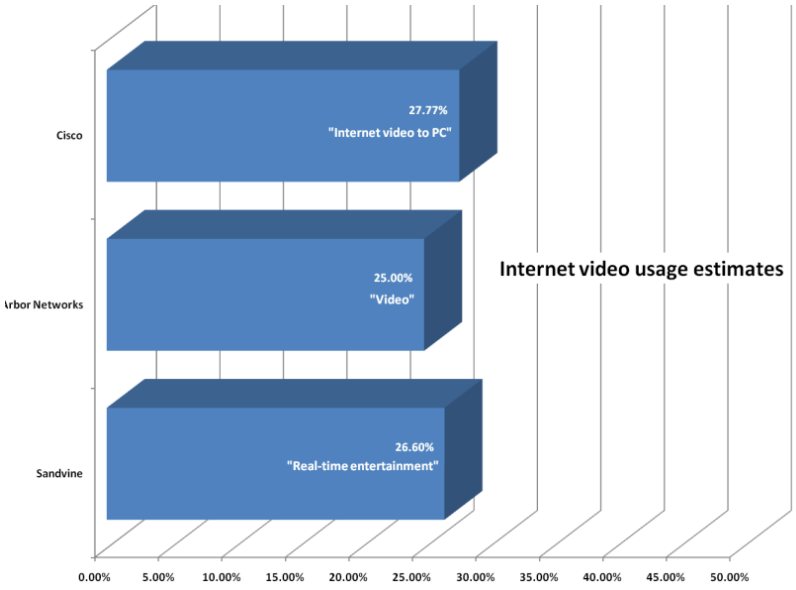
For the purposes of this study, Cisco's analysis is helpful as context but does not provide hard data based on the monitoring of actual internet traffic. However, Cisco also publish a Visual Networking Index **Usage Study** which draws data from over twenty ISPs worldwide serving a total of around 1m subscribers. This uses deep packet inspection to determine the type of data flowing into and out of each ISP.

Unfortunately, the amount of data publicly available from the Usage Study is low and in terms of categorising network traffic, only specific figures for file sharing usage are provided by the company.[38] This finds that **38% of global internet traffic can be identified as peer to peer**. The report also finds that "Nearly one-third of all file-

[38] The study also provides some data which states that the average broadband connection generates 11.4 gigabytes of Internet traffic per month and that the top 1% of broadband connections are responsible for more than 20% of total Internet traffic.

sharing Internet traffic is non-P2P. Web-based file-sharing, newsgroups, and FTP account for 32 percent of all file sharing traffic." This means that in total, **55.9% of all internet traffic is what Cisco term file-sharing**. However, the data is not broken down by protocol or type of traffic – for instance, it is not known what proportion of the 38% that is peer to peer file sharing is produced by bittorrent or eDonkey; or how important 'web-based' file sharing is, nor exactly which sites are listed under that definition.

Cisco does provide significant detail within the main VNI studies, allowing data estimates to be broken down by country, type of traffic, and for a number of years going forward through a customisable web-based tool. However, as these estimates are based on analyst predictions (and as they differ from that produced within the actual Usage Study (for instance, peer to peer is listed as 31.7% in 2009 rather than 38%), their methodology makes them unsuitable for inclusion in this report. It is worth noting that the estimate for video streaming bandwidth use is very similar to that produced by Sandvine and Arbor, as the chart shows. Cisco defines this as "internet video to PC" and estimate it at 27.7% of all internet usage. This is relatively close to the estimates from Sandvine for 'Real-time entertainment' (26.6%) and Arbor's 'Video' category (c.25%) – though again, note that the figures from Sandvine and Arbor are based on actual monitoring data rather than analyst estimates.



3.5 iPoque: Internet Study 2008/2009

Monitoring period: "Two weeks", varied periods depending on location.

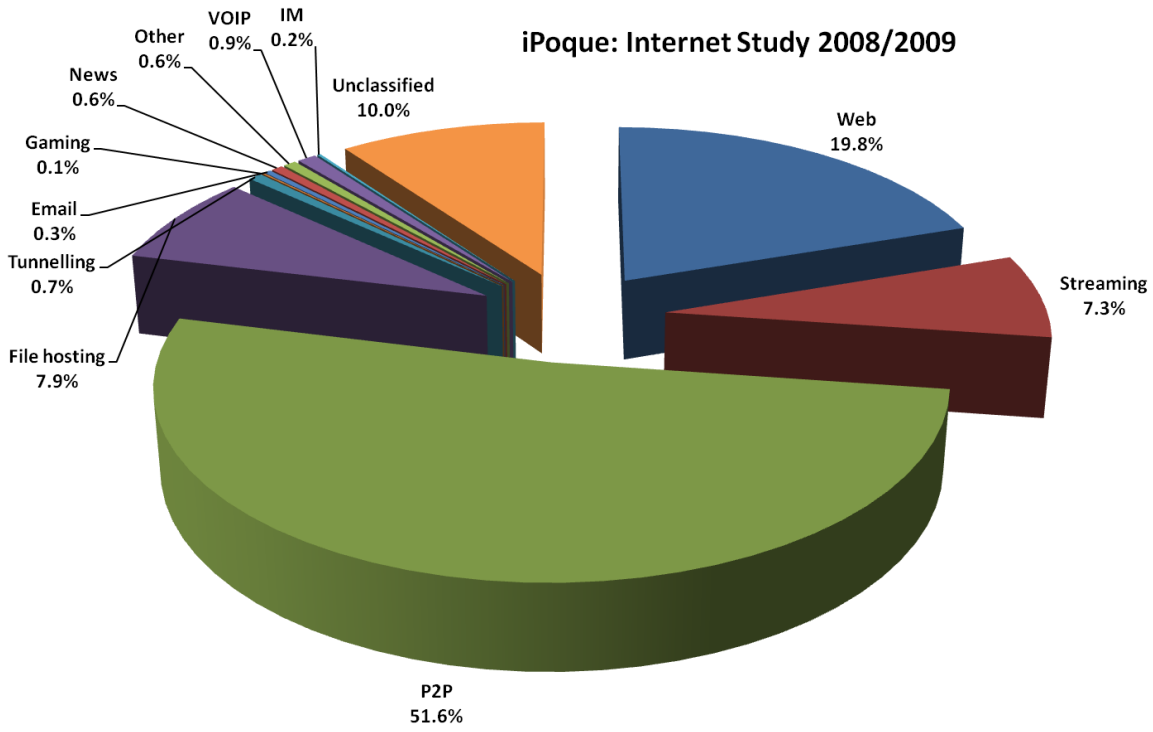
Monitoring locations: 11 monitoring locations; eight ISPs and three universities from eight regions: Africa, South America, Middle East, Eastern, Southern, and Southwestern Europe, Germany. No locations in the United States.

Number of subscribers: 1.1m

Amount of traffic monitored: 1.3 Petabytes (the smallest of the three studies where traffic amounts are known).

P2P traffic: 51.6%

Streaming video traffic: 7.34% (categorised as 'Streaming')



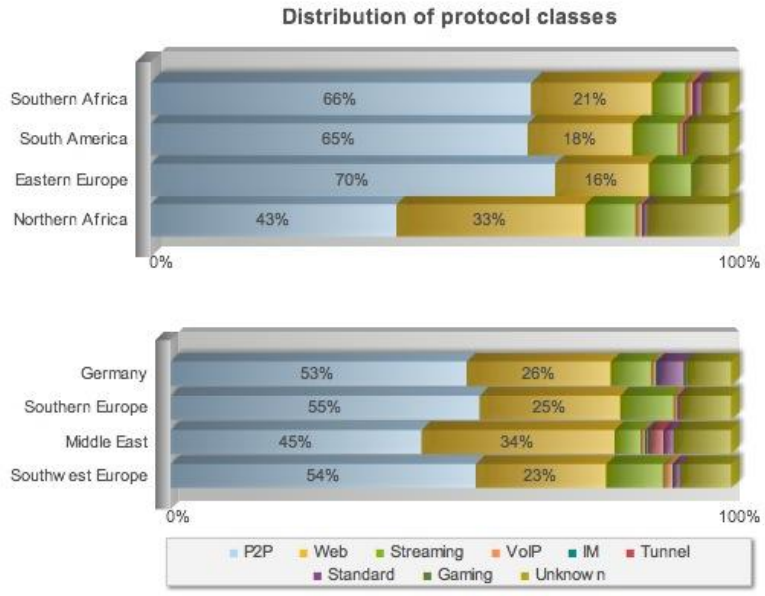
Other points:

- Peer to peer file sharing generates "by far the most traffic in all monitored regions" – from 43% in Northern Africa to 70% in Eastern Europe.
- Peer to peer traffic has dropped slightly as a proportion since their previous 2007 study. BitTorrent is the most popular single protocol.
- File hosting (direct downloads from cyberlockers) has increased to "up to 45% of all Web traffic" in some regions.
- "Rapidshare alone is responsible for 5 percent of the worldwide Internet traffic".

3.5.1 Methodology

iPoque is a German network monitoring and DPI solutions provider. They claim to be the leading European company in their field. The company has issued 'Internet Study' reports each year since 2007. The 2009 report is detailed in its results and discussion but based on a small amount of traffic[39] generated from eleven locations at different (unknown) time periods and which cover a relatively small number of users (1.1m subscribers in total compared to 24m for Sandvine). Each location uses iPoque's PRX Traffic Manager hardware which combines protocol detection with DPI and behavioural traffic analysis.

The eleven locations themselves are scattered around Africa, Europe, and the Middle East, with only one or two locations in each country. Three of the locations are universities where user profiles and bandwidth usage are likely to be significantly different to a consumer ISP. The study notes that the various issues with the data (amount gathered, locations, types of network, time period, number of subscribers) mean that "the results are not statistically representative".



3.5.2 Discussion

The chart above, taken from the iPoque report, again shows the significant variation from location to location of different types of internet usage but also shows substantially different results – particularly for P2P usage – compared to the other three main studies.

[39] Arbor's study is based on over 200,000 times as much data.

- P2P is the highest single category in every region, ranging from 45% in the Middle East to 70% in Eastern Europe, far higher than other studies.
- Web use differs from 16% in Eastern Europe to 33% in Northern Africa.
- The 'streaming' category (defined by iPoque simply as audio and video streaming) takes up anything from 5.8% to 10.1% depending on location but does not come close to the one-quarter of internet traffic identified by the preceding three studies.

It is possible that the locations studied by iPoque simply represent areas which show significantly different internet usage to those monitored by Sandvine, Arbor, or Cisco. Previous reports from iPoque have historically shown much higher P2P usage than other monitoring companies: given the commercial focus of the company on the detection of file sharing protocols (and their equipment does appear able to detect an enormous range of protocols), it is also possible that iPoque may be able to detect some traffic which other monitoring companies might miss or be able to more accurately identify protocols. However, the variation is such that this cannot be the sole reason for the differences.

In summary, the iPoque report indicates that peer to peer traffic is very high in most of the monitoring locations from which they have obtained data while streaming is lower than shown in the other three studies. However, it is difficult to generalise from their findings to other locations and, in particular, to other countries. iPoque has good knowledge and capabilities in identifying different protocols but as a study of use in determining bandwidth make-up worldwide and in the United States, the report must be used with caution.

3.6 Focused studies

Two recent academic studies of network usage were also uncovered. Each examine only a single ISP and as such, the ability to generalise from the results will be difficult but each provide some findings worthy of discussion.

3.6.1 Maier et al (2009) - On Dominant Characteristics of Residential Broadband Internet Traffic

Maier et al. studied traffic for 20,000 subscribers from a major European ISP within a single urban area at various points during the second half of 2008 and the first half of 2009.

The study found that HTTP comprised 57.6% of all traffic with bittorrent responsible for 8.5% and eDonkey for 5% of traffic. At least one-quarter of all HTTP traffic carried Flash video with a further 7.6% carrying other video.

Just fifteen domains accounted for 43% of all HTTP traffic (and therefore 26% of all internet traffic). A single cyberlocker / direct download provider was responsible for 15.3% of *all* HTTP traffic (related to this, 14.7% of all internet traffic was in the form of RAR archives, commonly used in cyberlocker or newsgroup downloads).

Rank	Domain	Fraction of Traffic
1	Direct Download Provider	15.3%
2	Video portal	6.1%
3	Video portal	3.3%
4	Video portal	3.2%
5	Software updates	3.0%
6	CDN	2.1%
7	Search engine	1.8%
8	Software company	1.7%
9	Web portal	1.3%
10	Video Portal	1.2%

3.6.2 Erman et al. (2009) - Network-aware Forward Caching

Erman et al. examined internet traffic from 100,000 broadband subscribers from three states from a single broadband provider in the United States. The data analysed was taken at regular points from February 2007 to September 2008. The authors concluded that “HTTP… is increasingly being used to handle most of the Internet’s tasks such as distribution of software, updates, patches, and multimedia, and by P2P applications”.

- 66% of internet traffic was HTTP; the web is “the workhorse for data delivery”
- Video was a large portion of HTTP and around 22% of all traffic
- 12.3% was P2P (though this portion could be up to 17% given issues with identification)

3.6.3 Layton and Waters, Internet Commerce Security Laboratory (April 2010) - Investigation into the extent of infringing content on BitTorrent networks

This study was on the surface similar to the investigation pursued in Part A of this report into infringing content on bittorrent. The authors gathered data from a range of bittorrent trackers and collated the information, then looked at the most popular 1,000 individual files.

The authors found that 43% of the sample of 1,000 bittorrent swarms was films, 29% was television episodes, and 16.5% was music.[40] The proportion of torrents infringing copyright was estimated at 89% with no evidence of legitimate usage found in the torrents within the top three categories (film, television, and music).

However, the study had significant methodological flaws and as such, Envisional believes it should not be considered as valid for the purposes of this report.

- The authors chose the most popular 1,000 torrents in terms of **number of seeds** rather than number of downloaders. It is common for fake files or malware to have seed numbers artificially boosted in order to attract downloaders. Little or no work appears to have been done in weeding out the fake files, resulting in a peer count of over 117m seeds across only 1m torrents (compared to just 17m peers in *total* for all 1.8m torrents tracked by PublicBT in Part A).
- There was an issue with **domain pseudonyms** for common trackers. Some of the tracker names used in the data gathering actually point to an IP address for a different (and more popular) tracker altogether[41].
- There are a number of instances where the **reported data** stretches credulity: for instance, at the point of their analysis in April 2010 the most popular file was listed as a pirated version of the film *The Incredible Hulk*. This film was released in 2008 and was not one of the most popular that year, yet the data produced by the authors state that one version for the film had over one million peers, both a level of popularity that is difficult to believe for a film of this age and an absolute number of downloaders that is higher than any single bittorrent swarm ever recorded by Envisional. For example, the number of seeds in the most popular swarm for the final episode of the television program *Lost* – believed to be the highest-seeded bittorrent swarm ever seen – was never above 100,000 at any one time, according to Envisional's own monitoring.

[40] http://www.icsl.com.au/files/bt_report_final.pdf

[41] For instance, the tracker address "tracker.ilibr.org" points to the PublicBT tracker: a query to the ilibr.org tracker is actually sent to the PublicBT tracker instead. With both the Ilibr tracker and the PublicBT tracker included in the data gathering, the same information is being gathered twice. Further, *two* versions of the ilibr.org tracker are included on two different ports - yet these both point to PublicBT and will end up querying the same tracker twice (the port numbers make no difference in this aspect).

3.7 Summary: Bandwidth Usage

As the preceding discussion makes clear, navigating through studies of internet traffic in order to attempt some level of consensus is challenging. With no established or accepted methodology, classifications, or measurement techniques, the analyst depends on the detail provided in each study to assign confidence and gain understanding.

Each of the four main studies discussed have methodological issues of a greater or lesser extent.

- **Sandvine's** report is detailed but the amount of traffic on which the analysis is based is not provided. Given that the focus is upon three weeks of analysis across 24m ISP subscribers, the data volume should be significant. Further, the methodology is outlined clearly and the company was also willing to discuss their approach and send further data when requested.
- The **Arbor** study is based on a volume of data which dwarfs all other studies but detection is poor and while a smaller dataset is analysed to allow more precise measurement of certain sectors, confidence is obviously affected.
- **Cisco** provides only a few data points. Their main VNI reports provide granular data for a wide range of applications and countries yet rely on analyst predictions rather than data measurement. The focus is much more on predicting network growth than on detailing traffic for a particular time period.
- **iPoque's** report relies on a limited sample of data from varied dates across a small range of monitoring locations in less developed internet markets.

Apart from Arbor who do not analyse traffic in this manner, all studies show significant regional variation. Internet usage in North America is clearly not the same as in Latin America or Europe or Asia. The variations shown for instance by iPoque across monitoring locations in the same small region demonstrate that there can be large variations between countries (and likely within countries, also). Envisional's own monitoring data for networks like bittorrent and eDonkey show differences between countries in usage of those protocols.

With the limitations of each study in mind, it does seem possible to generate some broad conclusions and estimates about internet traffic using the data provided.

3.7.1 The importance of the web

- Standard, daily, routine **web browsing** – to Google, Facebook, the BBC, Wikipedia, Twitter, Amazon, eBay, Flickr, blogs, forums, and so on – is responsible for around **one-third of all internet traffic**. It may be difficult to be more precise than this: so many applications and sites employ the web for distribution or storage of content that categorisation becomes difficult. Sandvine and Cisco appear to ensure that most web traffic which is not web-page based (such as video streams, file hosting downloads, and so on) are categorised separately but the two studies diverge significantly over how much traffic is then left: Sandvine posits 33.7%; Cisco's VNI study estimates just 18.2% (a figure which also includes email and instant messaging data). Arbor finds that 42% of internet traffic is 'Web' while iPoque estimates anywhere from 16% to 34% depending on location (with

Web: 33%

this figure including file hosting sites). The smaller Maier and Ermann studies find around 35% of traffic to be non-video HTTP traffic. For this report, the amount of web usage is held to be 33% of all internet traffic.

- In the **United States**, the maturity of the web and its place as home to so many applications which have extended the use of the web – Google, YouTube, Facebook, Twitter, and so on – mean that relative web use in the US may be higher than that observed worldwide. Both Sandvine and Cisco (the only two of the four studies to analyse the US or North America separately) report or estimate slightly higher web use in the country.

- Beyond everyday web browsing, there are two other areas of web-based traffic which should be considered separately: **streamed video** (and to a lesser extent, audio); and **file hosting** or cyberlockers.

    - **Video content**, particularly streamed video, is one of the major components of internet traffic, with much of it being transmitted through or sourced from HTTP communication. Three of the studies reach a broad level of consensus on the level of internet traffic which features streamed content: Sandvine's 'Real-time entertainment' category, Cisco's 'Internet video to PC' estimate, and Arbor's simple 'Video' category all place web-based video viewing at around **25%-28% of traffic** (and is assumed to be 26.5% for the purposes of further analysis in this report). Sandvine's category includes audio-only streams and Arbor's category is hardly defined at all but the figures are relatively close in agreement. This is an area where the iPoque study shows considerable difference to the other three reports. It is possible that streamed video is less important in the locations where their measurement technology is based but without further detail on the countries from which iPoque are reporting, this can only be speculation.

      Video: 25-28%

      On-demand video content appears to be consumed more highly **in the United States** (the home of YouTube and many other online video hosts) than in other regions: ComScore reported that over 173m internet users in the US watched more than 32bn videos during January 2010 alone, significantly higher figures than for users in Germany and France, for instance. With this in mind, an estimate for video usage in the United States as comprising 27%-30% of internet traffic can be made.

    - The use of central web-based **file hosting sites or cyberlockers** such as Rapidshare and MegaUpload can be significant depending on country. These sites seem to be more heavily used in Europe and less developed internet markets (such as the Middle East and Africa) than they are in North America. Sandvine estimate that cyberlockers are responsible for around 5.3% of all internet traffic, and this should be seen as a minimum – the company's list of sites included in the 'Storage and backup' category is far from exhaustive for cyberlockers. However, no other cyberlockers are as large as Rapidshare and Sandvine provides detailed traffic analysis for that site and for MegaUpload. Thus while actual cyberlocker usage may be higher than Sandvine's figure, it is likely not much higher.
      iPoque believe that Rapidshare alone contributes 5% of all traffic and that cyberlockers overall are responsible for 7.9% of traffic though this comes from countries where cyberlocker usage appears to be relatively high.
      Cisco do not delineate this area specifically but do estimate 'non-P2P' filesharing (web-based file sharing, newsgroups, and FTP) at around 19%. It is reasonable to assume that most of this non-P2P filesharing will be from cyberlockers as newsgroups and FTP are shown in other studies to be around

      File hosts / cyberlockers: 7%

1% of all internet traffic and little more. As with their estimate for peer to peer usage (see below), Cisco are therefore estimating a much higher level of cyberlocker usage.
Arbor are fairly quiet on this issue, stating only that MegaUpload was found to be responsible for around 0.6% of all internet traffic in early 2009.
Analysis of the overall data available leads to a cautious estimate that central file hosts like cyberlockers are responsible for around **7% of internet traffic**.
Data from Sandvine – the only source of information on this area – show relatively low usage of the two main cyberlockers for users from North America. Given this, an estimate of cyberlocker usage for the United States of **3%** is acceptable.

3.7.2 Peer to peer remains significant

- **Peer to peer applications** have traditionally been considered to take up a very large amount of internet traffic: studies from 2005 found that more than half of all internet traffic used peer to peer. That may have been correct at that time but as noted above, there has since been a resurgence of the importance of the web to internet users at the same time as the internet has become increasingly a video-based medium. This is not to say that peer to peer traffic is declining in absolute terms.
  Determining how much internet traffic is peer to peer is more difficult. The proportion varies from study to study and, within those studies, from region to region, but it is likely that at least 20% of all internet traffic comes from peer to peer applications. Sandvine's figure is 20.4% worldwide and this may be slightly low. The list of P2P protocols included in their study is not exhaustive, though does include the major networks. However, both iPoque and Cisco place P2P usage much higher: the former at over 51% and the latter at 38%. iPoque's figure can only be taken as evidence of P2P usage in the particular locations they monitor. Cisco's figure is from a relatively small sample of 1m subscribers but accords with the higher figure they estimate from analyst predictions.

  Peer to peer: 25%

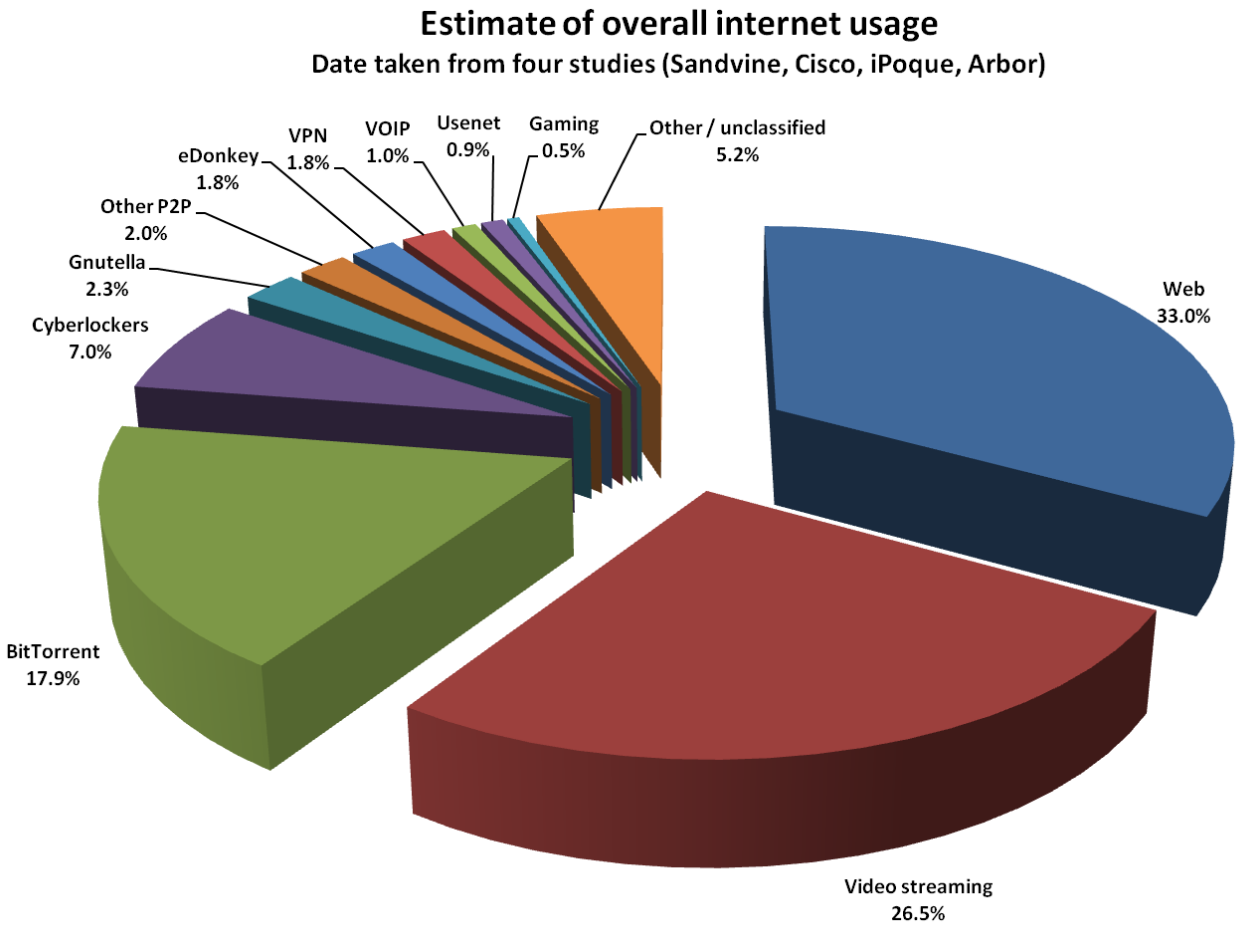
  Given these issues, this analysis estimates P2P usage worldwide at **25%** of all internet traffic. On this reading, bittorrent uses around 17.9% of all internet bandwidth.[42]

- The **United States** appears to be one of the lowest relative users of peer to peer worldwide: Sandvine measure aggregate (downstream and upstream) peer to peer traffic at 18.5% in North America and 14.6% for downstream, mostly through bittorrent. Similarly, Cisco's estimate falls from 31.7% for worldwide P2P usage to 23.9% for the United States alone. There is thus less of a gap between the two studies to reconcile. Assuming US P2P usage to be around **20% of internet traffic** seems reasonable with bittorrent at 14.32% and other peer to peer traffic accounting for just over 5%.

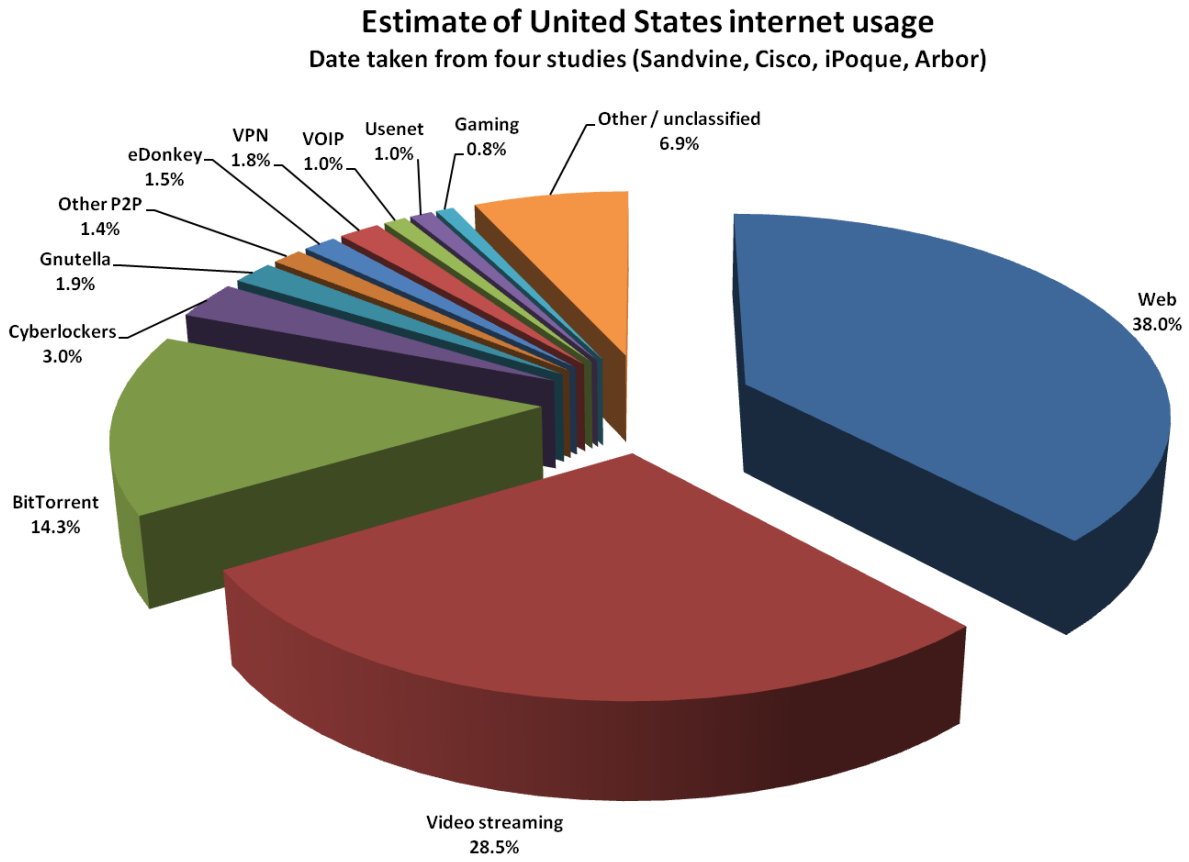
[42] Only Sandvine provide an overall figure for the amount of network traffic for which bittorrent is responsible: 14.1% (or 71.6% of all peer to peer traffic). If Sandvine's peer to peer estimate of 20.4% is taken as slightly low and the figure of 25% is assumed for all peer to peer data, then the overall figure for bittorrent would be extrapolated to 17.89% of all internet traffic.

3.7.3 Overall estimate

The chart below uses Envisional's own analysis experience and internet intelligence to draw together the four monitoring studies in order to produce an overall estimate for internet bandwidth usage. Web traffic and video streaming (most of which takes place through the web or via HTTP) makes up almost 60% of all bandwidth. BitTorrent provides another 17.9% with peer to peer overall contributing 25% of internet bandwidth. Areas of internet usage such as VPN tunneling, voice over IP, and gaming, are estimated by each of the four monitoring companies to contribute much smaller amounts of overall bandwidth.



In the United States, the higher relative use of the web and video streaming means that these two components are responsible for two-thirds (66.5%) of all bandwidth. BitTorrent usage is slightly lower at 14.3% with other peer to peer protocols contributing a further 5.7% of internet bandwidth. Cyberlocker usage is estimated to be lower in the US than elsewhere in the world, while Gaming and Usenet consumption is very slightly higher.



Part C of this report brings together these overall estimates from Part B with the analysis of common piracy arenas found in Part A to provide an estimate of the amount of internet traffic overall which is believed to be infringing.

piracy.intelligence@envisional.com

4 Part C: Drawing the data together

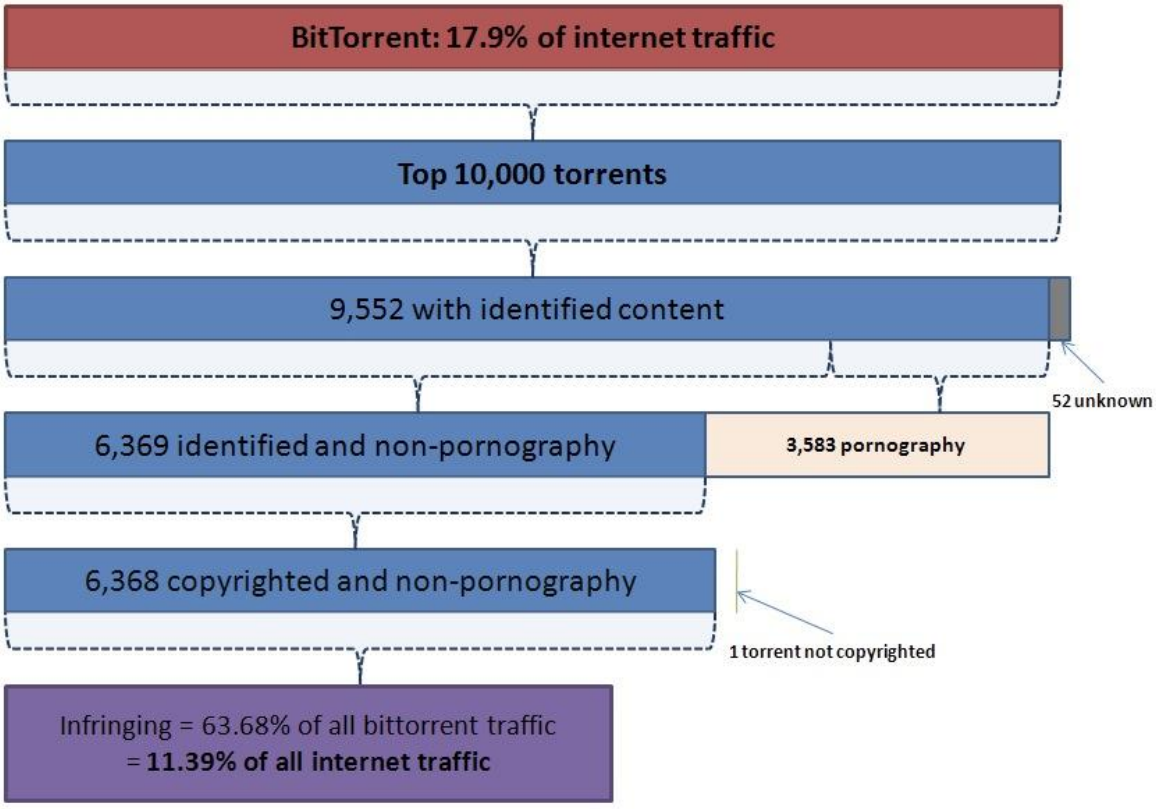
4.1 Introduction

Part A of this report examined a range of common internet arenas where pirated activity is often found and attempted estimates of the level of infringing activity found within each. Part B critically assessed four studies that attempted to determine the amount of overall internet bandwidth used by different protocols and types of content.

This final part of the report brings together these two parts in an attempt to provide an overall estimate for the amount of all internet traffic likely to be infringing. Each of the common piracy arenas examined in Part A will be summarised in turn.

4.2 Estimates of infringing use

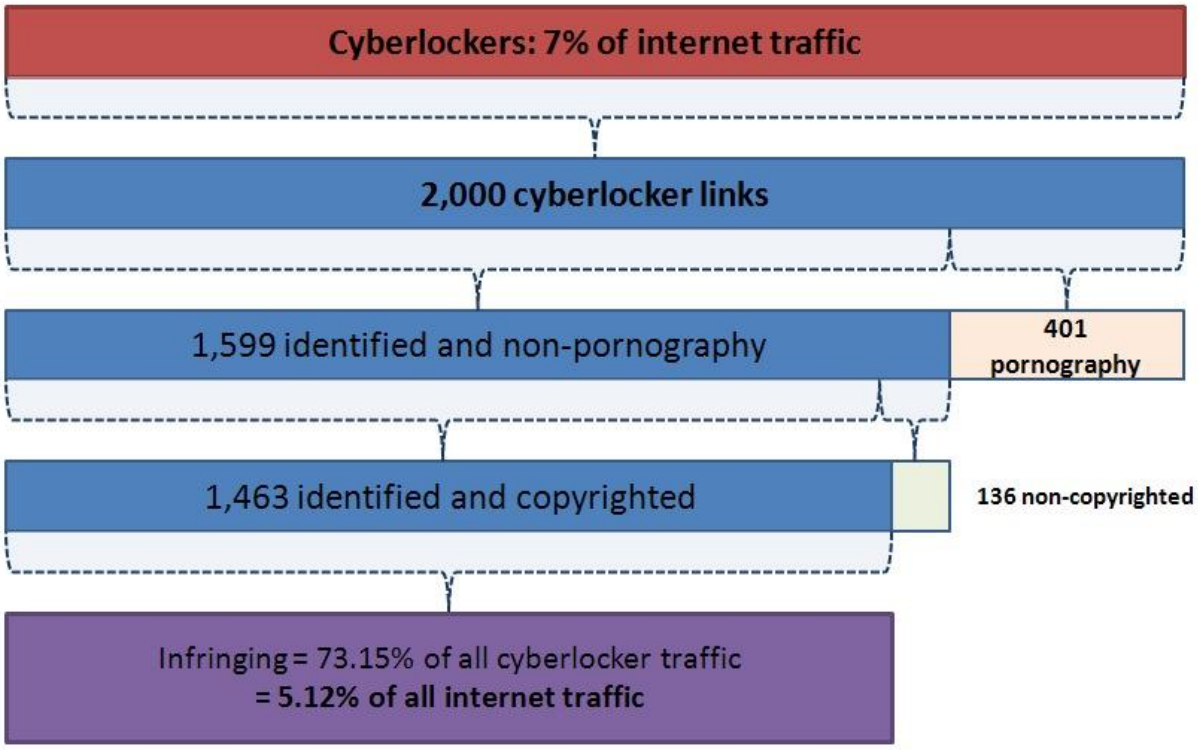
4.2.1 BitTorrent



The chart for bittorrent starts with the estimated 17.9% of internet bandwidth which is believed to be bittorrent. The amount of that bandwidth deemed to be infringing is estimated by reference to the analysis of the most popular 10,000 torrents held on PublicBT (found in Part A of this report). 63.68% of these torrents were found to be infringing and not pornography. This means that 63.68% of the internet bandwidth consumed by bittorrent can be estimated to be of infringing content, resulting in a final estimate that **infringing use of bittorrent is responsible for 11.39% of all traffic on the internet**.

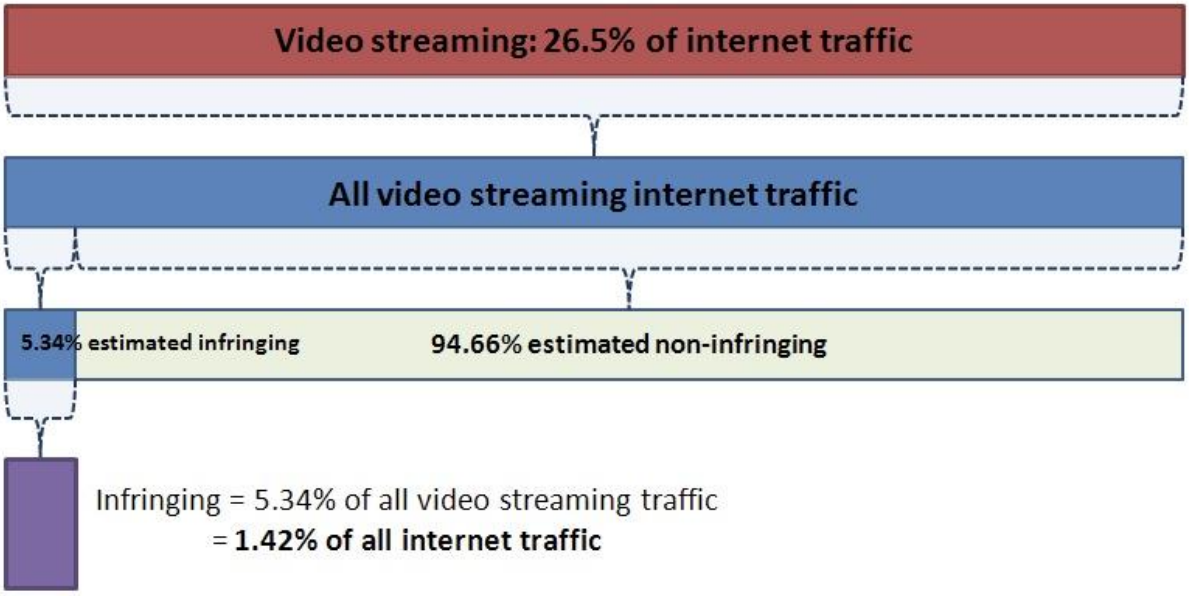
4.2.2 Cyberlockers

Cyberlockers are estimated to be responsible for 7% of all internet traffic. The estimations produced in Part A lead to a belief that around 73.15% of traffic to cyberlockers is related to infringing content. With an estimated overall internet bandwidth usage of 7% down to cyberlockers, this leads to an overall estimate for **infringing use of cyberlockers as 5.12% of all internet bandwidth**.



4.2.3 Video streaming

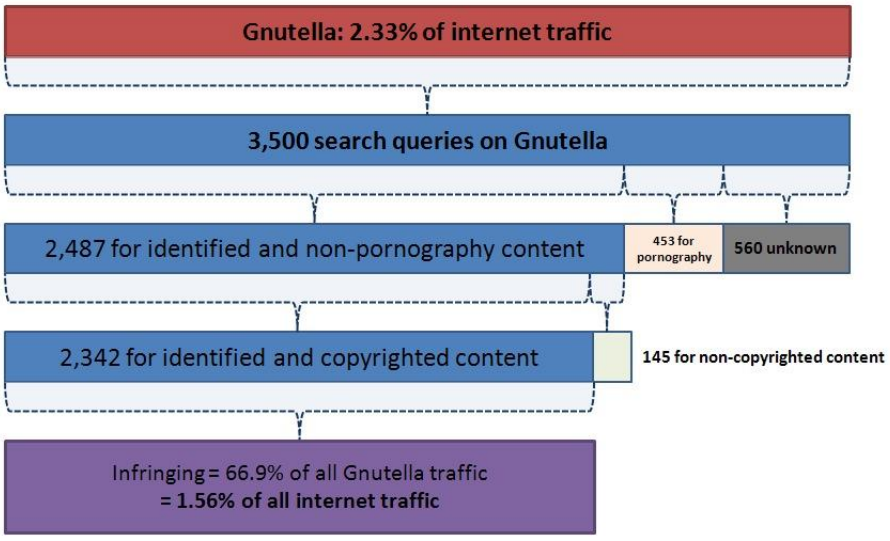
As Part A showed, the largest proportion of video streaming usage is legitimate and non-infringing. The studies discussed in Part B also demonstrated that video streaming traffic is the fastest growing area of internet consumption and is already responsible for more than one-quarter of all internet usage. As such, despite only 5.34% of all video streaming traffic estimate as infringing, this still amounts to **1.42% of all internet traffic**.



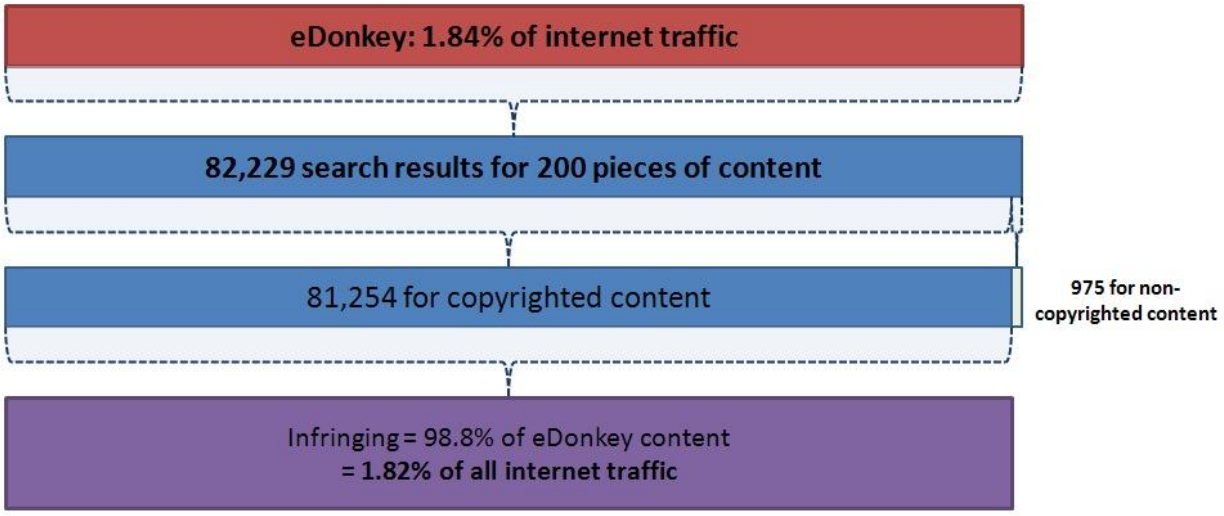
4.2.4 Other piracy arenas

Three other common piracy arenas were analysed in Part A: Gnutella, eDonkey, and Usenet.

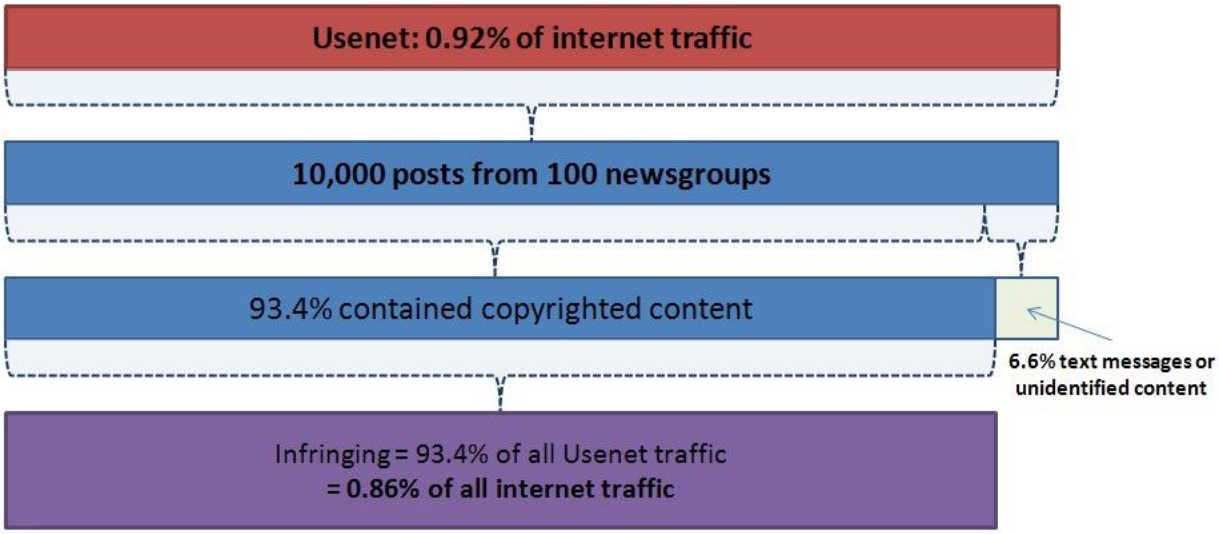
Gnutella is believed to be responsible for around 2.33% of internet traffic worldwide. With 66.9% of content searched for on the network estimated to be infringing and non-pornography, this leads to an estimate of **1.56% of internet traffic contributed by infringing content on Gnutella**.



eDonkey is heavily used in continental Europe, though it has declined in usage over the last two to three years after a series of successful anti piracy actions. The estimate in Part A is that 98.8% of eDonkey content is infringing. With 1.84% of internet traffic believed to be eDonkey, this would mean that **1.82% of all internet traffic is infringing content on eDonkey**.



Part A estimated the proportion of **Usenet** content that was infringing at 93.4%. The best estimate possible from the four studies in Part B found that Usenet made up 0.92% of all internet traffic. This would produce an overall estimate for the amount of infringing internet traffic from Usenet of 0.86%.



piracy.intelligence@envisional.com

Other P2P or file sharing networks also exist which are not explicitly covered within this research, such as Ares, DirectConnect, Kad (a sister network to eDonkey), Gnutella2 (used by clients like Shareaza), and MP2P (used by Piolet and Blubster), for instance. The four monitoring studies lead to an overall estimate for peer to peer usage other than the networks already discussed above of **2.02%**.[43] It will be assumed that infringing use of these networks is similar to the average infringing use of the networks analysed here in more detail: 78.94%. This would lead to an estimate of **overall internet use contributed by infringing content on these networks of 1.6%**.

Other types of internet traffic may also be used for infringing purposes. For instance, unauthorised copyrighted content might flow across VPN traffic and some VPN services like Relakks in Sweden exist primarily to hide file sharers from detection. Infringing content might also be transferred across email or be downloaded from normal web sites or blogs, for instance – though this would usually be small pieces of content such as music files rather than anything larger. However, estimating the size of this infringing traffic is extremely difficult, though experience means that the amount is likely to be small relative to the overall amount of bandwidth estimated for each type of traffic. As such, infringing content in these other areas is discounted for the purposes of this report, though this should not be taken as an indication that they are not used for the purposes of infringement.

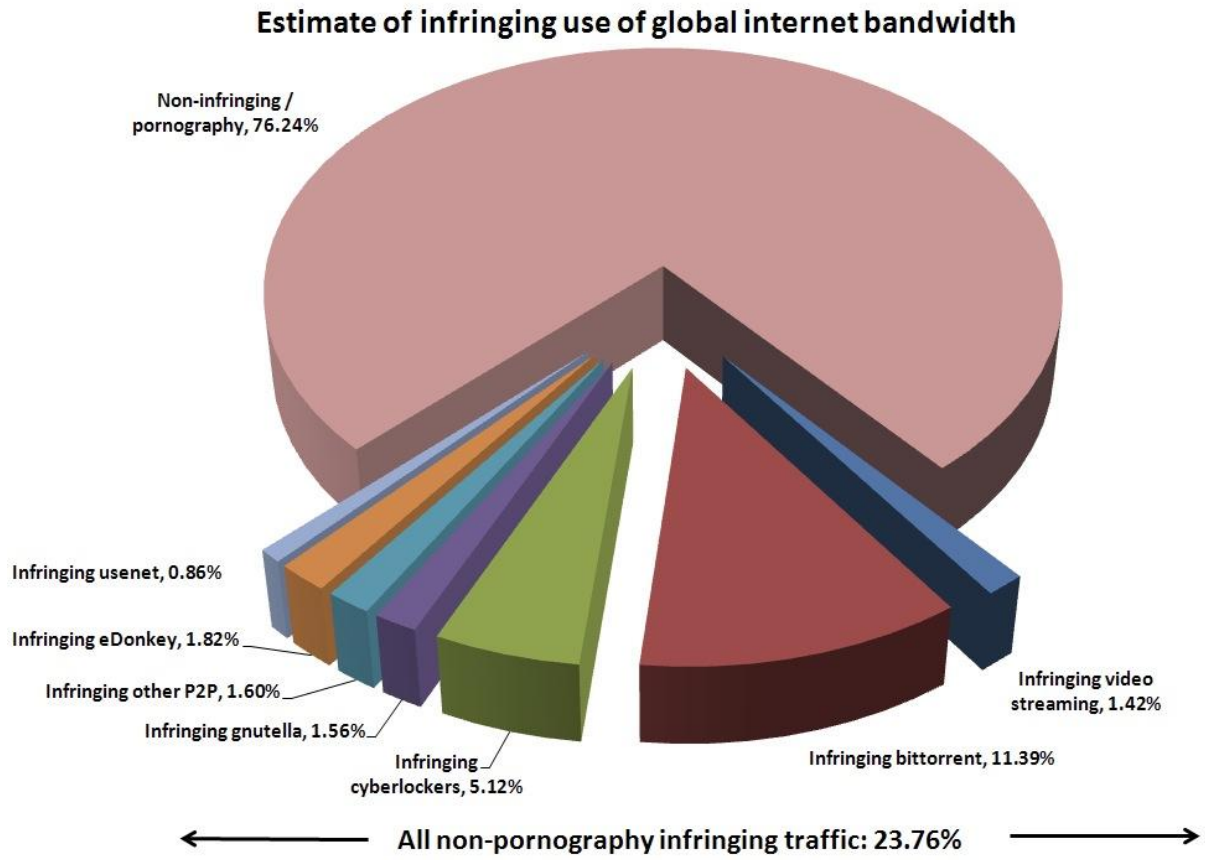
[43] A figure derived by taking the overall estimate for peer to peer traffic and subtracting the networks already identified (bittorrent, eDonkey, and Gnutella, for instance).

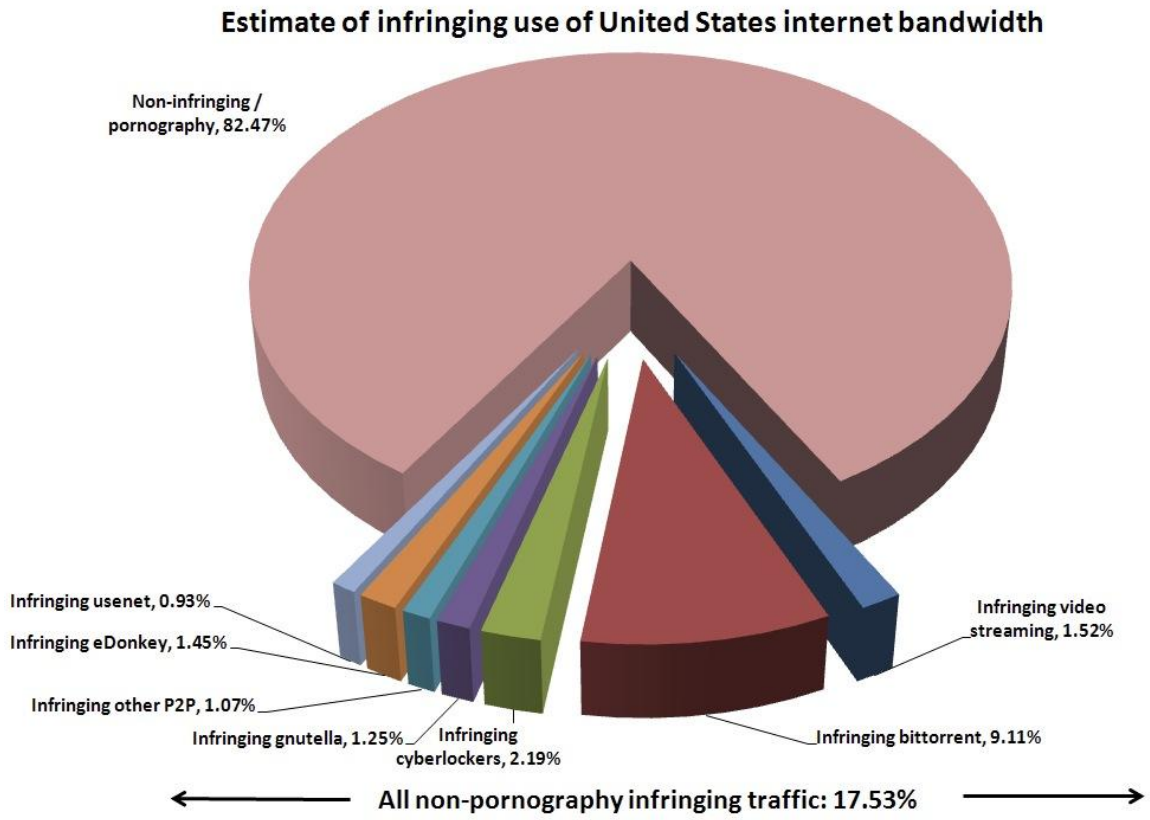
4.3 Summary

This report attempts to produce an estimate for the proportion of traffic which crosses internet that infringes copyright. Using studies of overall internet usage and analysis of common arenas through which content is transferred on the internet, the report finds that it is possible to calculate that a minimum of **23.76% of all internet bandwidth is devoted to the transfer of infringing and non-pornographic content**.

In the United States, the transfer of infringing and non-pornographic content is estimated to be responsible for a minimum of **17.53% of all internet bandwidth**.

The charts below show the overall estimate for the amount of global internet bandwidth which is believed to be infringing (and not pornography) and the overall estimate for the amount of United States internet bandwidth.





These estimates must, obviously, be issued with numerous caveats, both about the quality and accuracy of the data offered by the monitoring companies which estimate overall internet usage and about the ability to precisely quantify the proportion of infringing content on each arena of the internet. Methodological issues abound in both areas. Yet even given the limitations of the data available, Envisional believes that the estimates produced in this report are more accurate than any that have been published before. This report draws together the data in a way that allows, for the first time, the organisations which can help shape the ways in which users interact and obtain content to understand how much of the internet is devoted to the distribution and consumption of infringing material.

Piracy Intelligence

Envisional Ltd

piracy.intelligence@envisional.com