Copyright Violation on the Internet: Extent and Approaches to Detection and Deterrence

Submitted in partial fulfillment of the requirements for

the degree of

Doctor of Philosophy

in

Engineering and Public Policy

Alexandre M. Mateus

B.S., Computer Science and Engineering, Instituto Superior Técnico
M.S., Engineering Policy and Management of Technology, Instituto Superior Técnico
M.S., Engineering and Public Policy, Carnegie Mellon University

Carnegie Mellon University
Pittsburgh, PA

August 2011

Acknowledgments

I would like to acknowledge the financial support provided by Fundação para a Ciência e a Tecnologia (Portuguese Foundation for Science and Technology) through the Carnegie Mellon Portugal Program under grant SFRH/BD/27350/2006 and by the Information and Communication Technologies Institute (ICTI) project entitled "Web Security and Privacy: Weaving Together Technology Innovation with Human and Policy Consideration" (CMU-PT/SE/0028/2008).

I would firstly like to thank Professor Jon M. Peha, my advisor in this research, for taking me as his Ph.D. student, for all his support and advise, for his questions, for the countless hours dedicated to this research, and for his nearly infinite patience in guiding me to find my own questions and answers instead of just showing me the solution.

I would like to thank my co-advisor, Professor Pedro Ferreira, for his advise, for his questions and suggestions, and for his support of my research and his efforts over the last 7 years to make me a better researcher.

I would like to thank the members of my Ph.D. committee: Professor Jon M. Peha (chair), Professor Pedro Ferreira, Professor Lorrie Faith Cranor, Professor Michael D. Smith, Professor Francisco Lima, and Professor Michael J. Madison, for their interest in my research, for insightful discussions and for their questions and suggestions that contributed greatly to improve the quality of this research.

I would like to show my deepest appreciation to all the Professors in the Department of Engineering and Public Policy at Carnegie Mellon University who I had the chance to interact with, from whom I had the chance to learn, and who I consider myself lucky to have met.

Finally, I would like to show my gratitude to all my colleagues in the Ph.D. Program in Engineering and Public Policy. Without their company, support and friendship finishing the Ph.D. would have been a much harder task.

On a personal level, I am grateful to my family who has always done all they could to support my aspirations and help me reach my objectives. I thank my sister Sofia and brother Nuno for their love and patience, and my father Joaquim and mother Madalena, for their infinite ability to love and to give, and for having transmitted to me some of the values that always made and will always make me admire them.

Finally, I would like to thank my friends, for making work easier, for making sure that my life was not only work, and for filling my life with joyful moments. Without forgetting many others, I leave a special word of appreciation to Aleecia McDonald, Alexandre Ribeiro, Ana Cláudia Costa, Ana Mateus, Ângelo Mendes, Anu Narayanan, Bea Dias, Carolyn Denomme, Cristiano Casimiro, Inês Azevedo, Janice Tsai, Jennifer Logue, João Castro, João Graça, Maria João Rodrigues, Rodrigo Belo, Sandra Conduto, Sofia Taborda, Sónia Borges and Sónia Pedrosa.

Abstract

This research uses data collected from a university campus network via Deep Packet Inspection (DPI) monitoring and from the largest public BitTorrent tracker to characterize the extent of unauthorized transfers of copyrighted content using Peer-to-Peer (P2P) and to evaluate the effectiveness and limitations of DPI in detection of such activity, both to provide a perspective of how much copyright infringement happens using P2P and to inform those seeking to deploy DPI technology.

Use of P2P and transfers of copyrighted content were widespread on campus. In Spring 2008, 40% of students living on campus were detected using a P2P protocol, 70% of which were observed attempting to transfer copyrighted material. In late 2010, we estimate that over 800 million copies of content were transferred globally using BitTorrent per day, with an estimated number of transferred songs 13.1 times greater than worldwide sales of songs, and estimated number of transferred movies 6.8 times greater than worldwide box-office sales and 16.4 times greater than U.S. DVD and Blu-ray sales. Most transfers were from a small number of very popular titles that were widely available for sale. We found no evidence that use of P2P to transfer content without violating copyright was common both on campus and global BitTorrent. This indicates that copyright law is violated frequently using P2P, and while we cannot quantify how P2P transfers translate to lost sales, it is reasonable to assume some sales are lost due to P2P.

Focusing on effectiveness of DPI, after a couple weeks of monitoring DPI found up to 80% of detected P2P users attempting to transfer copyrighted content. In the short term, DPI could be effective to assess which network users transfer copyrighted content using P2P given some weeks of monitoring. However, limitations such as not being able to detect users of encrypted P2P can reduce DPI's effectiveness in the long term. Using behavioral classifiers that we implemented and that can detect encrypted BitTorrent from traffic summaries, we found students shifting from unencrypted to encrypted BitTorrent in the 2007-2008 academic year. If this trend continues, effectiveness of DPI for enforcement can be significantly hindered.

Table of Contents

List of Tables

List of Figures and Illustrations

"Ó rodas, ó engrenagens, r-r-r-r-r-r eterno!
Forte espasmo retido dos maquinismos em fúria!"[1]

Álvaro de Campos, Ode Triunfal

[1] *Oh wheels, oh gears, eternal r-r-r-r-r-r! // Strong withheld spasm of raging machinery!*

1 Introduction

The Internet is increasingly being used to obtain content, in particular audiovisual media (Cisco 2010). Peer-to-Peer (P2P) technology enables cost-effective distribution of content online by facilitating transfers of information between hosts (peers) that are part of a self-organizing overlay network supported by the IP network. At the same time, P2P raises significant issues in copyright protection and network management. P2P networks are used to transfer copyrighted content without permission from copyright holders, who claim that such activity has a heavy negative impact on their revenues (RIAA 2007). However, the actual dimension of copyright violations using P2P is far from being a settled matter, and there is still ongoing debate regarding how P2P affects the industries that produce and distribute copyrighted material (Oberholzer-Gee and Strumpf 2009). At the same time, P2P transfers add up to a significant share of all traffic exchanged on the Internet (Zhang 2009). Despite decreasing, the share of P2P traffic out of all Internet traffic was as much as 18% in 2009 (Singel 2009), and this can impose a heavy burden on networks and be problematic for Internet Service Providers (ISPs) managing congestion.

Deep Packet Inspection (DPI) network monitoring is one of the various technologies proposed over time to detect and prevent illegal file sharing using P2P, as well as to help network management deal with P2P traffic. DPI can both detect P2P traffic on the network and detect whether such traffic carries unauthorized copyrighted material. However, use of DPI for either copyright protection or network management is surrounded by controversy (US. Congress 2009), and to date, there is no thorough assessment of how DPI performs when used for detection of online copyright violations, and of the potential for circumventing DPI detection if it is deployed at a large scale.

This research aims to fill the above gaps by focusing in two main objectives. The first is to characterize the extent to which P2P is used to perform unauthorized transfers of copyrighted content. This objective motivates two studies, one focusing on P2P usage in universities and another focusing on the amount of copyrighted content transferred nowadays using BitTorrent. The second objective is to evaluate how well

existing network monitoring technologies, in particular Deep Packet Inspection (DPI), can detect P2P transfers and whether they carry copyrighted content, and what the implications are for online copyright enforcement. This second objective motivates one study focusing on evaluating the performance of DPI for detection of transfers of copyrighted content using P2P in a university campus.

The remainder of this dissertation is organized as follows. Chapter 2 presents the motivation for our two research objectives and describes each of them in more detail. Chapter 3 overviews current technologies used for online copyright infringement and the main policies that have been discussed or enacted in this area. Chapter 4 presents our first study, which characterizes the extent to which P2P is used to transfer copyrighted content in a college campus. Chapter 5 presents our second study, which evaluates how well DPI detects unauthorized transfers of copyrighted material via P2P. Chapter 6 is dedicated to our third study, which quantifies the overall amount of copyrighted content transferred using BitTorrent, the main P2P network in use nowadays. Finally, chapter 7 summarizes the main conclusions and policy implications extracted from the three studies, and chapter 8 leaves pointers for future work in the areas explored in this research.

2 Objectives

2.1 Characterize the Extent of P2P Usage for Transfers of Copyrighted Content

The first objective of this research is to characterize the extent to which P2P is used to perform unauthorized transfers of copyrighted content. Fulfilling this objective will inform policymakers about the magnitude of P2P and online copyright infringement and whether there is need for intervention. This objective is pursued in two separate studies: our first study (chapter 4), which traces the evolution of P2P activity and transfers of copyrighted content in a university campus; and our third study (chapter 6), which estimates the overall supply of content and number of copies transferred using BitTorrent, today's most popular P2P network.

The motivation to focus on a university campus in our first study comes from the fact that university students are considered to be among the main users of P2P for transfers of copyrighted content. Previous assessments of the extent of P2P and illegal transfers of copyrighted content using P2P on university campuses concluded that college students are among the biggest users of file sharing, with over half of college students engaging in P2P file sharing (Lamy, Duckworth, and Kennedy 2007) and accounting for 21% all P2P users (Guess 2008; Oster 2008). Assessments also revealed that a significant share of students' media libraries was composed of music obtained from P2P (U.S. Congress 2007a; L Smith et al. 2007) and that college students obtained more of their music from P2P than the rest of the population (Lamy, Duckworth, and Kennedy 2008). Such results have drawn attention to P2P in university campuses, as made clear by the fact that the U.S. Congress held at least six hearings on online copyright infringement in universities between 2003 and 2007 (U.S. Congress 2003; U.S. Congress 2004; U.S. Congress 2005; U.S. Congress 2006; U.S. Congress 2007a; U.S. Congress 2007b) and discussed possible interventions to deal with it (Bangeman 2007; Fischer 2007).

The fact that P2P on university campuses became a mainstream issue in legislative circles calls for a deeper analysis of the dimension of P2P in college campuses. Especially given that previous studies presented above were performed by means of surveys, which cause results to depend on the memories and openness of survey respondents, how survey instruments are designed, and how the subjects are selected. This is particularly relevant in this case, given that the subject in question constitutes illegal activity, and some respondents may refrain from disclosing their behavior in surveys. Our study presents results from a quantitative assessment of online media transfers based on actual observation of P2P exchanges on a college campus. Thus, not only are our results independent of whether or not survey respondents fully disclose their behavior, it is also possible to access information that Internet users may not know, such as the volume of P2P transfers or the time of such transfers.

In our first study, we trace the evolution of P2P activity and transfers of copyrighted content in a university campus over the course of three academic semesters to help assess the need for intervention, and observe the demographics of P2P users to understand whether possible interventions can be targeted using demographics. We characterize what content is transferred using P2P to understand users' preferences and how they change over time, which can inform content providers of what they are competing against, as well as to identify possible drivers for P2P usage. Finally, we compare use of P2P to legal distribution of content to contribute towards understanding what the impact of P2P is on sales of content and what factors make users decide for the legal or the illegal option.

BitTorrent has become the main P2P network in use nowadays (Bangeman 2008b), and by many accounts it is used mostly to transfer copyrighted content without permission of the copyright holder (Envisional 2011). Such unauthorized transfers of copyrighted content are responsible for billions of dollars in lost sales and thousands of lost jobs according to representatives of the music industry (U.S. Congress 2007a). There is a growing body of literature attempting to assess whether file sharing does indeed lead to a decline in sales of content, particularly in sales of music and video. However, as summarized in a recent working paper (Oberholzer-Gee and Strumpf 2009), different authors present contradictory results. Most papers focusing on this subject argue that P2P file sharing contributes to the

decrease in music and video sales (Hong 2007; Liebowitz 2008; Michel 2006; Rob and Waldfogel 2006; Zentner 2006), with displacement rates ranging between 3.5% (for movies) and 30% (for music), i.e., each music title downloaded through P2P displaces sales of 30% of a music title. But other articles argue that P2P has positive effects on music sales (Andersen and Frenz 2008; Gopal and Bhattacharjee 2006). Accounts of P2P file sharing in these articles are based on self-reported data collected by means of surveys or on use of proxies such as Internet penetration. A third set of authors, which use actual measures of file sharing, argue that P2P transfers are unrelated to changes in content sales (Bhattacharjee, Gopal, et al. 2007; Oberholzer-Gee and Strumpf 2007; MD Smith and Telang 2008; Tanaka 2004). Our research sheds some light into this question. In our first study, through observation of student usage of P2P and of the iTunes Store, we contribute empirical evidence about the extent to which P2P users still purchase content online. In our third study, we estimate the number of copies of copyrighted content transferred using BitTorrent, the first input for the sales displacement equation, which had not been well quantified before.

Previous studies that quantified how much content is available in P2P failed to estimate how much of that content is actually transferred by users (Envisional 2011; Layton and Watters 2010). Furthermore, such estimates were obtained based on the raw number of peers in swarms, which makes them imprecise because they fail to account for churn (failed peers) and thus grossly overestimate the amount of users downloading content. In our third study, the estimates of the number of unauthorized transfers of copyrighted content performed using BitTorrent are one step ahead of previous studies. We estimate both supply of content, i.e., how much content is made available in BitTorrent, as well as demand for that content, or the number of copies of content actually transferred. And we take into account churn caused by failed leechers, which allows us to reach more accurate estimates of the amount of content transferred in all swarms managed by today's most popular BitTorrent tracker.

In our third study, we estimate the number of transferred copies of copyrighted content and break down such estimates by type of content to compare them to content purchases from legal outlets. This puts our estimates in perspective, providing a better understanding of how serious an issue copyright violation

using P2P still is. We also assess which technical characteristics of content users prefer (different methods of video digitalization, video resolutions and audio bit rates). This can serve as an input for designing policies to deal with online copyright infringement since content formats and other technical characteristics can influence the performance of technical measures against illegal transfers using P2P. In particular because current technologies for detection of transfers of copyrighted content, like DPI, are more effective detecting specific types of content, or content with specific characteristics.

2.2 Evaluate the Performance of Deep Packet Inspection in Detection of P2P Transfers of Copyrighted Content

The second objective of this research is to assess how DPI performs in detecting P2P traffic and transfers of copyrighted content using P2P. Fulfilling this objective will inform policymaking on whether and how DPI can be useful for enforcement of online copyright. This objective is pursued in our second study (chapter 5), which uses data collected from a university campus network to assess various aspects DPI's detection of P2P activity and of transfers of copyrighted content using DPI.

DPI network monitoring technology has the ability to look at the content of information packets exchanged over the Internet and identify whether copyrighted content is being transferred in those packets. This makes it one of the technologies that can possibly be used for detection of unauthorized transfers of content online using P2P. However, use of DPI network monitoring to detect P2P transfers of copyrighted content faces some open opposition, with arguments ranging from privacy-invasion and freedom of speech concerns to more technical questions (US. Congress 2009). One strong argument against using or mandating solutions that rely on DPI for online copyright enforcement rests on the fact that there is no assessment of how effective DPI is in detecting the target behavior. In particular because the most popular P2P networks support traffic encryption, which can prevent DPI from detecting P2P traffic and from assessing whether content is copyrighted. Our second study evaluates the performance of DPI in

detection of P2P transfers of copyrighted content, taking the above issues into account, to provide input for the law and policymaking process.

Our second study assesses how effective DPI is in detection of P2P activity and transfers of copyrighted content performed using P2P and what is the impact of possible detection errors. Effectiveness is one of the main factors to take into account when determining whether or not to deploy a new enforcement technology, especially in this case, where detection errors may lead to unwarranted penalties to non-violating users. The study focuses also in comparing DPI's detection of transfers of audio and video, two of the types of content most transferred in P2P networks, and about which copyright holders have been most vocal. Such comparison is relevant because technical characteristics of video make it more difficult to detect than audio, and the popularity of video content in P2P networks has been growing in recent years. Finally, the study assesses the extent to which encryption is being adopted in P2P by comparing detection of P2P using different network monitoring technologies. Understanding the extent to which P2P encryption is being adopted is important to estimate how effective future deployments of DPI technology will likely be. Since DPI cannot detect transfers of copyrighted content performed using encrypted P2P, if a significant share of P2P users adopt P2P encryption, and thus go undetected, this can lead to a significant decrease of the effectiveness of DPI when used for copyright protection.

3 Overview of Technologies and Policies Pertaining to Online Copyright Infringement, Detection and Deterrence

The purpose of this section is to present a brief summary of the operation of the principal technologies used nowadays to distribute content online, of means of detection of copyrighted content made available online, of means of deterrence of illegal transfers of that content, and of counter-measures that Internet users can adopt to avoid being detected transferring copyrighted content. The section ends with a brief summary of policy and legal approaches to online copyright violations put in place in the last decade.

3.1 Technologies for Content Distribution

Three main types of technologies are used nowadays for distribution of copyrighted content without the permission of copyright holders: Web-based technologies, Usenet Newsgroups, and P2P technologies. Next, we describe how each of those technologies work, as well as the processes of making content available and obtaining content using each of them, and we present existing estimates of the amount of copyright content being transferred in each case.

3.1.1 Web-based Technologies

Web-based technologies distribute content either using bulk file transfer protocols that allow users to download the requested files for later consumption, or using streaming protocols that allow the users to consume the content in real time. The first approach is implemented by Direct Download Link (DDL) websites (also called cyber-lockers or one-click hosts), which let users upload content that then becomes available via a URL that the user shares with other users who will download the shared content. The second approach is implemented by Streaming websites, the most well known example of which is perhaps YouTube, where users upload videos that then become available for other users to watch right on their browsers. Both approaches can be used to share content that can be legally made available

online, and in many cases they are (e.g., YouTube hosts many user-generated videos that do not infringe on any copyright), but they are also used to host copyrighted content illegally and make it available to a wide range of users that obtain it without the permission of copyright holders.

3.1.1.1.1 DDL Hosts

DDL hosts are web-hosting platforms that allow users to upload files, which then become available for download via a URL (link) returned to the user who uploaded the content. Originally, such platforms were developed to overcome limitations in transferring large files over the Internet, in particular large email attachments since most email service providers limited the size of message attachments to a couple of MB. Using DDL hosts, a user uploads a file and then shares the content URL with the third parties with whom the file is to be shared, and those third parties download the file from the website.

The key to using DDL hosts for generalized spread of copyrighted content lies in the distribution of the content's URL to all the users possibly interested. This is achieved through forums and index sites where such URLs are posted and organized according to the type of content they point to, thus making it relatively easy for interested users to find links to the content they seek.

This type of service started gaining popularity in 2005, with offers from several startups, but only a few popular sites made it to today. The most widely known DDL hosts nowadays are *rapidshare.com* and *megaupload.com*, and both sites operate a tiered business model that gathers revenue from both advertisements and subscription fees. In the lowest tier, users do not need to register or pay a subscription fee in order to download a file, but the offered functionality is rather limited. Users have to wait several seconds before starting the download, they are bombarded with advertisements, and they are restricted in the number of simultaneous downloads, in the maximum transfer speed per download, in the total amount of downloads per day and in the total traffic exchanged per day, among others. Registering with the website alleviates some of these restrictions, and paying a subscription fee allows for further relaxation of the restrictions.

Since the hosting services are offered using regular websites, both the upload and download of files take place using traditional bulk file transfer protocols, such as HTTP or FTP. While it is fairly straightforward to implement end-to-end security on both protocols using SSL, to date none of the most popular DDL hosts supports access via encrypted connections. Top DDL hosts generate a large volume of traffic, which they load balance by redirecting users to addresses in a pool of IP addresses they operate.

3.1.1.1.2 Streaming Websites

In streaming websites, instead of downloading the content for later consumption, users watch the video or listen to the music directly on their browser. To that extent, streaming websites offer the simplest way for the user to obtain video or music content online, since no software other than a web browser is required, nor is any particular IT savvyness, and, apart from an initial buffering period, the user can start watching/listening to the content almost immediately after the load of the webpage.

Nowadays there is a large number of websites that stream content in a large scale, in particular video content. Examples of popular video streaming websites are *YouTube*, *Vimeo*, *DailyMotion*, *VuReel.com*, *DivxDen.com*, *MegaVideo.com*, *Hulu*, *Netflix* or *ABC.com*, among many others. These can be organized in two dimensions depending on who feeds the content to the website and on the amount of content that can be watched legally on the website, as portrayed in figure 1.

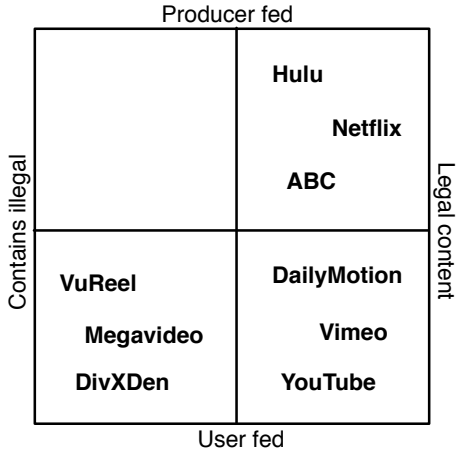


Figure 1. Classification of video streaming websites depending on who uploads the content and whether the uploaded content can legally be watched on the website.

Producer-fed websites are typically set up by the content producers themselves (ABC, for instance) or by companies that acquire the rights to stream the content online from producers (Netflix, for instance). Such websites typically realize revenue by means of advertisements included in the video, thus providing the user with a TV-like experience on their computers, or by means of subscription fees. User-fed websites are typically set up in a way that is similar to DDL hosts. Users upload content using the website and receive in return a URL to the webpage where that content can be viewed. In this case the sources of revenue are also advertisements and, in some cases, subscription fees.

Most user-fed streaming websites have terms of usage that explicitly state that illegal content or copyrighted content, without the permission of the copyright holder, should not be uploaded. However, users may eventually upload such content, which can lead to those websites making available content that infringes on copyrights. As we discuss in section 3.2, there are ways to proactively detect when copyrighted content is uploaded, and websites could devise other ways of holding users accountable for uploading infringing content (via authenticated identities, for instance), but few video sharing platforms take such measures. In figure 1, a particular website is on the lower left or on the lower right quadrant depending on the how much legal vs. illegal/copyrighted content is typically made available overall in the website. Streaming websites on the lower right quadrant are used mostly to share video authored by the users, which does not violate copyright. As for the websites on the lower left quadrant, many uploaded videos are copyrighted movies, TV shows, or other videos, made available without the authorization of copyright holders. In both cases, users can find the videos they want to watch via search or browsing of indexes, which organize and make available the URLs of web pages where the content can be viewed. Typically, websites on the lower right index their own content and the search occurs in the website itself. As for websites on the lower left, the indexing and search functionality is provided by separate index websites, which often organize content made available by several streaming websites (examples of such index websites are *surfthechannel.com* or *yidio.com*).

3.1.1.1.3 How much copyrighted content is transferred

Both in the case of DDL hosts and in the case of streaming websites, it is difficult to produce a good estimate of the amount of copyrighted content that is transferred without the permission of copyright holders.

Several measures show that traffic generated by DDL hosts amounts to a considerable percentage of overall Internet traffic. In their 2008-2009 Internet Study (Schulze and Mochalski 2009), iPoque reports that DDL hosts contributed to up to 10% of overall Internet traffic. But, looking at different geographical areas separately shows a wide range of usage of DDL, which goes from under 20% of all HTTP traffic in South Africa or in Eastern Europe, to over 40% of all HTTP traffic in South America and Southern Europe. As for specific DDL providers, in their 2007 report (Schulze and Mochalski 2007), iPoque found that *Rapidshare.com* and *MegaUpload.com* were the top players, generating respectively 55% and 17% of all DDL traffic. Other sources corroborate some of these findings. For instance, in a recent report (Labovitz, Iekel-Johnson, et al. 2009), Arbor Networks found that *MegaUpload.com* contributed significantly to overall web traffic, since its change of service provider made the new provider's share of overall Internet traffic increase more than ten times from under 0.05% to over 0.5% of overall Internet traffic. As for the amount of copyrighted content shared using DDL hosts without permission from copyright holders, there is very little empirical evidence. One recent study analyses a sample of 2000 files shared in DDL hosts and estimates that over 90% of those contain copyrighted content (Envisional 2011).

Concerning streaming video, the most recent statistics show that it accounts for a considerable amount of Internet usage. In terms of traffic, streaming video is the fastest growing category of Internet traffic (Labovitz et al. 2009). In terms of users, a recent survey by the Pew Internet & American Life Project shows an increase in the percentage of US adult Internet users who watched or downloaded video online from 57% in 2007 (Madden 2007) to 69% in 2010 (Purcell 2010). In the latter period, 32% of users reported watching movies or TV shows, but only 10% of users paid to watch or download video. This does not imply that the users who transferred movies or TV shows without paying were transferring

them without the authorization of copyright holders, in particular since there are significant free and legal offers of movies and TV shows online (e.g., Hulu.com or ABC.com). But, since websites that offer streaming of movies and TV shows illegally are easy to use, users don't have to wait for the video to completely download before starting to watch it, and that for most Internet users such websites are ambiguous as to whether their offer is legal or illegal, it is only fair to expect that a significant percentage of the users who watch video online are actually obtaining some of it from illegal sources.

3.1.2 Usenet Newsgroups

Another way of sharing content online is by posting it to Usenet newsgroups, which are supported by a set of distributed servers, each storing a copy of the content that is made available at each moment. Usenet newsgroups allow users to post messages under threads organized around particular subjects, which then become available to all other users. They preceded the web as a communication technology, with the RFC defining the format of exchanged messages dating back to 1983 (IETF 1983), and the RFC specifying the underlying communication protocol, the Network News Transfer Protocol (NNTP), dating back to 1986 (Kantor and Lapsley 1986). In order to post to and read messages from Usenet newsgroups, a user needs to have access to a news server, and to run client software that can connect to that server and download/upload news messages.

While Usenet was originally developed to distribute text messages, developments in data-packaging and encoding mechanisms soon made it possible to include binary content in Usenet posts. However, uploading and downloading binary content to Usenet is a complicated task that needs to be performed by specialized software. For upload, a large binary file needs to be compressed and broken down in a set of small-size archives. Such archives are then analyzed in order to generate a set of auxiliary files with information that will allow recovering the original content in case some of the archives cannot be downloaded from the newsgroup. Finally, each archive and each auxiliary file is uploaded individually to the newsgroup as a different post. For download, a user needs to obtain all the archives from the different posts in the newsgroup, or most archives and the auxiliary files needed to reconstruct the rest, and then

go through the inverse process of decompressing and re-joining the information in the archives in order to reconstruct the original file.

Usenet was a very popular means of communication before the Web became popular, but its popularity decreased after that. However, it managed to sustain a number of users that still allowed for newsgroup servers to be maintained by most ISPs until recent years. The fact that this was a somewhat underground community away from the spotlight of Internet hype allowed for most Usenet activity to remain under the radar of the masses for many years. In particular, Usenet has always been used to illegally share copyrighted content, but only in recent years, and in particular due to technological developments and new business models, did it become a major outlet for illegally distributed copyrighted media.

On the technological site, the main impediments to using Usenet broadly to transfer large binary files came essentially from difficulty in finding the desired content, which is divided among hundreds of news messages each containing only a small piece of the file. Traditional Usenet clients didn't offer much help in this task because they were tailored to allow users to browse through newsgroups and read recent news messages related to a set of topics of interest, not to find a set of messages containing the pieces of a large binary file. Such impediment was overcome by the appearance of online services (e.g., *newzbin.com*, *binsearch.info*) that continuously index the messages posted to Usenet and allow users to perform keyword searches which result in lists of references to the messages that contain the various pieces of a given binary file (called *collections*). Still, after finding the collection corresponding to the desired file, the user would be left to the cumbersome task of transferring each individual news message and later reconstructing the original file. This task was eased by the specification of a metadata file format (NZB file[2], developed by *newbin.com*) that aggregates information about all the messages in a collection, thus making it easy for software clients to transfer and reconstruct the content files from those separate messages. This file format was soon adopted by Usenet indexing services and by a new generation of news clients that specialize in fetching binary files from Usenet. Nowadays, downloading a large binary

[2] http://docs.newzbin.com/index.php/Newzbin:NZB_Specs

file from Usenet takes two easy steps: first the user visits an index website, searches for the content and downloads the respective NZB file, and then the user opens the NZB file using a compliant newsreader client, which will download all the necessary pieces and reconstruct the file[3].

On the business side, the main development happened as ISPs slowly started to phase-out their free Usenet services. Maintaining a Usenet server when a fair amount of the messages exchanged actually contain binary data, some of which belonging to copyrighted content that is posted without the permission from copyright holders, can represent both a considerable cost in storage and bandwidth as well as possible exposure to litigation from copyright holders. Perhaps partly due to these reasons, ISPs slowly started to phase out their free Usenet service, making room for a set of independent Usenet service providers with offers particularly appealing to users wanting to download binary content. Table 1 presents the main Usenet service offers nowadays in the market. It shows that, besides allowing multiple simultaneous, and possibly encrypted, connections[4] from each client, which altogether add up to high transfer speeds, such providers maintain large repositories of legacy content[5] posted to Usenet, allowing users to obtain that content for monthly fees ranging somewhere between $8 and $15 for the lowest tier of service (but going all the way up to $30 for upper tiers).

Table 1. Feature comparison of most popular metered Usenet news service providers[6].

	Lowest Price (per month)	Capacity (GB)	Price per GB/month	Retention (days)	Simultaneous connections	256-bit SSL
Newsguy	$8.33	50 GB	$0.17	365	32	Free
UseNeXT	$10.27	50 GB	$0.21	600	30	Free
100ProofNews	$8.95	40 GB	$0.22	170	20	Free
Newsgroupdirect	$13.58	60 GB	$0.23	200	20	Free
Newsdemon	$15.99	50 GB	$0.32	225	20	Extra
NNTPjunkie	$9.95	30 GB	$0.33	365	32	Free
Giganews	$12.99	35 GB	$0.37	428	20	Extra
Thundernews	$9.99	25 GB	$0.40	465	20	Free
Easynews	$14.97	30 GB	$0.50	170	20	Free
Newsgroups	$12.95	25 GB	$0.52	150	20	Free

[3] In fact, recent newsreader clients already allow searching indexes directly within the client and start the actual content downloads automatically, thus turning these two steps into a single step performed within only one application.

[4] The more simultaneous connections the more pieces of a binary file the user can download in parallel, thus the higher aggregate download bandwidth the user will achieve.

[5] Server's retention rates mean the number of days of messages that the servers maintain. In this case, the servers with highest retention rates store over 1.5 years of Usenet messages and respective content.

[6] Adapted from http://www.newsadmin.com/newsservers_compare_metered.asp on June 30, 2010.

Hence, in order to obtain copyrighted media from Usenet newsgroups, a user needs to obtain the appropriate newsreader software with support for NZB metadata files, contract service from a suitable Usenet service provider and have knowledge of the index websites where to search for available content. While there are free news clients, clients that support SSL encryption and that can perform downloads using NZB files, expand the transferred archives, recover from errors using parity files, and reconstruct the original binary, are available for prices ranging from $20 to $35. To add to this fixed price, there will be the service subscription fee of $8 to $30. Hence, while transferring copyrighted content from Usenet newsgroups has become relatively easy in the latest years, and arguably safer due to transport encryption, it comes at a non-negligible price to end-users.

3.1.2.1.1 How much copyrighted content is being transferred

Computing accurate statistics on the amount of copyrighted content transferred using Usenet newsgroups is a difficult task due to the distributed architecture of Usenet: since there are many servers spread across the planet, in order to collect information on the number of times each particular file is downloaded from Usenet, it would be necessary to trace the number of times such file was downloaded from each of those servers. The best available information on Usenet newsgroup usage comes from a set of publicly available sources[7] that provide generic figures on number of posts and number of bytes uploaded and downloaded from a subset of Usenet servers. The main problem with such figures is that they come from only a subset of all available servers (without reporting which servers are in that subset) and that they offer raw counts of files uploaded and downloaded without considering factors such as spam messages, for instance, which are widespread in newsgroups.

Nevertheless, looking at the numbers provided by one of such sources (*www.newsadmin.com*) shows that over a day (June 26, 2010) there were close to 30 thousand binary files posted to the top 20

[7] Examples are www.newsadmin.com, www.usenetstats.com, news.anthologeek.net

newsgroups (in terms of posted files) typically used for transfers of large copyrighted binary files[8]. This, despite being a rough measure, shows that there is plenty of supply of content on Usenet, a fair share of which is likely copyrighted content posted illegally. As for the demand, numbers from the same source show that the top 20 newsgroups (in terms of unique accesses) had close to 80 thousand unique accesses in the same day. This number, while more dependent on the sample of servers used[9], shows that the high supply of binary content in newsgroups is matched with a high demand for those files. A recent study (Envisional 2011) sampled the last 100 files posted to each newsgroup in random sample of 100 newsgroups and found that over 90% of those files contained copyrighted content. Hence, there is much activity happening in Usenet, with most of the content made available in the newsgroups being copyrighted.

3.1.3 Peer-to-Peer (P2P) Technologies

In the content-sharing context, Peer-to-Peer (P2P) designates a set of protocols that allow end-hosts to interact with each other and form networks that are used to transfer files. Such networks are called Peer-to-Peer networks because the relationships between connecting hosts are horizontal, i.e., there is no defined client-server hierarchy, and each peer acts as a client at times and as a server at other times.

P2P networks have many uses, some of which are legal and others illegal. Use of P2P for distribution of content is advantageous for content providers because they don't need to provision all the bandwidth, storage or computing power necessary for all clients to access the content, since all users share the burden of transmission. As such, there are many legal content distribution services that use P2P as the underlying data transmission technology (e.g., Vuze, Pando networks). However, there are also many examples of use of P2P to transfer copyrighted content without the permission of copyright holders, in

[8] alt.binaries.classic.tv.shows, alt.binaries.comics, alt.binaries.dvd, alt.binaries.dvdrs.pw, alt.binaries.dvds, alt.binaries.e-book.technical, alt.binaries.games, alt.binaries.games.wii, alt.binaries.hdtv, alt.binaries.hdtv.x264, alt.binaries.movies.divx, alt.binaries.mp3, alt.binaries.multimedia, alt.binaries.multimedia.anime.highspeed, alt.binaries.psp, alt.binaries.sounds.mp3.complete_cd, alt.binaries.teevee, alt.binaries.tv.big-brother, alt.binaries.tv.deutsch, alt.binaries.x264

[9] The number of files posted to a subset of servers eventually converges to the number of files posted in the entire Usenet as posts are flooded throughout the network of servers. In the case of unique accesses, each server can report only its number of accesses.

particular in today's general-purpose P2P networks. The first P2P network used for content sharing was Napster, which debuted in 1999 and soon grew to accommodate 26 million users worldwide (Jupiter Media Metrix 2001). Napster was shut down in 2001 as the result of a lawsuit over illegal sharing of copyrighted content, but by that time a number of other P2P networks already existed and attracted an increasing number of users. Nowadays, the leading P2P networks in terms of number of users and amount of exchanged content are indisputably BitTorrent and Gnutella, followed from afar by other less popular networks, such as eDonkey, Ares or DirectConnect (Bangeman 2008b).

In order to transfer content using a P2P network, a user needs to take three main steps: connect to the network; search for the peers sharing the desired content; and actually obtain the content from those peers. The main differences between P2P networks lie on the order by which these steps occur and on the particular way each of them is implemented. Next, we detail the process of making content available and of downloading content from the two top P2P networks in use nowadays: BitTorrent and Gnutella.

In BitTorrent[10], indexing of available content is done outside the network, using websites that gather collections of metadata files (torrent files), each describing how a particular set of files can be transferred and containing the addresses of trackers that manage the sharing of the respective content. A tracker is a facilitator of BitTorrent transfers, which maintains a list of peers sharing the content at a given time, and when contacted by a peer that wishes to transfer the content, returns a list of other peers that that peers can connect to in order to obtain the content. Hence, the first step a user has to take in order to transfer content using BitTorrent is to search for the desired content using an index website. After obtaining the torrent file from an index website, the user opens it with a BitTorrent client who contacts the trackers listed in the file, obtains the addresses of other peers sharing the desired content, and then contacts those peers to obtain the content. Each file is typically divided into small pieces that are obtained individually from different peers and then re-joined in order to reconstruct the original file. This allows for a peer to concurrently obtain multiple parts of each file and thus achieve high overall transfer speeds even if each

[10] We mean to provide a brief, simple and general description of how the BitTorrent protocol operates. For more specific information, as well as recent developments in the protocol, please consult http://wiki.theory.org/BitTorrentSpecification

individual piece is transferred at a low rate. In BitTorrent, as soon as the peer obtains a complete part of a file, that part is made available for sharing in the network, making it so that peers typically upload content at the same time as they are downloading it.

In Gnutella[11], search for content happens inside the network. First the user needs to join the network by having his Gnutella client connect to a peer that is already participating in the network[12]. Once connected to the network, a peer only knows the addresses of a set of neighbor peers. Searches for content happen by flooding a query throughout the network, starting by querying the peer's neighbors, who forward the query to their neighbors and so on. Once the query reaches a node that has the desired content, that node replies to the original peer signaling that the content was found. From this point, the requesting peer contacts the peer that has the content in order to initiate the transfer. Recent versions of the Gnutella protocol allow for a peer to obtain different parts of a file from different peers, similarly to what happens in BitTorrent. Many of the design decisions that lead to the architecture of the Gnutella network were motivated by the objective to make the network as decentralized as possible, so that it would not be easily taken down (as happened to Napster).

The process of uploading original content is also different depending on the network. In the case of Gnutella (and most other P2P networks) the user only needs to indicate, in the P2P client, which of the content available locally is to become shared in the network. The P2P client indexes that content and starts responding positively to queries that match any of the locally shared files. As a result, the pool of content available in the network is the union of the pools of content that each user is sharing.

In BitTorrent, the client software provides some support for uploading original content, but the user needs to take a set of explicit steps in order to share content. First, the user first needs to create the torrent file containing the metadata relevant for the transfer, that is, at least a description of how the content being

[11] We mean to provide a brief, simple and general description of how the Gnutella protocol operates. For more specific information, as well as recent developments in the protocol, please consult http://wiki.limewire.org/index.php?title=GDF

[12] The address of such peers can be found in different ways, for instance by consulting lists of peers that are typically online, by using lists from previous sessions, etc.

shared is divided into pieces and an enumeration of the trackers that will control the swarm[13]. Most BitTorrent clients can generate torrents automatically given the original files to share and the list of desired trackers. After having the torrent file, it is necessary to announce to the trackers that the content is available for sharing, which is another task that BitTorrent clients also perform automatically. At this point, all the client needs to do is start seeding the file (i.e., inform the tracker that it contains a full copy of the file and that it is ready to share it with other peers) and file is in fact being shared online. However, other peers need to be informed that the file is available for sharing if they are to express interest in obtaining it. Hence, a final step is to announce that the content was made available and provide other possibly interested users with the torrent file so they can join the swarm. This is typically done by posting the torrent file, along with information describing the content being shared, to index websites, who collect, organize and allow searching for information about content currently being made available via BitTorrent (e.g., thepiratebay.net, mininova.com, btjunkie.com, etc.).

Hence, while Gnutella is a self-contained decentralized network that relies only on the connected peers to function, BitTorrent needs infrastructure other than that offered by the peers, in particular, it needs trackers to manage the transfers[14] and index websites or some other form of spreading the information about the content that is available. While the need for centralized infrastructure such as trackers and index websites represents a vulnerability of BitTorrent when compared to Gnutella, it also makes it easier for users to find content and more efficient for transfers to actually take place (in particular for large files). That is perhaps one of the main reasons why BitTorrent is nowadays the undisputed leader in P2P networks. And while over the last years there have been several attempts to take down BitTorrent by taking down these central points of failure, the BitTorrent ecosystem has shown a remarkable ability to overcome such attempts, either by replication of the centralized resources[15] or by further developments

[13] A swarm is the term used in BitTorrent to refer to the set of peers connected to each other for the purpose of sharing a particular instance of content.

[14] There are new forms of operation of the BitTorrent protocol that no longer rely on trackers. While these are gaining popularity, most of BitTorrent transfers nowadays still go through trackers.

[15] There are at the moment a large number of open trackers providing service to all BitTorrent users (examples include OpenBitTorrent, PublicBT, myTorrentTracker, BitTrk and TheHashDen), a large number of closed invitation-only trackers that focus on specific types of content, and a large number of index websites that organize and allow searching for content made available,

that make those central resources less fundamental for the operation of the network[16]. Finally, on the client software side, P2P users have ample choice of both BitTorrent and Gnutella clients for Windows, Mac OS and Linux operating systems.

3.1.3.1.1 How much copyrighted content is being transferred

There are no precise figures on how much copyrighted content is transferred overall using P2P. Some of the available figures come from assessments of the extent of P2P usage to illegally transfer copyrighted content on university campuses. By means of surveys, such studies found that over half of college students engaged in P2P file sharing (Lamy, Duckworth, and Kennedy 2007) and that college students amounted to 21% all P2P users (Guess 2008; Oster 2008). Our own research, presented in chapter 4, uses data from actual network monitoring on a U.S. university campus and finds that, in April 2008, close to 40% of college students used P2P and 70% of those transferred copyrighted material at some point in a one-month period. Such percentages were actually down from April 2007, when 51% of students were detected using P2P and 81% of those transferred copyrighted content via P2P.

The above figures look at the activity and habits of college students, and while they cannot be generalized to the entire population, they do show that, at least in this environment, there is a considerable amount of transfers of copyrighted content using P2P. As for statistics for the general population, most existing reports were found to rely on inaccurate assumptions (N Anderson 2009) or to simply perform incorrect calculations (Oster 2008). This does not mean that P2P usage is not high in the population at large. For instance, *thepiratebay.com*, the largest BitTorrent tracker at the time, reported reaching 25 million users in late 2008[17], and these were only the users of one tracker in one particular P2P network.

---

both open websites that serve the public at large (examples include thepiratebay.org, btmon.com, mononova.org, fulldls.com, btjunkie.com, torrents.net and vertor.com) or closed servers paired with closed trackers that serve each specific sharing community. On top of index websites, there are meta-search services that index the content made available in multiple index websites, such as torrentz.com, which provides meta-search in over 30 index sites.

[16] Such developments include the use of Dynamic Hash Tables (DHT) to substitute the function of trackers, or the use of magnet links to substitute the function of torrente files.

[17] http://thepiratebay.org/blog/138 (November 15, 2008)

One proxy for how much content is transferred using P2P is the volume of traffic that is generated, with the obvious caveat that not all P2P traffic corresponds to illegal transfers of copyrighted content, as there is legal material circulating in P2P networks. Accounts of the P2P traffic are also surrounded by a great deal of uncertainty. First, because the amount of P2P traffic exchanged varies depending on the network location where the measurement takes place; different ISPs, countries, cities or neighborhoods, will most likely show different amounts of P2P traffic. And also because the most popular P2P protocols allow users to encrypt traffic, and detection of encrypted P2P traffic is all but trivial.

Concerning the amount of exchanged P2P traffic, a survey of available literature on traffic monitoring and measurement shows a wide range of estimates in the period between 2002 and 2006, but with an increasing trend in the overall percentage of P2P traffic, going from 21.5% in 2002 to 35% – 93% in 2006 (Zhang 2009). More recent measurements show that the overall percentage of P2P traffic started decreasing. According to Cisco's Visual Networking Index, the percentage of P2P traffic fell from 60% in the end of 2006 (Cisco 2008) to 39% in the end of 2009 (Cisco 2009), and is predicted to drop to 17% of Internet traffic by 2014 (Cisco 2010). This drop in percentage happens despite the fact that the total amount of bytes exchanged using P2P continues to increase, but at a slower rate than the increase in video traffic, the latest killer-app on the Internet. Other studies, while reporting slightly different figures for the total percentage of P2P, also report this decreasing trend in latest years (Labovitz et al. 2009).

Finally, recent studies looked at swarms managed by public BitTorrent trackers and attempted to estimate how much copyrighted content was transferred. One of such studies estimates that close to two thirds of BitTorrent traffic is "non-pornographic copyrighted content shared illegitimately" (Envisional 2011). Our own research, presented in chapter 6, estimates that an average of 800 million copies of content are transferred using BitTorrent per day, and finds only 0.02% of those copies to be titles that can be shared legally in P2P.

Hence, while it is hard to find precise figures, there is clearly a great amount of copyrighted content transferred using P2P. And while P2P traffic is decreasing as a percentage of Internet traffic, it is still

increasing in terms of exchanged bytes. This shows that, while P2P may step down as the principal technology for exchanging copyrighted content in the future, it is still the main technology for that purpose nowadays and is likely to remain a significant player for years to come.

3.2 Methods of Detection of Transfers of Copyrighted Content

Copyrighted content being shared illegally online can be detected in three principal ways. It can either be detected at the time it's made available, via preemptive detection (and filtering), it can be discovered after it has been made available, via Internet search that finds the content in the platforms where it is made available, or it can be detected at the time it is being transferred by users via network monitoring.

Preemptive filtering works by detecting whether content is copyrighted when it is first uploaded, and acting upon it in case it cannot be legally made available online. This type of detection can only be performed by the operator of the infrastructure to which the content is being uploaded, which makes this method particularly suitable for use with web-based distribution technologies In fact, the best examples of usage of this method come from web-based distributors, such as YouTube (King 2007) and other video streaming websites (Gannes 2009), which detect whether content is copyrighted at the time of upload, and try to preempt copyright infringement. In the case of Usenet newsgroups, preemptive filtering could also be achieved, but the fragmented nature of uploads would make it harder to implement – each file is broken down into tens or hundreds of small compressed pieces uploaded independently, possibly to different servers in the distributed infrastructure. Perhaps due to the inherent difficulty of implementing such an approach, there are no examples of preemptive filtering in Usenet newsgroups. In the case of P2P, since there is really no central entity or entities to which the content is first uploaded, preemptive filtering cannot be implemented on the content itself. However, in the case of BitTorrent, it would be possible to perform preemptive filtering in the centralized parts of the distribution process, such as preventing torrents for copyrighted content from being made available by index websites.

Content identification is typically performed in an automated fashion so that preemptive filtering can deal with high volume of uploads. Technology for identification of copyrighted audio and video is nowadays offered by various companies (Gannes, 2009). Identification happens by matching unique features of each uploaded file to a database containing features of known copyrighted titles, which are typically made available by copyright holders that have interest in acting upon the content they own that is uploaded to such websites by third parties. In the case of YouTube[18], for instance, copyright holders provide information that allows the identification of each of their titles together with the specification of the policy to follow in case uploads of that title are detected. Such policies are specified in a computer-readable meta-language to allow a fully automated process, and can go from requiring that the content is not made available at all, to more lenient approaches that allow the copyright holder to realize revenue from advertisements associated with the uploaded title.

Another type of detection is implemented by searching the Internet for copyright protected content and whose transfer constitutes copyright violation. Such searches typically focus on index websites, either of URLs pointing to content made available on DDL or streaming websites, for torrent files providing access to BitTorrent swarms, or for nzb files indexing content available on Usenet. The fact that index websites are built for the specific purpose of making it easy to find content made available online makes them particularly easy to search for the purpose of detection of copyrighted content. Pointers (URLs, *torrent* files or *nzb* files) are found by searching for the name of the desired content, and to make sure such pointers actually lead to the copyrighted material they advertise, part of the content is often transferred to confirm that it is actually part of the advertised media title. This type of detection allows for discovering which outlets make copyrighted content or pointers to copyrighted content available, but it cannot unravel which Internet users transferred such content. As such, it can only be used to act upon those outlets, typically via the mechanisms available in the law that allow copyright holders to request the removal of infringing content.

[18] YouTube's content identification system actually gives copyright holders multiple options when their content is detected being uploaded. Among others, copyright holders may have the content removed, promoted, or even try to monetize it via advertisements. More information on YouTube's Content ID system can be found at http://www.youtube.com/t/contentid

A final method uses network monitoring, which can detect transfers of copyrighted content as the content is being moved through the network. It can be performed in a passive fashion, by observing the traffic that flows through IP networks, or by actively participating in file sharing networks.

Passive network monitoring works by introducing a monitoring device in the network at a point where the traffic carrying copyrighted content can be observed. The monitoring device analyzes traffic using deep packet inspection (DPI) in order to sort out media transfers from other types of Internet traffic, after which it verifies whether such media transfers contain copyrighted content or not. When the media is found to be copyrighted, the monitoring device can collect the IP addresses of hosts involved in the transfer. To identify whether transferred media belongs to copyrighted material, the monitoring device applies content identification technology (similar to the one used in preemptive filtering) to pieces of the content gathered as they are being transferred. Due to storage and speed limitations of monitoring devices, and in particular when a high volume of traffic is being monitored, the pieces of content collected are typically a small fraction of the full title being transferred, which can impact the success rate of content identification (for instance, the collected piece of content may correspond to the use of a fraction of a movie to make a review of that movie, which is clearly a case of fair use, but that the identification process could not know because it didn't have visibility of the full length of the media file). The amount of copyrighted content detected using this technology depends on the visibility that the monitoring device has of the monitored network. In particular, monitoring devices need to be placed at point in the network where traffic carrying copyrighted content is likely to go through (such as the aggregation points of sub networks or connection points between private networks and the Internet). This means that the entity conducting network monitoring needs to either control the network or needs to obtain authorization from the network operator to deploy the monitoring devices.

Active infiltration of P2P networks, or Swarm Infiltration (SI), consists of using a modified P2P client that connects to the P2P network as a regular peer, but whose purpose is to collect information about other peers sharing particular copyrighted titles. This technique exploits the fact that, in order to function, P2P networks need to make the information about the peers sharing content at any given time accessible to

other peers. Hence, this type of monitoring can be conducted by anyone who can connect to the P2P network, since there is no need for privileged access to specific points of the underlying IP network. SI has been used mostly by copyright owners as a means of collecting information about P2P users illegally transferring their content.

Both passive and active network monitoring produce information about the IP addresses of users detecting transferring copyrighted content. Detected IP address and the time at which the detection occurred can be then used to identify the user responsible for the activity, a task that needs the cooperation of the user's ISP. Such information can then be used to take action against detected users.

3.3 Means of Deterrence of Illegal Transfers of Copyrighted Content

Several methods can be used to try to deter Internet users from obtaining copyrighted content for free in an illegal fashion. We overview the main methods that involve a technological approach, either by preventing transfers from taking place (Traffic blocking or IP address blocking), or by attempting to discourage users from engaging in further transfers (poisoning and malware injection).

Traffic disruption aims to prevent a user from completing the transfer of a title identified as copyrighted by terminating the connection between the hosts transferring the file. Disruption happens at some point in the route between the two computers involved in the transfer and can only be performed by someone who has access the stream of packets flowing between the two ends. It is typically paired with detection via network monitoring. Actual disruption can happen in different ways. One way is to drop subsequent packets from TCP connections once they are identified as transporting copyrighted content. This technique requires privileged access to the network such that there is the ability to drop packets at that point. Furthermore, since packets are effectively lost en route to the destination, this will cause the sender to repeat transmission of those packets, which may lead to a considerable amount of traffic between the sender and the disruption point until the sender times out and abandons the connection. Another possible way of disrupting traffic is to direct forged TCP RST (reset) packets to one (or both) ends of the

connection in order to make the end hosts terminate the connection. This requires less privileged access to the network, since it is only necessary to observe packets from the actual data flow in order to convincingly forge one of them.

IP address or hostname blocking prevents users in a sub network from accessing certain locations on the Internet. This is typically achieved using a firewall deployed at the point where the sub network connects to the Internet. Such a technique is only effective if the IP addresses or hostnames that are to be blocked are well known and don't change often, which makes it appropriate in the case of Web-based distribution or even Usenet newsgroups, but almost useless in the case of P2P (except for preventing users from obtaining torrent files, which are typically downloaded from centralized servers).

Malware injection or poisoning attempt to discourage users looking to obtain copyrighted content online by introducing forged content in the distribution channels. Such injection can have the goal of making it more difficult for users to obtain the content they seek, or it can actually look to distribute different content (typically malware). In both cases, users are discouraged because they end up transferring content that looks like what they are seeking but that ends up not being it, which takes them back to the starting point of looking for sources for the content that they seek once again.

3.4 Counter-measures Against Detection

Users seeking to transfer copyrighted content online without being detected can take advantage of different counter-measures available nowadays, which are directed mostly to prevent detection via network monitoring, both passive and active. The two main counter-measures in use nowadays are protocol and content obfuscation and use of Virtual Private Network (VPN) services.

Protocol and Content obfuscation counter-measures allow users to conceal the protocols they are using and the content that they are transferring from passive network monitoring. Such measures can be general and apply to any type of Internet traffic, as in the case of use of general-purpose SSL encryption, or can be specific to certain protocols, as in the case of BitTorrent encryption. In both cases, the payload

of packets is encrypted, which prevents DPI monitoring from unraveling the content transferred therein, and thus from detecting which protocol is being used to exchange the packets and from extracting pieces of content being transferred in order to detect whether it belongs to copyrighted titles or not.

General purpose SSL encryption is implemented as a layer on top of communications and can be used with almost any protocol. In particular, it can be applied to transfers from web-based content distribution services, to transfers from Usenet newsgroups and to P2P transfers. Most of today's Usenet servers support this type of encryption in transfers of content, and it is also the type of encryption implemented in the Gnutella P2P network. In both cases, it's up to the user to choose whether to activate encryption or not. In the case of BitTorrent, encryption is an extension to the protocol itself and users can choose to activate it for all communications or only for communications started by their end (in this last case, BitTorrent clients still reply to unencrypted requests, and thus may still exchange content using unencrypted sessions detectable by network monitoring).

Virtual Private Network (VPN) services allow users to establish encrypted tunnels to providers' servers through which all their traffic is sent and received. The provider's servers relay all the user's traffic to/from the final destination, achieving two main effects. First, since all traffic is tunneled and encrypted, it cannot be detected via passive network monitoring at any point in the path between the user and the VPN provider's server. And second, the other hosts the user is communicating with only see the IP address of the VPN provider's server. In particular, this defeats active network monitoring via Swarm Infiltration since the peers in the P2P network only see the external IP address of the VPN server and not the IP address of the actual user. General-purpose VPN providers have existed for a long time, but both due to lawsuits by the music industry and due to new legislation requiring identification of Internet users found transferring copyrighted content (in France and the U.K., for instance), in recent years offers started to target P2P users and gaining popularity (Sjoden 2009; Cheng 2009; Van der Sar 2009a).

3.5 Policy and Legal Approaches to Online Copyright Violations

Transferring copyrighted content using the Internet without the authorization of copyright holders is illegal in most countries in the world, save specific exceptions. In the last decade there has been significant activity both in terms of lawsuits against Internet users illegally transferring copyrighted content online and against companies whose products facilitated such transfers, as well as in legislative terms both in the U.S. and abroad. This section presents a brief summary of that activity.

By U.S. Copyright law (17 U.S.C.), except where "fair use" provisions apply (17 U.S.C. §107), transfers of copyright-protected works without permission from the copyright holder are infringements of the holder's rights (17 U.S.C. §106, §501). Both those who transfer the copyright-protected works and those who aid and support such transfers can be held liable for copyright infringement. This means that both P2P users and P2P developers may be accused of copyright infringement. Concerning ISPs, the Digital Millennium Copyright Act (DMCA) (105th Congress 1998) has provisions limiting ISP liability under certain circumstances, but to obtain such "safe harbor" protection, ISPs must respond to subpoenas and identify subscribers accused of a violation (17 U.S.C. §512).

This legislative framework has allowed copyright holders to pursue legal action against those infringing their rights using the Internet. In the past, the music industry, through the Record Industry Association of America (RIAA), used these legal provisions in several lawsuits against P2P companies (Macavinta 1999; Oswald 2006) and users (EFF 2008). To unveil the identity of users, RIAA traditionally used the subpoena mechanism in DMCA. When users were university students, since early 2007, the music industry started utilizing "pre-litigation settlement letters" requesting that infringing students be identified and that the letter be forwarded to them (Buskirk 2007). Since these letters were not legally binding, some universities ignored them, while others forwarded them to students (EFF 2008). Upon receipt of the letters, students could avoid court action and settle the case using the phone or a website[19]. The industry ended their

[19] https://www.p2plawsuits.com

lawsuit campaign against P2P users in the end of 2008 (McBride and E Smith 2008) announcing plans to start collaborating with ISPs for copyright protection. More recently, law firms, both in the USA and in the UK, started to use the subpoena mechanism in DMCA to reveal the identity of thousands of users sharing independent or adult movies. After identifying the users behind each IP address, the law firms demanded that each person paid a settlement fee of a couple thousand dollars via a website[20] to drop the lawsuit (N Anderson 2010b; N Anderson 2010a).

Illegal file sharing has also been debated at the legislative level, in the U.S., Europe, and elsewhere. In the U.S., Congress held at least six hearings on online copyright infringement in universities between 2003 and 2007 (U.S. Congress 2003; U.S. Congress 2004; U.S. Congress 2005; U.S. Congress 2006; U.S. Congress 2007a; U.S. Congress 2007b), discussing possible interventions to deal with it (Bangeman 2007; Fischer 2007), and enacted the Higher Education Oportunity Act (HEOA, 110th Congress 2008) that has provisions requiring universities to use technological approaches to deal with online copyright violations on campus networks. Outside of the U.S., the focus was mostly on ISPs, particularly in the European Union, where France (Parlement Français 2009) and the United Kingdom (U.K. Parliament 2010) approved legislation requiring that ISPs participate in identifying users accused of illegally transferring copyrighted content online, and that can eventually lead to revoking Internet access to repeated infringers. Besides legislative initiatives, in recent years, various countries including the U.S.A., Canada, the E.U. and Japan have been negotiating the Anti-Counterfeiting Trade Agreement (ACTA Negotiators 2010), an agreement that aims to establish international standards for enforcement of intellectual property rights, and that considers forms of dealing with illegal online file sharing such as requiring ISPs to identify Internet users accused of transferring copyrighted content.

[20] www.copyrightsettlement.info

4 P2P on Campus: Who, What and How Much

This chapter focuses on online exchanges of media in university campuses, with a focus on exchanges using P2P and the iTunes Store. It seeks to fulfill three main objectives. First, to quantify the extent of P2P usage and transfers of copyrighted content using P2P on campus, how these are changing over time and how they vary by the demographics of users, to help assess the need for intervention. Second, to characterize the content that students are obtaining from P2P, both to better understand the evolution of the P2P "market" for digital content, which can inform content providers of what they are competing against, as well as to identify possible drivers for P2P usage. Finally, to shed light on the impact of P2P transfers in revenues of copyright holders, which depends on the extent to which those transfers displace sales of content. This is done by quantifying the extent to which media is obtained from P2P and from the iTunes Store and YouTube, and by correlating usage of these sources.

The extent of P2P and illegal transfers of copyrighted content using P2P on university campuses has been previously assessed. By means of surveys, such studies found that over half of college students engaged in P2P file sharing (Lamy, Duckworth, and Kennedy 2007) and that college students amounted to 21% all P2P users (Guess 2008; Oster 2008). Moreover, these studies revealed that a significant share of students' media libraries was composed of music obtained from P2P (U.S. Congress 2007a; L Smith et al. 2007) and that college students obtained more of their music from P2P than the rest of the population (Lamy, Duckworth, and Kennedy 2008). Results from such survey-based studies depend on the memories and openness of survey respondents, how survey instruments are designed, and how the subjects are selected. This is particularly relevant in this case, given that the subject in question constitutes illegal activity, and some respondents may refrain from disclosing their behavior in surveys. This chapter presents results from a quantitative assessment of online media transfers based on actual observation of P2P exchanges on a college campus. Thus, not only are the results herein independent of whether or not survey respondents fully disclose their behavior, it is also possible to access information that Internet users may not know, such as the volume of P2P transfers or the time of such transfers.

In the remainder of this chapter we present an overview of our data collection methodology (section 4.1), followed by a summary of collected data and introduction of definitions used throughout the remainder of the chapter and in chapter 5 (section 4.2). Next we present obtained results, first drawing the general picture of P2P usage and how it is evolving on campus, followed by a breakdown of P2P figures by demographics, then focusing on the content being transferred using P2P, and finally on how usage of P2P to obtain media compares to usage the iTunes Store and YouTube (section 4.3). The chapter ends with a summary of findings and policy implications (section 4.4).

4.1 Data Collection Methodology

Research presented in this chapter and in chapter 5 was conducted using data collected in the scope of the Digital Citizen Project (DCP) at Illinois State University (ISU). The aim of the DCP was "to significantly impact illegal piracy of electronically received materials, using a comprehensive approach to confront pervasive attitudes and behaviors in peer-to-peer downloading of movies, music, and media" (Illinois State University 2008). In February 2007, a team engineers and social scientists from Carnegie Mellon University (CMU) began conducting research on the dissemination of copyrighted material on the ISU campus, which comprised the collection of network monitoring data. This section describes the methodology utilized for collection and anonymization of the network data used in this research.

4.1.1 Network Monitoring

Monitoring data was collected from the ISU campus network, which serves the entire campus population. The campus network connects to the Internet by means of two Internet Service Providers (ISPs) and Internet 2. In order to maintain appropriate levels of service for applications considered crucial to the university's everyday operation, ISU uses traffic shaping at the point where the campus network connects to ISPs. Shaping works by allowing specific classes of application to use up to a predefined percentage of the available bandwidth, but no limit is imposed on the amount of traffic generated by each network user.

The campus network is divided into several sub-networks. ResNet is the sub-network that students connect to in their dormitories. ResNet users purchase network access from ISU, which allows them to connect to one wired connection each in their dorm room. Each such connection has is assigned an IP address which is fixed for the entire semester. Students are not allowed to setup wireless 802.x networks in the residence halls, and this policy is enforced by the ISU network management staff.

Network monitoring was performed using three types of commercially available monitoring appliances: two were different types of deep packet inspection (DPI) appliances, and the third one consisted of Netflow collectors. The DPI appliances used were Packeteer PacketShaper[21] (Packeteer) and Audible Magic CopySense[22] (AM). Both devices log relevant attributes of transmissions between users inside the campus network and outside parties, for traffic routed using the commercial ISPs. Packeteer had already been deployed before the DCP project started, and was used for traffic shaping as described above. It classifies communication sessions according to the type of traffic that they carry in over 500 classes[23], without retaining any actual content of the communications sessions.

AM was deployed on the network to enforce ISU policy before CMU started collaborating with the DCP. It uses header information to identify P2P communication sessions, and then attempts to identify copyrighted media transferred in those sessions in real time. Identification of copyrighted media is performed by matching the content found in detected P2P sessions against a database of audio fingerprints of copyrighted media titles[24], or against hash codes[25] of files known to contain copyrighted content. The device records the copyrighted titles that were matched in the database without permanently

[21] Packeteer was since acquired by Blue Coat, for more information on the features of Packeteer PacketShapper (now Blue Coat Packetshapper), refer to http://www.bluecoat.com/products/packetshaper/

[22] For more information on the features of AM CopySense, refer to http://www.audiblemagic.com/products-services/copysense/

[23] Classes include, among others, common protocols, services, Peer2Peer networks and content distribution networks. A detailed list of the classes available in the Packeteer version used for data collection can be found in (Packeteer 2007).

[24] One technique used by AM to identify copyrighted material is audio fingerprinting. AM collects a sample of the audio track of the material that is being transferred and extracts relevant and unique characteristics of that audio (which are format- and encoding quality-independent). These are then compared against the database with the audio characteristics of known copyrighted titles.

[25] In most P2P networks, each file that is shared is identified using a unique hash code calculated based on the contents of the file. This short code (128 or 256 bytes) guarantees that the same file (i.e., the same content) is identified in the network independently of different filenames that it may have. The hash code is used by AM to identify copyrighted material because it allows for faster comparisons and earlier detection than the technique based on audio fingerprinting.

retaining any portion of the transmission. In cases when the content of P2P sessions cannot be matched against anything in the database, and when the sessions contain metadata describing the content being transferred (typically the name of the file being transferred), AM records such metadata.

AM logs information on communications in the form of *events*. An event corresponds to one or more consecutive TCP or UDP[26] sessions between a pair of peers in a P2P network. All the TCP or UDP sessions in an event are either identified as being associated with the same copyrighted media title, or cannot be associated with any media title in AM's database. Hence, an AM event means that two peers in a P2P network, one inside the ISU campus and another one outside, were detected by AM exchanging or attempting to exchange information (either belonging to an identified copyrighted media title, or information that could not be identified as belonging to any copyrighted media title present in AM's database) over a set of consecutive TCP sessions or consecutive UDP sessions.

4.1.2 Connection of Monitored Activity to Users and Devices

All monitoring appliances produced logs of network activities that contain an IP address in use by a device on campus. In the case of data collected by AM, identification of devices and users associated with such IP addresses was implemented using data from several network management databases also collected from the ISU network. For each data record, the identification of the device (i.e., the device's MAC address[27]) from the recorded IP address was performed via lookups in the DHCP[28] lease logs using the IP address that performed the monitored activity and the time when the activity occurred.

Users that performed each activity were identified by querying different network management databases, depending on the type of network connection used to perform the activity. Such queries returned the

[26] UDP sessions are actually pseudo-sessions, with consecutive UDP packets being aggregated in the same pseudo-session if they occur within a time interval that is lower than a predefined threshold.

[27] Media Access Control address, a 48-bit identifier that is (virtually) unique to every device that connects to an IP network.

[28] Dynamic Host Configuration Protocol, a protocol used by devices in a network to obtain a lease for a unique IP address and information about several other parameters necessary to connect to the network. IP addresses are assigned to requesting devices for a period of time and the lease information is typically stored in a log.

University Login Identification (ULID)[29] of the user that registered the device detected performing each online activity, thus it is assumed that such user was the one responsible for the activity. The ISU directory database provided the remaining information about the user associated with each ULID, namely the user's birth year, gender, major, role (student, staff, faculty), and university title (freshman, sophomore, junior, etc.).

Netflow data and data collected by Packeteer did not go through either of these processes, so it contains only the recorded IP addresses.

4.1.3 Privacy of Monitored Users

The collection of monitoring data was performed according to the DCP policy guidelines, which include, but were not limited to, the following measures to protect the privacy of monitored users. Data collection was performed at ISU only by ISU staff. The only output from monitoring appliances provided to researchers at CMU was an anonymized version of the collected data. To make it impossible to unveil personally identifiable information such as the ULID of a person, an IP address, or a MAC address, such fields were removed. Some were replaced by pseudonyms generated using a one-way 256-bit hashing function[30]. Both the data collection process and the generation of pseudonyms were performed in an automated fashion without human intervention, so no human ever saw the raw data, and the keys used in the hashing function were destroyed. Monitoring and anonymization were controlled by members of the network management team at ISU, which could have access to the raw data anyway, and they were precluded from analyzing the anonymized data. CMU researchers who analyzed the resulting data were not allowed to observe raw data prior to anonymization, thus being unable to connect any of the data to a specific person, computer, or location on campus. Both the ISU Institutional Review Board (IRB) and the CMU IRB approved this research and data collection procedures.

[29] University Logon ID, a unique identifier assigned to each person in the ISU campus.

[30] Function F(K,X) → Y that, given a key K and an argument X, generates Y, a 256-bit long representation of X. F minimizes the probability that different X arguments will return the same Y. Furthermore, it is, in practical terms, impossible to map back from Y to X.

4.2 Summary of Collected Data and Definitions

Monitoring appliances collected data for about one month in each of the Spring 2007, Fall 2007, and Spring 2008 academic semesters. In each of the periods, AM collected a log of the *events* described above. In Spring 2007 Packeteer collected hourly summaries with total amount of bytes and number of communication sessions entering and exiting the ISU network, broken down by protocol/application. In Fall 2007 and Spring 2008, it collected one individual record per detected communication session, which is a Netflow v.5 record[31] augmented with identifiers of the protocol/application used in the communication. Table 2 presents a brief summary of the data collected by each appliance in each period.

Table 2. Summary of data collected in the three monitoring periods by AM and Packeteer.

	Spring 2007	**Fall 2007**	**Spring 2008**
Number of people living on campus	6,544	6,764	6,763
Time span of AM data	03/31 to 04/30	09/01 to 10/04	02/12 to 04/27
Full hours / days with AM data	648 / 25	654 / 26	1,747 / 60
Number of AM events collected	24.6 million	22.2 million	58.1 million
Time span of Packeteer data	04/01 to 04/30	08/30 to 10/01	03/07 to 05/01
Full hours / days with Packeteer data	720 / 30	735 / 29	858 / 31
Number of Packeteer events collected	hourly summaries	3.3 billion	4.3 billion
Hours / days with both AM and Packeteer data	642 / 25	541 / 20	770 / 24

Data collected through network monitoring is always dependent on the vantage point where monitoring devices are deployed. In this case, both AM and Packeteer were deployed at the point where the campus network connects to commercial ISPs, which means that they only detected communication sessions in which one party was inside the campus network and another party was in the external Internet. Hence, none of the data collected by AM or Packeteer contain records of intra-campus communication sessions nor of communications routed through Internet 2.

[31] For a list of fields typically contained in a Netflow v.5 records consult https://hypersonic.bluecoat.com/packetguide/7.3/info/netflow5-records.htm

4.2.1 Definitions

Analysis in this chapter and part of the analysis in chapter 5 deals with two main types of activity: the usage of P2P and the usage of P2P to transfer copyright-protected media. We will use the following definitions in both chapters:

- A **P2P activity** is as a communication event detected by AM or Packeteer, in which information is transferred using a P2P protocol.
- A **P2P user** is a network user detected doing at least one P2P activity in one monitored period.
- A **Detected Attempt to Transfer Copyrighted Media** (**DATCoM**) is a detected AM event corresponding to a transfer or transfer attempt, using a P2P protocol, of media identified as being protected by copyright.
- A **DATCoM user** is a user who is detected doing at least one DATCoM in one monitored period.

Not every DATCoM is a copyright violation as defined by U.S. copyright law (17 U.S.C.). For instance, in some DATCoMs, users may be downloading material that will be used in particular ways that fall under the "fair use" doctrine (17 U.S.C. §107). It is impossible to tell whether the copyrighted material in each DATCoM was used in any way that can be considered "fair use"; we leave such considerations outside the scope of this research. Also, the fact that detection may occur by matching the hash code in the P2P request to a database of hash codes of copyrighted content allows for the existence of some DATCoMs that correspond to P2P requests that never got a reply, in which no actual copyrighted content was transferred. However, for such a request to exist, one of the parties had to advertise that she was making copyrighted content available[32], and the other party had to search for that content and instruct her P2P client to download it. Obtained results do not change significantly if such "empty" DATCoMs are disregarded because nearly all DATCoM users and copyrighted titles were detected in multiple DATCoMs. At least one of these DATComs contained enough bytes to actually correspond to a transfer,

[32] Whether or not making copyrighted content available constitutes a copyright violation is currently the subject of legal dispute beyond the scope of this paper (Bangeman 2008a).

as opposed to a failed request. Hence, while not all DATCoMs detected on campus are copyright violations, most of them probably are, and they are good indicators that users engaged in transfers of copyrighted content using P2P networks.

A DATCoM represents an attempt to transfer content, without distinguishing downloads from uploads. There is no distinction between uploads and downloads because activities detected by AM do not contain conclusive information about direction of transfers. In legal terms, there is a difference between uploading and downloading copyrighted content, which would make it extremely relevant to analyze the extent to which students upload content to peers outside campus or download it from them. Such findings could also be important in terms of quantifying the economic impact of P2P. However, the available data does not allow one to draw significant conclusions regarding downloads versus uploads.

4.3 Results

This section presents the results obtained from analysis of the collected data focusing on students that live on campus. ISU is a public college that was initially established as a teacher's college. Nowadays its focus is broader but it still offers many education-related majors. It has about twenty thousand students, 88% of which are undergraduate (Illinois State University 2006). ISU's residence halls house 34% of its students and accounted for over 96% of all P2P activity detected on campus.

4.3.1 Extent and Evolution of Detected P2P Activity

Each of the three monitoring periods observed represents part of an academic semester in the one-year period between Spring 2007 and Spring 2008. In Spring 2008, the latter of those monitoring periods, P2P usage, particularly to transfer copyrighted content, was widespread on campus. As table 3 shows, in Spring 2008 about 40% of students living on campus were observed performing P2P; of these, 70% were detected transferring copyrighted content over P2P, averaging over four distinct copyrighted movies, songs or TV shows in a day.

Table 3. Percentage of students detected performing P2P and DATCoMs, and of number of copyrighted titles detected per student detected overall in the Spring 2008 monitoring period (95% CI in parenthesis).

	Out of students living on campus	Out of detected P2P users	Out of detected DATCoM users[33]
Students detected in P2P	39.5% (38.3% - 40.7%)		
Students detected in DATCoM	27.6% (26.6% - 28.7%)	70.0% (68.2% - 71.7%)	
Copyrighted titles detected per student in the period	7.9 (7.12 - 8.62)	19.9 (18.12 – 21.75)	28.5 (25.99 - 30.96)
Copyrighted titles detected per student per day	0.24 (0.22 - 0.26)	1.82 (1.71 – 1.93)	4.35 (4.16 - 4.55)

Over the one-year period leading to Spring 2008, P2P activity showed a generalized decrease. This is clear from figure 2.a, which shows the declining daily percentage of users detected engaging in P2P and transferring copyrighted content, and from figure 2.b, which plots the decrease in the daily average number of copyrighted titles being transferred per student living on campus or per detected DATCoM user. Daily averages allow comparing time slices of the same duration for each monitoring period, which makes them adequate for inter-period comparisons[34]. Figures for the duration of each monitoring period are not meaningful for such comparisons due to the different number of monitored hours in each period.

Hence, despite the observed decrease over time, P2P usage is still widespread on campus, as are transfers of copyrighted content using P2P. The number of P2P users detected on campus in the later Spring 2008 period, while falling below the 50% reported by the RIAA (Lamy, Duckworth, and Kennedy 2007), is still in line with previous survey results reporting widespread use of P2P in other university campuses.

[33] Some titles were detected being shared by some users over several days, therefore the overall number of copyrighted titles detected in the period for each user is not equal to the sum of the number of titles detected in each day.

[34] In order to draw meaningful comparisons between monitoring periods, and since the period durations are different, we averaged over sub-periods with similar duration of 1 day, i.e., 24 consecutive hours of monitoring data starting at midnight. Some days were excluded from the calculations, namely Spring break and Easter or Labor Day weekend, because the percentage of students present on campus was much lower.

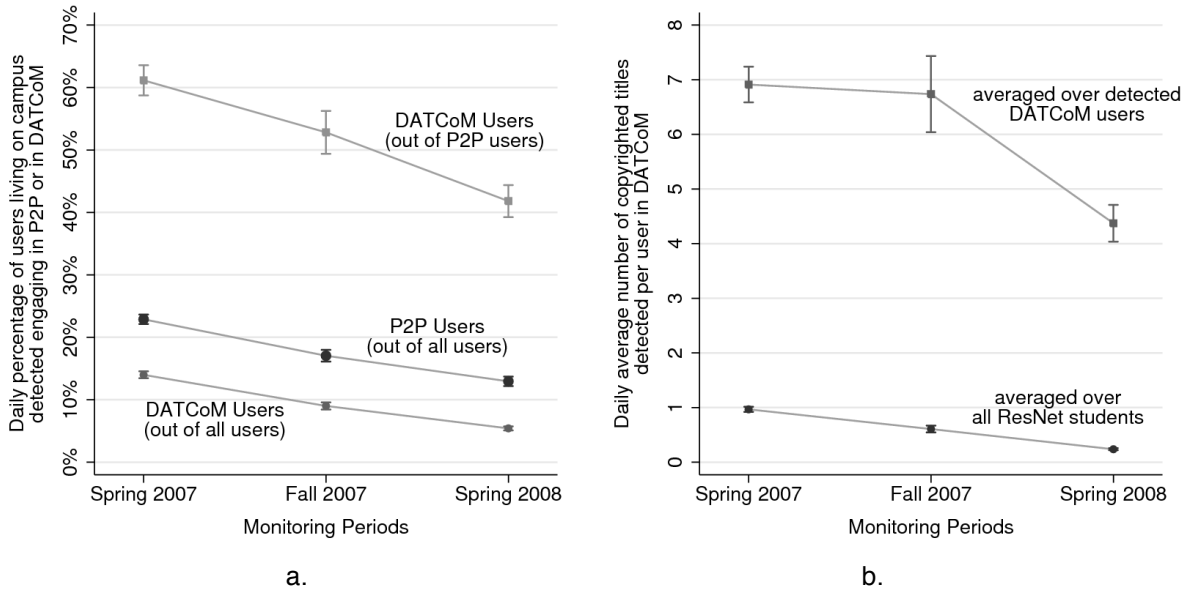


Figure 2. (a) Average daily percentage of students detected engaging in P2P out of all students living on campus, and of students detected engaging in DATCoM, out of all students living on campus and out of detected P2P users in each day. (b) Daily number of copyrighted media titles detected in DATCoM, averaged over all students living on campus and over students detected engaging in DATCoM in each day. Caps show 95% confidence intervals (CI).

4.3.1.1 Interpreting the Results

While we cannot know the exact reason behind the decrease in the percentage of students detected engaging in P2P and transferring copyrighted content between Spring 2007 and Spring 2008, we can offer various plausible hypotheses to explain this event. One hypothesis is that some students did abandon P2P, either for legal purchases of media, or for illegal methods of obtaining content online for free[35]. Another plausible hypothesis, which is confirmed by analysis detailed in chapter 6, is that some students stopped being detected because they activated P2P traffic encryption. Students could abandon P2P or turn to encryption if they knew they were being monitored, and ISU students probably knew that was the case from reading a series of articles about piracy in the ISU student newspaper that warned about plans for network monitoring (Froemling 2006; J Smith 2006; Swasko et al. 2007). However, given that all such articles were published before Spring 2007, the first monitoring period, we believe this does

[35] Other methods of obtaining content online for free, such as video streaming websites and Direct Download Link (DDL) hosts, were becoming popular at the time of monitoring. While those sites have the potential to divert many users from P2P, from very limited analysis pertaining to use of Rapidshare.com (today's top DDL host) we found that very few users downloaded content from that website. Hence, we believe that DDL hosts were not responsible for the observed decline in P2P usage.

not explain the observed decline. In fact, during the monitoring periods, ISU students had no reason to believe that the likelihood of monitoring increased.

Assuming the possibility that some users activated P2P encryption, then the percentage of P2P users detected on campus should be interpreted as a lower bound because the monitoring appliances used for data collection were unable to classify encrypted P2P traffic as being P2P traffic. The precise impact of P2P encryption in the current results depends on whether or not users were knowledgeable of that technique and willing to activate it[36]. Use of technology to limit quality of service or network capacity available for P2P, or to impose some type of punishment on alleged copyright violators, can provide incentive for users to activate evasive measures such as encryption, but the question remains whether such an incentive is enough for users to act. While further technical advances have begun to yield cost-effective network monitoring tools that can detect encrypted P2P traffic, no form of network monitoring can observe the content transferred within such traffic in order to determine whether it is copyrighted.

The percentage of DATCoM users and number of copyrighted titles detected per user should also be interpreted as lower bounds because, under several circumstances, appliances used for data collection fail to detect whether content transferred within P2P traffic belongs to copyrighted works. A detailed account of such circumstances is provided in chapter 5.

In conclusion, the data reported above are lower bounds on both general P2P activity and P2P activity related to copyrighted-content taking place on campus. And the observed decreasing trend may reflect a decrease in actual activity, but may also reflect an increase in use of techniques to conceal online activity from detection via network monitoring.

[36] Encryption of P2P traffic is achievable simply by activating a feature available in most modern BitTorrent and Gnutella clients, the two most popular P2P networks currently in use.

4.3.2 Demographics of P2P Users

Variations of the above figures over demographics (i.e., by gender, class, birth year, area of major[37] or IT savviness[38] of students) and across the three monitoring periods, show decreasing trends in the percentage of detected P2P users, percentage of detected DATCoM users and number of copyrighted titles detected being transferred per user in each demographic subgroup (figures 2 through 5), with only small but sometimes statistically significant differences between demographic subgroups. The fact that activity spans all demographics is another aspect of how P2P and transfers of copyrighted content were and remain widespread on campus.

Differences among subgroups of the campus population could provide insight into possible drivers for P2P use, or be useful from an intervention perspective as a means of targeting actions to reduce the volume of transfers of copyrighted content. The relative uniformity of P2P usage and transfers of copyrighted content among demographics means that, using demographics only, it is not possible to identify subgroups of the population that are substantially different in terms of P2P usage or engagement in DATCoMs so as to allow targeting of interventions aiming to reduce P2P or DATCoMs on campus. In fact, demographics have very low predictive power for either the probability of being a P2P user, the probability of P2P users performing DATCoMs[39], or the number of distinct media titles per DATCoM user. This is in evidence in table 4, which summarizes regression results for three models estimated for the above outcomes of interest using as predictors the demographic variables under analysis[40]. Poor goodness of fit for all models shows that none of the outcomes is successfully predicted based on demographics.

[37] ResNet students were in 79 different majors. Based solely on their major, students were grouped by area of major, which captures the general scientific area of their majors.

[38] IT savviness is inferred from students' majors and it captures the propensity of students to be more IT savvy, which may lead to different online behavior.

[39] Models for probability of performing DATCoMs for all users were attempted, but resulted in poorer fits of the data.

[40] Categorical variables, such as birth year, enrolment semester and area of major were coded using binary dummies. For all models, the base case is that of a female in the 1st semester, born in 1989, in the General Student major, who is not an IT Savvy major.

Table 4. Description of regression models (dependent variable, possible values for the dependent variable, type of regression model and goodness of fit metric) used to assess the predictive power of demographics.

Model	Dependent variable	Value	Type	R^2
A	Probability of being a P2P user	1 for P2P users, 0 otherwise	Logit	0.023
B	Probability of being a DATCoM user for P2P users	1 for DATCoM users, 0 for other P2P users. Not defined otherwise.	Logit	0.055
C	Number of distinct media titles per DATCoM user	Log of the number of media titles per DATCoM user. Not defined for non-DATCoM users.	OLS	0.041

As for the trend over time, the relative uniformity across demographics in the evolution of P2P usage shows that whatever the incentives were for the observed decrease, they seem to have reached all demographics alike. This does not rule out the hypothesis that, for instance, while responding to the same incentives, students in some demographics turned to measures to conceal their activity, while others stopped P2P activity, resulting in the relatively uniform decrease observed.

While not useful for targeting purposes, differences between demographic subgroups can still provide relevant insight about P2P usage on campus. In particular, breakdowns by birth year and by class show that the incidence of detected P2P usage decreases from one monitoring period to the next for every birth year (figure 3.a) and for every class (figure 3.b). Similar decreasing trends are observable in the percentage of users detected engaging in DATCoM as well as in the number of copyrighted titles transferred per DATCoM user. The fact that P2P usage is already high in September for freshman, most of whom have just begun higher education, and that usage declines over time for every given birth year, support claims by higher education officials that students already come to college with entrenched P2P habits (U.S. Congress 2007b), and that it is not in college where they “learn” to use P2P. In addition, the fact that freshman in 2008 use P2P less than freshman in 2007, and the same is true for sophomores, juniors, and seniors, indicates that the demand for P2P is fading somewhat with each subsequent “generation” of students.

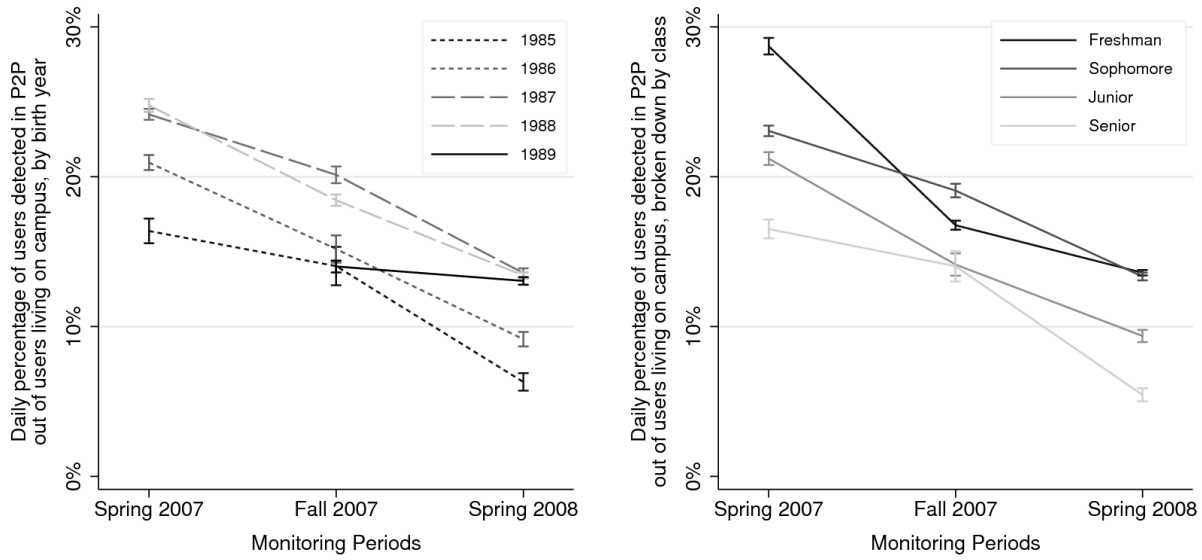


Figure 3. Breakdown of percentage of detected P2P users by birth year and by class. Error bars represent 95% CI. a) Detected P2P users, by birth year: daily percentage of students living on campus detected engaging in P2P broken down by birth year. b) Detected P2P users, by class: daily percentage of students living on campus detected engaging in P2P broken down by class.

Gender and IT Savviness[41] display similar patterns in terms of P2P usage[42], with a significantly higher percentage of males detected doing P2P than that of females (figure 4.a) and also a significantly higher percentage of IT savvy users than of Non- IT savvy users (figure 5.a). However, when it comes to the percentage of users engaging in DATCoMs or to the daily average number of copyrighted titles detected being transferred by the latter, then the roles invert and females and Non- IT savvy users take the lead (figures 4.b and 5.b). This flip can happen because males (or IT savvy students) who engaged in P2P used it more to obtain content that either can be legally transferred over P2P or that the monitoring appliances were unable to detect[43].

[41] IT Savvines is defined based on students' majors. Majors considered IT Savvy are signaled in Appendix A.

[42] The correlation coefficient between gender and IT savviness is 0.17 for all monitoring periods.

[43] Monitoring appliances used can fail to detect specific types of content, and users can take measures to prevent the detection of certain types of content. Both cases are further discussed in section 5.

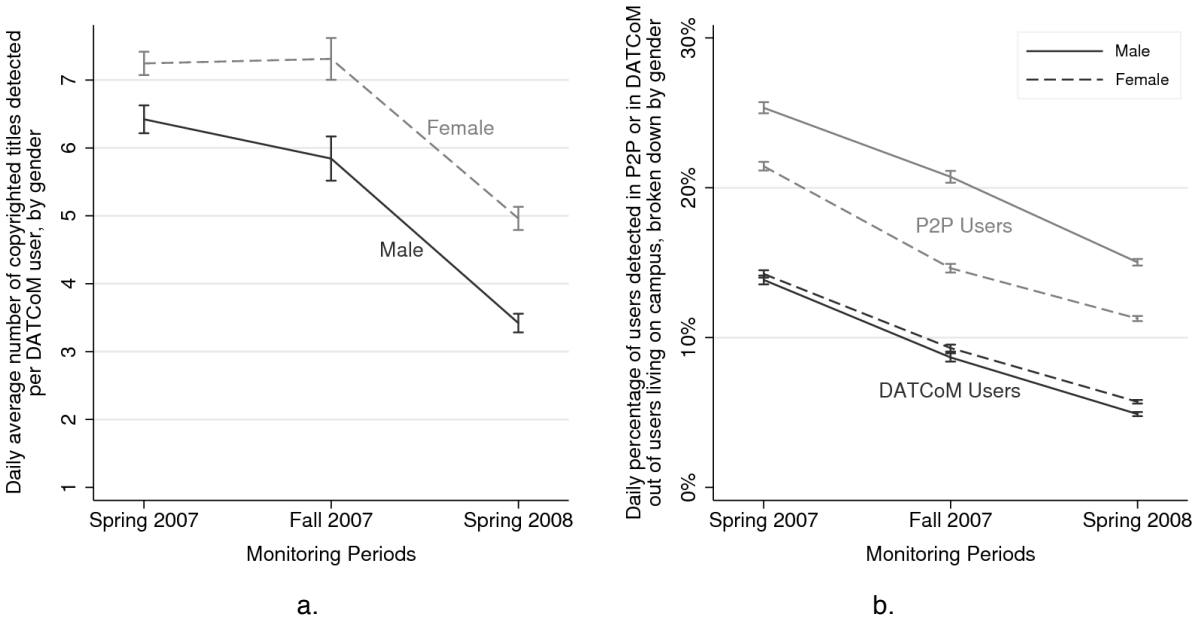


Figure 4. Breakdown of detected P2P activity by gender. Error bars represent 95% CI. a) Users detected in P2P or DATCoM: daily percentage of students living on campus detected engaging in P2P or in DATCoM. b) Copyrighted titles per DATCoM user: average daily number of copyrighted titles detected being transferred per DATCoM user.

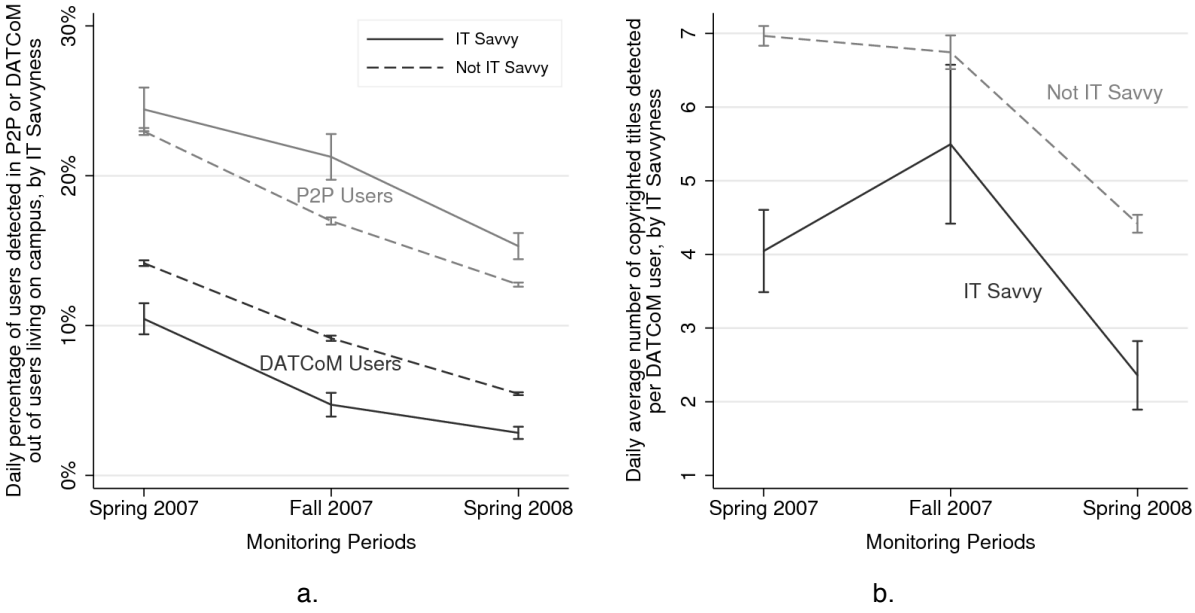
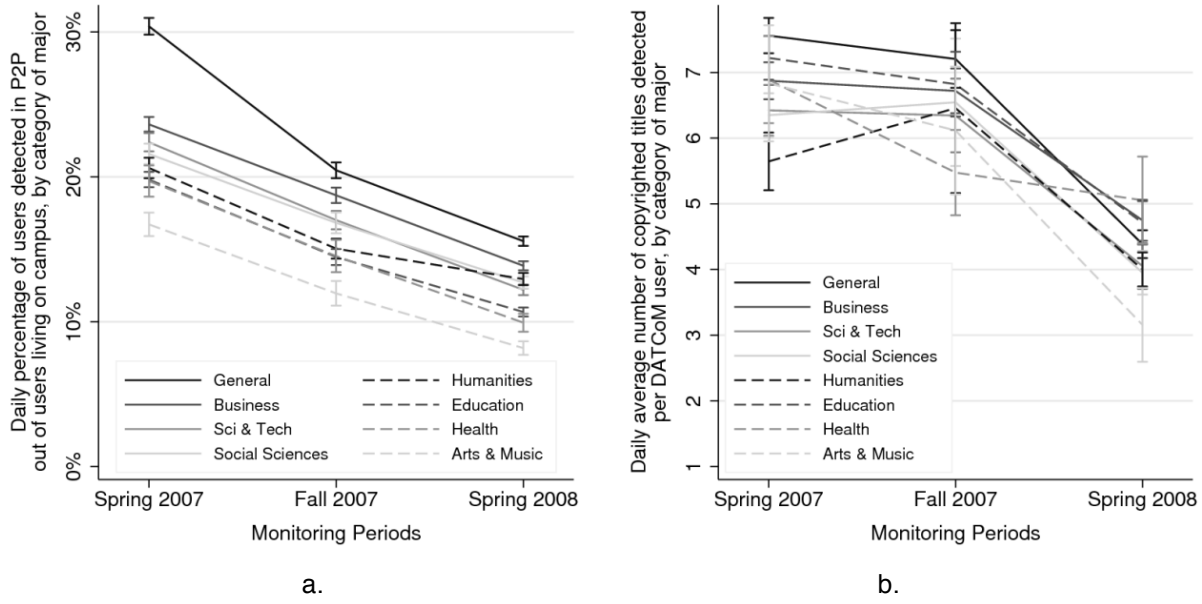


Figure 5. Breakdown of detected P2P activity by IT savviness. Error bars represent 95% CI. a) Users detected in P2P or DATCoM: daily percentage of students living on campus detected engaging in P2P or in DATCoM. b) Copyrighted titles per DATCoM user: average daily number of copyrighted titles detected being transferred per DATCoM user.

Finally, there are few differences in the percentage of students engaging in P2P or DATCoMs among different areas of major (grouping of majors by area of major is presented in Appendix A). From figure 6.a,

out of 8 areas of major, detected P2P usage was highest among General Student majors and lowest among Arts and Music majors (similar to users detected in DATCoMs), both with statistically significant difference from other Major areas. However, looking at the detected number of titles per DATCoM user (figure 6.b), the differences between majors fade and become insignificant in all periods. Hence, while General Student and Arts and Music majors are respectively more and less prone to engage in detected P2P, those that do transfer copyrighted content tend to transfer as many titles as students in other majors.



a. b.

Figure 6. Breakdown of detected P2P activity by Area of Major. Error bars represent 95% CI. a) Detected P2P users, by area of major: daily percentage of students living on campus detected engaging in P2P. b) Copyrighted titles per DATCoM user: average daily number of copyrighted titles detected being transferred per DATCoM user.

4.3.3 Content Transferred over P2P

This section presents a characterization of the types of content that students living on campus transfer over P2P, covering both the titles identified as copyrighted and those that could not be identified by the appliances used for monitoring. DPI technology used in detection of copyrighted content can only identify content from a predefined pool of titles. AM in particular can only detect songs, movies, and TV shows because the pool of identifiable content contains features only for titles of these types. This leaves out

other types of copyrighted content often found in P2P networks, such as software or adult content. While AM cannot tell whether content of the latter types belongs to titles whose transfer over P2P is not authorized, it can nevertheless observe and report which filenames are exchanged, by extracting those filenames from metadata contained within P2P transfers. Analysis in this section uses both media titles contained in DATCoMs, as well as filenames contained in communication sessions not identified as DATCoMs.

Overall in the three monitoring periods, AM detected over 36,000 distinct media titles in DATCoM from about 74% of detected P2P users, and over 100,000 distinct filenames in metadata not identified as copyrighted from about 85% of detected P2P users. As expected, titles detected in DATCoM were found to be songs, movies or TV shows, whereas metadata unveils transfers of many other types of content. Table 5 presents the breakdown of detected titles and filenames by type of content and shows that close to 20% of detected filenames indicate transfers of music albums, adult content, software, books or images.

Table 5. Percentage of copyrighted titles detected in DATCoM and filenames detected in Metadata overall in the three periods, and for each individual monitoring period, broken down by type of content. Some titles or filenames were detected in multiple periods. All columns add up to 100%.

	Titles in DATCoM *n = 36,313*	**Filenames in metadata** *n = 101,879*	**Titles + Filenames**		
			Spring 2007 *n = 64,102*	**Fall 2007** *n = 51,770*	**Spring 2008** *n = 46,896*
Unclassified		7.7%	3.3%	4.7%	7.8%
Song	99.5%	66.0%	83%	79%	70%
Album		2.3%	0.7%	1.3%	2.8%
Movie	0.4%	3.9%	2.2%	2.4%	3.6%
TV Show	0.2%	3.0%	2.1%	1.5%	2.4%
Other[a]		17.0%	8.7%	11.1%	13.4%

[a] *Other stands for: Adult content, Software, Books, Audiobooks and Images*

Table 5 also shows a great disparity between the percentage of movies and TV shows detected among DATCoM titles and those detected among metadata filenames. Such disparity, as demonstrated in chapter 5, is due to greater success of the content detection mechanisms in identifying copyrighted content within audio files than within video files. For this reason, titles detected in DATCoMs and filenames detected using metadata will be considered jointly in the remainder of this section. While filenames may advertise content that is not what is actually stored within the file, P2P searches work by

matching keywords to filenames. Therefore, the filenames are still useful for the analysis in this section, which is concerned with figuring out what content P2P users try to get from P2P.

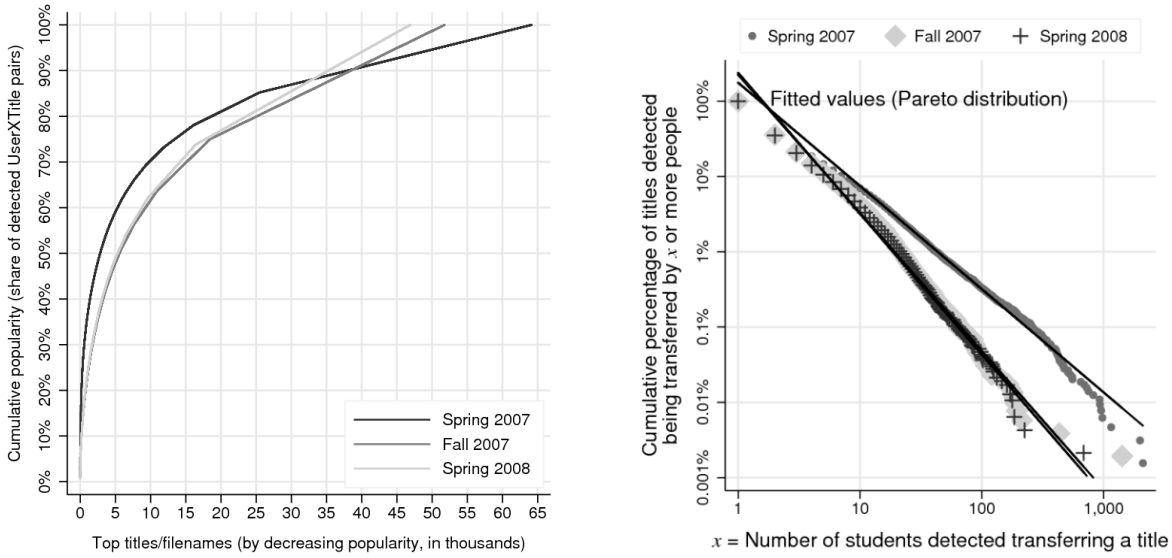
4.3.3.1 Distribution of Popularity of Music and Video

Knowing if P2P is used to transfer the latest blockbuster movie or top-selling single, or if it is used to transfer less popular media not widely available in stores can inform industry decision-makers working on alternantive ways to reach P2P users. For example, these alternative ways might include new marketing strategies to make content sales competitive with P2P, or expanding catalogs to make it easy for P2P users to obtain the material they seek from legal sources.

Figure 7.a plots the cumulative distribution of the popularity of titles/files detected being transferred over P2P. The concept of popularity used here is analogous to the one used for actual sales of content: popularity is defined as the percentage of transfers of a title/file (i.e., the percentage of User-Title/Filename pairs involving that title/file) out of the total transfers of all titles/files (i.e., out of the total pool of User-Title/Filename pairs detected) in each period. The distribution shows that a small set of the most popular titles is responsible for most transfers. For instance, in Spring 2007, only 5% of the files/titles represent more than 50% of all transfers. However, it is also clear that P2P caters to less mainstream taste. For instance, as much as 25% of transfers in Spring 2008 are of files transferred by a single person, and these correspond to as much as 65% of all detected files.

The distribution of the number of people detected transferring each title shows a heavy tail of unpopular titles in all monitoring periods. This is clear from figure 7.b, which plots the inverse cumulative distribution of the number of people detected transferring each title/filename in a log-log scale. It shows that the majority of titles/filenames were detected being transferred only by a few students. For instance, in all periods, over 90% of detected titles and filenames were transferred by less than 10 students each. The

lines in figure 6 result from fitting Pareto distributions ($P[X \geq x] \sim x^{-k}$) to the data in each period[44]. The fitted parameter (k) controls the shape of the distribution: the higher the value of k, the heavier the distribution's tail. In this case, as portrayed in the first column of table 6, k grows from Spring 2007 (1.37) to the later periods (1.84 and 1.87) showing that the tail of the distribution got thicker as time passed. This indicates that there was an increase in the percentage of titles on the low demand side of the distribution, which hints at a diversification of the media obtained through P2P, with P2P users' interest becoming more dispersed through a variety of less popular content and less concentrated in a few very popular titles.



a. b.

Figure 7. Cumulative Distribution Function (CDF) of popularity of detected titles/filenames in the different monitoring periods. a) Popularity of top titles: cumulative share of detected pairs of User-Title/Filename as a function of the rank of detected titles/filenames (example: in Spring 2007, the top 5,000 Titles/Filenames detected accounted for about 60% of all unique pairs of User-Title/Filename detected). b) People per title: markers represent the inverse cumulative distribution of the number of students detected transferring each title/filename, and lines represent the approximation to Pareto distributions (example: in any monitoring period, under 10% of the titles are detected being transferred by 10 or more students).

[44] The shape parameter for the Pareto distribution (k) was estimated using linear regression of the logarithm of the cumulative number of titles transferred by x or more people ($P[X \geq x]$) on the logarithm of the number of people (x), i.e., by fitting linear relationships similar to the ones presented in figure 7.b. The correspondence between the shape parameter of the inverse cumulative Pareto distribution (k) and the exponent of the underlying density power law (a) is given by $a = 1 + k$ (Adamic 2000; Adamic and Huberman 2002).

The heavy tail observed in the distribution of popularity of content transferred over P2P shows that P2P caters to students' interest in a wide variety of media titles. This means that the size of the catalog that is made available to customers is a factor that needs to be taken into account by anyone developing a legal media service if that service is to compete with P2P.

Looking separately at music and video content shows that over the 1-year period of monitoring, the distribution of popularity of songs remained relatively stable, while video saw a great shift towards a higher percentage of titles with only a few transfers each. This is visible in table 6, which presents the shape parameters resulting from fitting Pareto distributions separately for music (songs), and video (movies, TV shows, or adult content).

In Spring 2007, the shape parameter for songs (1.76) was greater than for video (0.62 and 0.64 for movies and TV shows, and 1.1 for adult content), meaning that the percentage of files in the low demand region of the distribution was greater for songs than for the other content types, i.e., that students' transfers of videos concentrated more in popular titles than their transfers of music. This is comparable to the distributions of sales of CDs and DVDs from Amazon.com and of movies ratings from Netflix. Estimates for Amazon.com sales of CDs and DVDs (M Smith and Telang 2007) show shape parameters for a Pareto distribution of 0.99 for CD sales and 0.62 for DVD sales[45]. Using data on title ratings obtained from the Netflix Prize data set[46] yields an estimated shape parameter of 0.66 (R^2=0.88) for Netflix movies[47]. Hence, much like what is observed in campus P2P in Spring 2007, video (whether it is DVD sales or Netflix ratings) seems to be more concentrated around popular titles than music.

[45] Actual estimates were based on the distribution of sales per sales rank ($sales \sim rank^{-b}$) and yielded values of $b = 1.01$ for CDs and $b = 1.61$ for DVDs. As demonstrated in (Adamic 2000; Adamic and Huberman 2002), going from estimates of b to estimates of k, the shape parameter of a Pareto distribution ($P[X \geq x] \sim x^{-k}$, which is equivalent to $sales \sim rank^{-b}$), is trivial, since $k = 1/b$.

[46] The Netflix prize dataset (Netflix 2009) comprises data on over 100 million ratings given by close to 500 thousand users to over 17 thousand movies. The dataset was published at http://www.netflixprize.com as training set for a competition promoted by Netflix and stopped being available for download when the competition ended.

[47] The distribution of popularity of Netflix titles is measured in terms of user ratings and does not reflect actual movie rentals, but it is a fair approximation since movies that are most often rated are also most often rented.

Moving forward from Spring 2007 to Spring 2008, we observe a great decline in the number of P2P users detected transferring video[48] (from 50% to 5.5% of P2P users detected transferring movies and TV shows, and from 27% to 7% transferring adult content). A decline was also observed for music, but it was much smaller (from 66% to 48% of P2P users detected transferring songs). This resulted in fewer people transferring each file/title on average, and in the case of video, the decrease was greater for popular titles. As an example, from Spring 2007 to Spring 2008, the number of students detected transferring the most popular movie or TV show at the time of monitoring decreased dramatically from about 1000 to 13 students, and for the most popular adult content file the decrease was from about 650 to 31 students, while the most popular song saw a milder decrease from 430 to 225 students. The result of such decline in terms of the distribution of popularity of transferred titles was that, from Spring 2007 to Spring 2008, the distribution of transfers of video changed considerably towards a heavier tail. In Spring 2008, the shape parameter estimate for the distribution of video increased significantly (to 2.53 and 2.54 for movies and TV shows respectively, and to 2.46 for adult content) while for songs the parameter increased only slightly (to 1.83). As such, the percentage of titles with low demand increased for both types of video but remained roughly the same for songs. This means that transfers of video over P2P started to dissipate over a wider range of titles instead of concentrating in a few popular titles, which once again highlights the importance of a large video catalog in a legal service that aspires to compete with P2P.

[48] Such a steep decline might be due to some monitoring issue. We considered such a hypothesis, but believe that not to be the case because the monitoring appliances used were the same in all periods. The only thing that changed between periods was the database of copyrighted titles, which is updated periodically with new titles. If it was not updated between periods, then more recent (and likely more popular) titles would not be detected as copyrighted. However, that would not influence the reported statistics, which are drawn from both titles identified as copyrighted and filenames detected in transfers that could not be identified as copyrighted, so an outdated database of titles would just shift detection from copyrighted titles to filenames in metadata, maintaining the general overall counts.

Table 6. Estimates of Pareto distribution shape parameters obtained by fitting the distribution of popularity of Song, Movie, TV Show, and Adult titles and filenames detected being transferred in each monitoring period. Each cell contains point estimates, 95% confidence intervals in parenthesis, and adjusted R^2 values.

	All Media	Songs	Movies	TV Shows	Adult
Spring 2007	**-1.37** (-1.39 to -1.35) *R^2=0.988*	**-1.76** (-1.81 to -1.72) *R^2=0.978*	**-.62** (-.69 to -.55) *R^2=0.805*	**-.64** (-.69 to -.60) *R^2=0.922*	**-1.10** (-1.15 to -1.06) *R^2=0.955*
Fall 2007	**-1.84** (-1.88 to -1.79) *R^2=0.982*	**-1.90** (-1.95 to -1.86) *R^2=0.985*	**-3.06** (-3.30 to -2.82) *R^2=0.993*	**-3.00** (-3.19 to -2.81) *R^2=0.995*	**-2.53** (-2.73 to -2.33) *R^2=0.970*
Spring 2008	**-1.87** (-1.91 to -1.83) *R^2=0.988*	**-1.83** (-1.88 to -1.79) *R^2=0.988*	**-2.53** (-2.84 to -2.23) *R^2=0.969*	**-2.54** (-2.82 to -2.27) *R^2=0.978*	**-2.46** (-2.69 to -2.22) *R^2=0.948*

Many factors may have caused the observed change in the video distribution, in particular when compared to the relative stability of the song distribution, none of which can be readily observed with the available data. One possible factor is the emergence of other sources of media, namely one-click hosts and video streaming websites, which increased in popularity during the timeframe comprising the monitoring periods. However, while we have insufficient data to exclude this possibility entirely, these online sources do not appear to be the cause. For one-click hosts and streaming websites to account for the shape change in the distribution of popularity for video, P2P users would need to be obtaining popular videos from those alternatives and using P2P only for less-popular videos. Although this is possible, there are arguments against this hypothesis. One-click hosts provide music as well as video. Given the stability in the music distribution, users would need to have downloaded popular video from the one-click host but not music. Streaming websites are devoted to video, so the hypothesis only makes sense if users got popular titles from those websites, and got unpopular video from P2P. This is plausible, but we do not have an account of what content was available or downloaded from one-click hosts or streaming websites at the time, so we cannot reach a conclusion in this case. Certainly, this hypothesis would be better grounded if Hulu.com existed at the time, since it typically offers only a few recent TV episodes, but Hulu.com only opened to the general public in March 2008, six months after the observed change in the popularity distribution for video.

Another possibility concerns the maturity of the P2P "market" for music, which has existed since 1999, and was not subject to much change over the course of the year, versus the still embryonic P2P "market"

for video, which continues to evolve rapidly. In particular, there were two significant trends that might have affected how people obtain video from P2P. On one side, the amount and variety of video content available in P2P (in particular through BitTorrent, the main P2P network used for video transfers) increased significantly during the monitoring period, which allowed users to more easily find the content they desired and gave them a wider choice of video content to transfer, thus catering to more diverse user preferences. On the other side, the bandwidth available to Internet users was continually expanding, making it less time-consuming to get large video files from P2P. These trends resulted in a decrease in the "cost" of obtaining a video title from P2P, where cost is measured in terms of the search time necessary to find the desired title in a P2P network (which decreases as more titles become available), plus the actual transfer time of the video file if the title is available (which decreases if there is more bandwidth available for the transfer). So, the fact that it was more expensive to obtain a video from P2P in Spring 2007 than in the subsequent periods could result in users sticking to only a few titles they really wanted in the earlier period, which yields a distribution more concentrated around popular content; while in later periods users could more freely experiment with a wider variety of titles, which would lead to the observed heavier tailed distribution of transferred content.

4.3.3.2 Other Content

P2P file sharing is used for multiple purposes, some of which are legal. Depending on the monitoring period, between 20% and 30% of users detected using P2P did not generate DATCoMs. It is unknown how many of these students transferred copyrighted material without being detected, and how many were involved only in legal use of P2P. This section tries to understand whether legal uses of P2P, or uses that may not be legal but that would not generate DATCoMs, can explain why so many P2P users did not generate DATCoMs. To do so, it focuses on particular content that would not generate DATCoMs when transferred using P2P, namely the Linux OS, the game World of Warcraft (WoW) and Adult content, which were identified using metadata.

Transfers of Linux and other open source software are among the most often mentioned legal uses of P2P and generate no DATCoMs. In all of the monitoring periods, the percentage of P2P users transferring files whose names indicate Linux[49] topped at about 0.2%. The breakdown of this percentage between users with detected DATCoMs and users without detected DATCoMs reveals proportions that are not significantly different from zero. Hence, there is no evidence to support the hypothesis that the transfer of Linux is a motivator for the use of P2P, even among users not detected using P2P to transfer copyrighted material.

World of Warcraft (WoW) is the most popular online multi-player game, with a market share of over 60% and over 11 million monthly subscribers (Magrino 2008; Williams 2009), and enjoys great popularity among college students. The game producers released patches to improve the game's functionality in all monitoring periods[50]. Game patches are distributed using the BitTorrent protocol[51], and thus WoW players are likely to be detected as P2P users because they transfer WoW patches after each update. Transfers of WoW patches were detected in all monitoring periods (decreasing from about 7% of detected P2P users in Spring 2007 to 4% in Spring 2008, and averaging at 4.4% of detected P2P users over all periods). The percentage of students detected transferring WoW but not detected in DATCoMs increased over time (from under 1% of P2P users in Spring 2007 to about 3% by Spring 2008) but remained lower than that of students detected transferring WoW alongside copyrighted music and video when accounting for the three monitoring periods altogether (1.4% of P2P users detected transferring WoW alone, versus 3% detected transferring WoW and engaging in DATCoMs). These figures show that, while there are students who play WoW and are detected doing P2P because they download patches for WoW, the greatest part of students that transfer WoW patches also use P2P to obtain copyrighted music and video.

[49] Files with names that include "linux" or specific major Linux distributions on the market: "fedora", "ubuntu", "suse", "red hat", "mandriva", "slackware" or "debian".

[50] WoW patch releases in the U.S.: v2.0.12 (April 3, 2007), v2.2.0 (September 25, 2007), v2.4.0 (March 25, 2008) and v2.4.1 (April 1, 2008).

[51] The game updater application, which is responsible for implementing the patch download mechanism, relies on the BitTorrent protocol to download game patches. This application is not the same as the one students would use to obtain copyrighted movies and music, although both applications generate traffic that is classified by DPI under the BitTorrent class.

All monitoring periods show significant percentages of P2P users transferring files whose filenames indicate adult content (although with a great decrease from a daily average of 27% of detected P2P users in Spring 2007 to 7.3% in Spring 2008[52]). Transfers of adult content did not result in DATCoMs because the database of copyrighted titles used for identification did not include features for this type of content in any of the monitoring periods. This means that users who transfer adult content but no copyrighted music or video can be detected as P2P users but not as DATCoM users. Nevertheless, detected transfers of adult content do not explain the percentage of users found engaging in P2P but not in DATCoMs, since, between 80% and 90% of users detected transferring adult content are still detected in DATCoMs (figure 8.a). In fact, DATCoM users detected transferring adult content are observed transferring more copyrighted titles on average than DATCoM users not detected transferring adult content, as shown in figure 8.b.

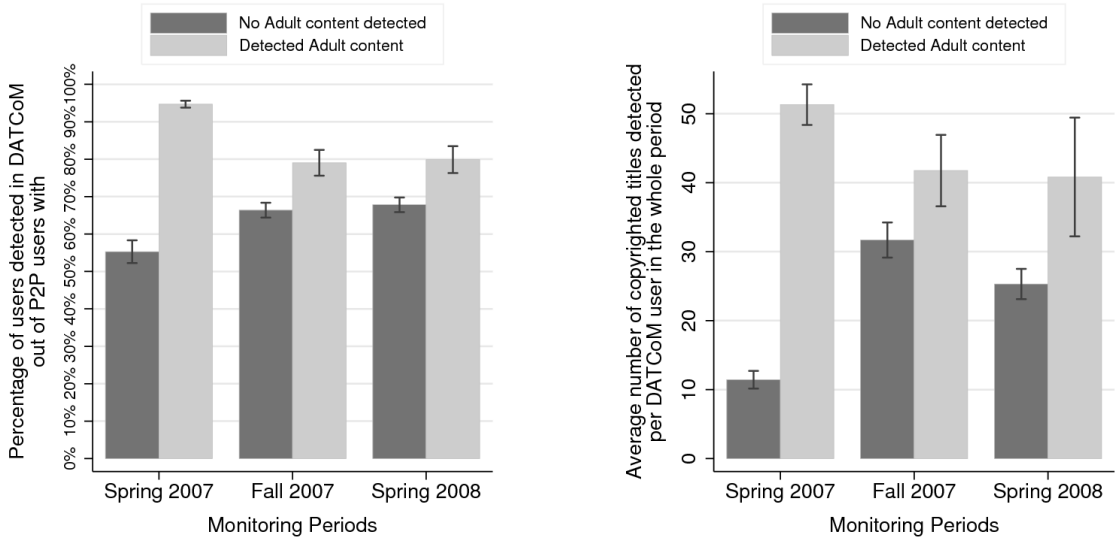


Figure 8. Breakdown of the percentage of DATCoM users and of the number of transferred titles, by whether users were detected transferring adult content or not. a) Percentage of DATCoM users: the percentage of P2P users detected in DATCoM overall in each monitoring period. b) Number of titles per DATCoM user: the average number of copyrighted titles detected per DATCoM user overall in each monitoring period.

[52] One possible explanation for the abrupt decrease in the percentage of P2P users detected transferring adult content is the growing availability of websites that offer free streaming of adult content, some of which is user-generated (Lyn 2008). Such websites are pointed out by the Adult movie industry as being responsible for a decrease in sales of adult movies, an effect that is also expected in P2P transfers of Adult movies.

4.3.4 Relationship between usage of P2P and iTunes

The impact of unauthorized P2P transfers of copyrighted content on the revenues of copyright holders is partly dependent on how much these transfers displace sales of content that could otherwise happen. While our data does not allow assessing how many sales of copyrighted content fail to happen due to P2P, it allows comparisons of media-related activity from popular online media outlets in order to assess whether the behavior towards obtaining media from legal online sources differs between users who perform P2P and users who do not perform P2P. This section compares media-related activity from P2P, the iTunes Store (iTS) and YouTube in Spring 2008 to assess whether P2P users also obtain media from the iTS and YouTube and whether there are differences between download of free content (song samples or YouTube videos) and content that is paid for (songs or iTunes videos).

4.3.4.1 Methodology

This analysis compares P2P and iTunes activity for each IP address in the ISU network. It uses the communication sessions that Packeteer detected for each IP address that were classified as iTunes traffic or as P2P traffic. In this case we believe that the activity detected for IP addresses is very likely to have been performed by a single network user; the reason for that is that IP addresses in ISU's residence halls were leased statically for the entire semester to each user who acquired a dorm room Internet connection. This allows accounting separately for each user connected to the network in dorm rooms.

However, collected data is composed by activity detected for all IP addresses on campus, some of which are not guaranteed to map to a single user during the entire monitoring period. For that reason, pre-processing had to be conducted in order to exclude two groups of detected IP addresses from analysis: addresses for which the Network Address Translation (NAT) could not be reversed, and addresses corresponding to short DHCP leases. The first group exists because Packeteer was deployed on the Internet side of the NAT at the border of the campus network, and thus all IP addresses from within campus were detected in the form of ISU-external translated addresses. This required post-processing at

the time of data collection to reverse such translation using the NAT logs, which was not always successful. As a result, 28% of all detected communication sessions were disregarded in the analysis because they could not be attributed to ISU-internal IP addresses, which made the sessions impossible to attribute to single particular devices[53]. The second group corresponds to IP addresses with short DHCP leases. These were not considered for analysis because they do not capture the behavior of single network users, but most likely of several users assigned to each particular address over time[54]. Such addresses amounted to about 9% of all detected addresses, but they represent negligible network activity in what concerns the analysis in this section (8MB of traffic per address on average in the entire period and percentages with P2P, iTunes or YouTube traffic detected not statistically different from zero).

In the case of iTunes communication sessions, to separate transfers of different types of media from the iTS, each detected inbound iTS communication session was classified into a media category using the following criterion[55]: sessions with less than 480KB were considered control traffic, sessions with 480KB to 1MB were considered sampling of music, sessions with 1MB to 25MB were considered downloads of songs, and sessions with more than 25MB were considered downloads of video[56]. Given that an iTunes user can generate traffic without entering the iTS[57], and analysis concerns song and video-related activities in the iTS, IP addresses were considered to have used the iTS if at least one sampling, song or video activity was detected. Such addresses correspond to 41% of addresses with detected iTunes traffic.

[53] NAT occurs at the IP level in the protocol stack, which makes it independent from anything higher in the stack, particularly transport protocols or application protocols. The ability to translate back using NAT logs maintains this independence. Hence, translation failures are not related to the type of activity contained in the events, which makes us expect these 28% bytes to be missing uniformly across protocols, thus not biasing results towards any type of activity in particular.

[54] Short DHCP leases were not very common in the ISU campus, occurring mostly on the wireless network, which was only available in few places on campus. DHCP-leased IP addresses appear in the data either only once for a short period of time (when the IP address is leased a single time in the monitoring period) or several times but never consecutively for more than the duration of the DHCP lease period (for IP addresses which are recycled by the DHCP server, and therefore leased multiple times in the monitoring period).

[55] This criterion was defined based on observation of the distribution of bytes per inbound communication session with detected iTunes traffic. That distribution displayed clear peaks around traffic volumes that indicate specific activities: around 480KB and around 960KB, equivalent to 30 seconds of a song at a bitrates of 128kb/s and 256kb/s respectively, likely corresponding to music sampling activities; centered around 4MB, likely corresponding to downloads of songs; and above 25MB, with a clear peak around 500MB, likely corresponding to downloads of videos.

[56] There are clearly other types of media that can be acquired from the iTunes Store, such as podcasts or iPod games (in Spring 2008 iPhone App store did not exist yet, hence there are no iPhone application transfers in the monitored events). We assume that the percentage of students that access these types of content was small.

[57] An example of such traffic is the download of album covers when the user transfers music from a CD to her iTunes music library.

YouTube inbound communication sessions were classified using the following criterion: sessions with less than 512KB were considered control traffic, sessions with more than 512KB were considered viewing of videos (which correspond to more than 15 seconds of video at YouTube's minimum encoding rate[58]). To capture activity from people who use YouTube to actually watch videos we consider only IP addresses for which at least one video viewing session was detected.

In the case of P2P communication sessions, since the focus of analysis is in the activity of P2P users that likely transferred some copyrighted media from a P2P network, only IP addresses detected transferring enough P2P bytes to constitute a copyrighted title are considered (a threshold that is set at 3 MB, roughly the amount of traffic necessary to transfer one song).

4.3.4.2 Results

Using the above criteria, IP addresses are classified as P2P or iTS users. Simple tabulations on the data show that use of P2P is found to be correlated with use of the iTS and YouTube. As shown in table 7[59], IP addresses detected engaging in P2P are more likely to be detected using the iTS and YouTube than IP addresses not detected engaging in P2P (and vice-versa). Clearly, to some extent, P2P and the iTS complement each other, a fact that can impact the revenues of copyright holders whose content is sold in the iTS in different ways. If a student uses P2P only when the content she is seeking is not available on the iTS, then P2P transfers of that content have no impact on iTS sales. On the opposite side, if a student uses the iTS only to sample content that she then gets from P2P, then all revenue from eventual iTS

[58] Before February 2009, YouTube supported video with at least 320x240 pixels, encoded at 200kb/s and audio encoded at 64kb/s, which means that the minimum data rate of a YouTube video would be 264kb/s.

[59] The percentages in table 7 refer to IP addresses with or without detected P2P activity. They are different from the percentage of students detected engaging in P2P reported in section 4.3.1 because they refer to different entities, as made clear in the methodology subsection above.

sales is lost[60]. Between the two extremes fall students who buy some content from the iTS and who get some content for free from P2P.

To investigate the relationship between content sampling and purchasing from the iTS and P2P usage, table 7 also breaks down iTS users between those who only sampled content and those who actually purchased content[61]. The table shows that about one third of P2P users still use the iTS, and that close to one quarter of P2P users still purchase content from the iTS, which means that, while use of P2P may reduce the number of people who purchase from the iTS, it certainly does not eliminate it. Thus not all iTS revenue is lost to P2P. Furthermore, P2P users are about three times more likely to use the iTS than non-P2P users, but they are four times more likely to use it only for sampling than non-P2P users. This means that P2P users are somewhat more likely to use iTS resources only to decide what content to obtain, possibly from P2P.

Table 7. Cross tabulations of detected P2P with detected iTS activity (broken down by iTS users detected only transferring content samples or detected transferring songs or videos) and of detected P2P with detected YouTube activity. Percentages in columns add up to 100%

		Not P2P	**P2P**
		76.4%	*23.6%*
Did not use the iTS	*84%*	89.1%	67.4%
Used the iTS only to sample content	*4.6%*	2.7%	10.6%
Used the iTS to purchase songs and videos	*11.4%*	8.2%	22.0%
Did not use YouTube	*61.6%*	66.9%	44.5%
Used YouTube	*38.4%*	33.1%	55.5%

Focusing on users who purchase content from the iTS at some point in the monitoring period, there is no statistically significant difference between P2P and non-P2P users when it comes to the percentage of users purchasing songs or videos, or the number of songs or videos purchased per user. As figure 9.a

[60] The fact that a user samples content from the iTS and then transfers all that content from P2P does not necessarily mean that all the titles transferred from P2P are lost sales. Due to budget constraints or due to willingness to pay for some content being below the price of that content, It is possible that the user would not acquire all the sampled content if she had no way to get it for free.

[61] Table 5 presents the percentages of IP addresses engaged in each activity. Although this is not identical to the percentage of users engaged in this activity, it is similar enough for us to reach conclusions. We believe the primary difference will come from the IP addresses of devices not permanently operated by a user, such as servers, which do not engage in P2P or iTS activity. However, in order for those addresses to have an influence that would alter our general conclusions they would need to comprise over 40% of all observed IP addresses, which does not seem plausible on a campus with twenty thousand students.

shows, about 90% of the users who purchase content from the iTS purchase songs and over 30% purchase videos, equally among P2P users and non-P2P users. Furthermore each P2P user who buys songs (or videos) buys as many on average as each non-P2P user, as depicted in figure 9.b.

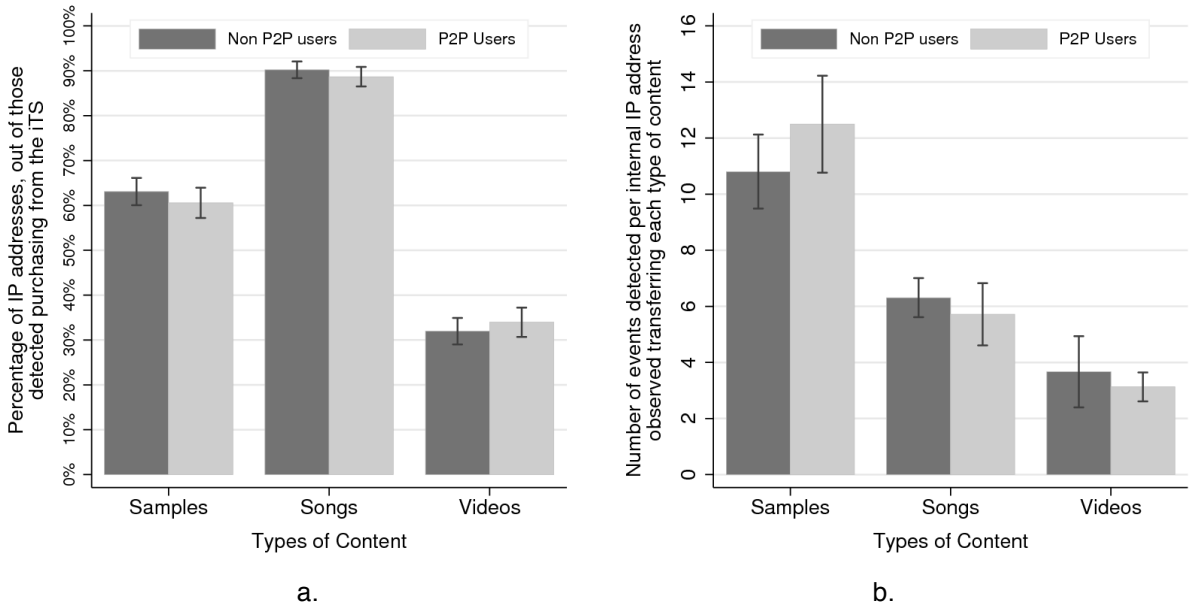


Figure 9. Percentage of IP addresses and average number of events per IP address detected in the iTS, broken down by type of content. (a) Percentage of IP addresses detected sampling music, downloading songs and downloading videos out of those detected using the iTS to purchase content, broken down by P2P usage. (b) Average number of samples, songs and videos downloaded per IP address detected downloading each of such media from iTS, broken down by P2P usage. Caps represent 95% confidence intervals.

It is not possible to tell precisely what impact P2P has on paid services from the detected activity alone, but there are certainly things to learn. There is evidence consistent with the hypothesis that some P2P users use iTS only to obtain free samples, since, out of iTS users, those who do P2P are somewhat more likely to use the iTS only to sample content for free without purchasing than those who do not do P2P. On the other hand, there is evidence that contradicts the hypothesis that P2P users view the ability to transfer content for free as a complete substitute for paid services, since a substantial fraction of P2P users also purchase content from the iTS. Moreover, purchasing behavior in the iTS is very similar for P2P and non-P2P users, i.e., P2P users who purchase content from the iTS do it in comparable percentages and download comparable quantities of songs and videos as non-P2P users who purchase content from the iTS.

4.4 Conclusions and Policy Implications

In this chapter we used data collected between Spring 2007 and Spring 2008 on a university campus to assess the extent and different dimensions of P2P usage on campus. We found that P2P activity and use of P2P to transfer copyrighted content were widespread on campus. In Spring 2008, the latest monitoring period, 40% of the students living on campus were observed using a P2P protocol. Moreover, while P2P can certainly be used effectively for legal purposes, there is evidence that many P2P users are violating copyright law. 70% of those observed using P2P were observed attempting to transfer copyrighted material, and most such attempts are likely to be copyright violations. On average, users detected transferring copyrighted material were observed attempting to transfer four distinct copyrighted media titles during a day of monitoring. Since not all such activity is observable with the Deep Packet Inspection (DPI) tools employed, the actual number of copyrighted titles transferred and the actual percentage of students engaged in these transfers is likely to be higher. Although some people might use P2P only to transfer content legally, we found no evidence that this was common. We considered transfers of the Linux operating system, of World of Warcraft (WoW) game patches, and of adult material (whose transfer over P2P may be legal for some titles but not for others), and found no evidence to support the hypothesis that a significant fraction of students use P2P to obtain those types of content and not to obtain copyrighted music and video as well.

The fact that a law is violated so frequently is good reason to consider a significant change. This could take the form of a change in policy, business practices, enforcement methods, technology, consumer education, or a combination of these. Although our results alone cannot tell us exactly what approach should be followed, they can help answer some important questions.

One such question is whether any particular intervention, such as a campaign to educate Internet users about copyright law, should be targeted at a specific group, or applied generally to all. At least among college students, we found no reason to target any specific group, because P2P users and users transferring copyrighted content were detected across all demographics, with fairly similar incidence

among different genders, ages, classes and majors. However, if education campaigns prove to be useful, there is reason to reach students even before they get to college. We find high incidence of P2P activity for freshmen in their first month on campus, followed by a gradual decrease in subsequent semesters. This indicates that students became P2P users prior to entering college, which is consistent with claims by higher education officials that students "learn" to use P2P at a younger age and already come to college with entrenched P2P habits (U.S. Congress 2007b).

Another question concerns whether and how the industries that produce and distribute copyrighted content could respond to the challenge of competing with a free but illegal alternative. Although it is impossible to quantify from our data how the number of transfers of copyrighted titles observed translates to lost sales, as this depends on many factors, it is reasonable to assume that some sales are lost. Nevertheless, contrary to the common belief that P2P users simply replace paid services with unpaid, our comparison of P2P usage to usage of the iTunes Store (iTS) showed that 22% of detected P2P users still made purchases from the iTS. Moreover, P2P users were more interested in content from the iTS than non-P2P users, and did a significant amount of purchasing; each P2P user who purchased from the iTS purchased on average as much as each non-P2P user who purchased from the iTS. The fact that close to one quarter of P2P users decide to pay for content at times, and to get content for free from P2P (risking legal action by copyright holders) at other times means that there must be factors other than price or risk of legal action weighing in such users' decision.

The variety of content available in each outlet appears to be one such factor. In order to offer its customers a range of content that is comparable to what we observed with P2P, a legal media distribution service would have to offer its users a large selection. We found that a small share of high demand titles accounted for most detected P2P transfers on campus, but also that a heavy tail of low demand titles added up to a significant share of transfers. This means that users' preferences go beyond the most popular titles, and that P2P gives such users access to an extensive catalog to suit those preferences. Unless legal alternatives also provide a large collection of content from which users can choose, they may turn to illegal forms of obtaining the content they desire. While providing such a large selection of content

may be possible for online services, it is typically not possible for bricks and mortar providers of paid content (CDs, DVDs), who are left at an inherent competitive disadvantage. It is probably worth exploring other factors that might influence a user's decision to choose P2P or a legal paid service for copyrighted content. Factors to consider include the convenience or ease of use of each service, ease of search, transfer speed, or usage restrictions in obtained content. In this way, those legally distributing content online may perhaps make their offerings more attractive in order to compete with the free but illegal alternative, at least for the roughly one quarter of P2P users who are still purchasing some media. This may reduce the impact of P2P use on the revenues of copyright holders, although considering that more than three quarters of P2P users were not seen purchasing from the iTS, at the time the top music retailer in the U.S. (Neumayr and Roth 2008), it probably would not eliminate that impact.

Finally, by looking at the evolution of P2P activity over time, we find what might be the beginning of a shift away from observable P2P. Although our observation of only three points in time does not allow for a strong conclusion, we did observe P2P activity on campus decreasing, both in terms of users detected using P2P (10% decrease) and of users transferring copyrighted material (20% decrease, out of detected P2P users), as well as in the average number of unique copyrighted titles detected being transferred per user in a day of monitoring (decreasing from seven to four titles per user detected transferring copyrighted content). Despite this decrease, the detected percentage of P2P users is still consistent with numbers reported by the entertainment industry (Lamy, Duckworth, and Kennedy 2007).

Such a decline can mean that students abandoned P2P for legal services, or for other methods of obtaining content online that can also violate copyright law, such as video streaming websites, direct download link providers (e.g., RapidShare), or Usenet newsgroups. Copyright holders clearly benefit from the continuing growth in use of legal options, but emerging illegal sources also matter, and the extent of their impact is an important open question that may affect the future of online copyright enforcement.

The observed decline can also be due to students escaping detection by using methods to conceal their P2P activity from network monitoring. This would have important implications for anyone considering use

of technology such as DPI to deter copyright violations. As with any technology, those considering deployment must determine whether intended benefits outweigh costs and any unintended side effects. Whether user behavior would ultimately reduce a system's ability to detect transfers of copyrighted material is certainly one of the factors to consider. DPI has limitations that can preclude it from detecting P2P activity and activity involving copyrighted content under certain conditions, one of which is if P2P users activate encryption, which is available in today's most popular P2P clients. It remains a question whether or not users will exploit these limitations to evade detection. In chapter 5 we investigate this use of encryption, considering other possible technological approaches to detecting P2P traffic using network monitoring.

5 DPI as a Tool for Detecting Unauthorized Transfers of Copyrighted Content

This chapter assesses how well DPI performs in detection of P2P activity and transfers of copyrighted content performed using P2P, and how its performance compares to the performance of other technologies that also use network monitoring data to infer the activities in which network users engage. Obtained results allow for a better understanding of the extent and evolution of P2P usage on campus presented in chapter 4, and provide input for the law and policymaking process.

Analysis presented herein focuses on three main issues. First, it assesses how effective DPI is in detecting unencrypted P2P transfers, users of unencrypted P2P and users of unencrypted P2P that transfer copyrighted content. Effectiveness is one of the main factors to consider when determining whether or not to deploy a new enforcement technology, especially in this case, where detection errors may lead to unwarranted penalties to non-violating users. Second, analysis compares DPI's detection of transfers of different types of copyrighted content, in particular audio and video. Audio and video are particularly relevant to detect because these are currently the most common types of content transferred in P2P networks, and the types of content about which copyright holders have been most vocal. Furthermore, popularity of video content in P2P networks has been growing in recent years. Finally, analysis assesses the extent to which encryption is being adopted in P2P, which prevents DPI from effectively detecting P2P traffic, and consequently from detecting transfers of copyrighted content. The dominant P2P networks and software clients in use today offer encryption features that can be turned on by users. If stricter measures for detection and deterrence of illegal transfers of copyrighted content based on DPI are put in place, it is possible that a greater share of users will adopt encryption as a counter-measure to escape detection, and thus go undetected. Such a scenario can lead to a significant decrease of the effectiveness of DPI when used for copyright protection.

The remainder of this chapter is organized as follows. Section 5.1 focuses on assessing the effectiveness of DPI's detection of unencrypted P2P activity and transfers of copyrighted content using unencrypted P2P, taking into consideration the effects of potential detection errors if DPI is used as a tool for enforcement. Section 5.2 covers differences in detection of different types of content, in particular differences in detection of copyrighted audio and copyrighted video. Section 5.3 compares DPI detection of hosts engaging in BitTorrent to detection using alternative detection methods that rely on the network monitoring data collected from campus to identify BitTorrent hosts. The latter methods are used to assess whether there were users on campus that engaged in encrypted BitTorrent, and whether their number increased from Fall 2007 to Spring 2008. Finally, section 5.4 presents a summary of the findings in this chapter and their policy implications.

5.1 Detection of P2P and Transfers of Copyrighted Content on Campus

Collected monitoring data covers three periods, each one a part of an academic semester in the 1-year period between Spring 2007 and Spring 2008. Over this year, detected P2P usage and detected P2P transfers of copyrighted content decreased considerably, as detailed in chapter 4. The average daily percentage of students detected engaging in P2P decreased from 23% to 13% of students living on campus. Out of students detected engaging in P2P, the average daily percentage of those detected performing DATCoMs decreased from 61% to 42%, and the average number of distinct copyrighted titles detected being transferred by each of the latter decreased from 6.9 titles in Spring of 2007 to 4.4 titles in Spring of 2008.

These trends portray the evolution of P2P activity and transfers of copyrighted content over P2P detectable on campus. However, monitoring technology likely failed to detect P2P traffic in some cases, and it erroneously detected transfers of copyrighted content over P2P in other cases, either by failing to detect copyrighted content being unlawfully transferred (false negatives) or by erroneously flagging as

infringing content being lawfully transferred over P2P[62] (false positives). This section focuses on detection errors performed by the appliances used for monitoring on campus, with the goal of informing lawmakers and stakeholders about how accurate and effective current cutting-edge DPI instruments are for copyright enforcement, given that accuracy and effectiveness are two factors that should be taken into account when deploying this type of detection technology, or when adopting a policy that mandates its deployment.

5.1.1 Limitations of DPI in detection of P2P and Copyrighted Content

As discussed in chapter 4, the percentage of P2P users detected on campus should be interpreted as a lower bound because the monitoring appliances used are unable to detect traffic as being P2P if that traffic is encrypted. This is the first limitation of current DPI technology, and while further technical advances may yield cost-effective network monitoring tools that can detect whether users are engaging in encrypted P2P, no form of network monitoring can determine whether the content transferred within encrypted traffic is copyrighted.

At the time of monitoring, the two most popular P2P networks, BitTorrent and Gnutella, supported encryption. Although software clients for such networks do not come with traffic encryption activated by default, in most clients, activating encryption is as simple as navigating to the client's preferences dialog and activating a checkbox. The impact of use of encryption in P2P detection rates depends on whether or not users were knowledgeable that encryption is available and willing to activate it. Several possible incentives may drive P2P users to activate encryption, in particular, the use of network monitoring to limit the quality of service or network capacity available for P2P, or to impose some type of punishment on alleged copyright violators, can provide such incentive. The extent to which existing incentives are sufficient for users to act remains a question. We will consider different technologies that can detect hosts

[62] In legal terms, there is a difference between whether it is possible to detect copyrighted content versus content that is not copyrighted, and whether it is possible to detect content whose transfer over P2P is not authorized versus content whose copyright holders allow transfers over P2P. Copyright infringement occurs only when content is copyrighted and it is transferred without the authorization of the copyright holder.

engaging in encrypted P2P in section 5 to determine the extent to which encryption was being used on campus.

The percentage of DATCoM users and number of copyrighted titles detected per user should also be interpreted as lower bounds. First, due to encryption, which prevents appliances from detecting P2P traffic, and consequently, any content transferred therein. But even within unencrypted P2P, monitoring appliances failed to detect some copyrighted content transferred over P2P. This conclusion derives directly from the method used for detection of copyrighted content. To detect whether content is copyrighted, today's DPI technology uses content matching, a method that works by extracting pieces of content being transferred over P2P and comparing specific features of that content to features of known copyrighted titles. Typical features used are hash codes[63], which can be used to identify any type of content, and audio fingerprints[64], which can be used to identify audio and video. The appliance we used for monitoring takes advantage of these two types of features. Below we detail the circumstances in which content matching fails to detect copyrighted titles.

One of the circumstances that cause content matching to fail to identify copyrighted content is when the features for that content are not present in the database of identifiable titles, in which case the features extracted from transferred content do not match anything in the database. Hence, the more complete the database of identifiable content is, the higher the rate of identification possibly achieved by content matching. However, several tradeoffs must be weighed when building such a database, in particular, the tradeoff between the number of features in the database and the time it takes to search for a match or the storage space required in the monitoring appliance to hold the database. Furthermore, the collection of features to introduce in the database may not be a trivial task, as it requires detailed information about a

[63] See footnote 25 in chapter 4.

[64] See footnote 24 in chapter 4.

wide range of content, and in the case of hash code features, it requires inventorying the many different ways in which each copyrighted title shared in P2P networks is packaged[65].

In the particular case of AM, the content matching appliance we used for monitoring, the content of the hash code and audio fingerprint databases is not public information, which means that the pool of identifiable media is unknown. However, that pool is known to contain only features of songs, movies and TV shows, and it is updated regularly with newly released titles[66] using input from the music, movie and television industries. This means that AM cannot detect other types of media known to be exchanged using P2P, such as software or digital books. And for music, movies and TV shows, it is fair to expect higher sales titles to be better represented in the database, both because those compose the industry's high-revenue fringe and because they are more likely targets of piracy (Van der Sar 2009b).

Content matching also fails to identify copyrighted content when it cannot extract the required features from the media being transferred. Content matching needs to extract enough content from the P2P transfer to allow a meaningful comparison to features in the pool of copyrighted titles, an amount of content that varies depending on the features being compared and on the type of content. When comparing hash codes, only the hash code needs to be extracted from the communication (which is typically transmitted in protocol messages and not with the actual content). For audio fingerprint comparisons, a few seconds of audio are needed and those often correspond to several kilobytes, if not megabytes in the case of video. This can be problematic since files exchanged in P2P are most of the times broken down in small pieces, each transferred in a different communication session, which can make it difficult to extract a large enough piece of media being transferred over a single P2P communication session.

[65] One characteristic of hash codes that identify content being transferred in P2P networks is that they remain the same independently of the file name, because they are calculated based on the actual bytes in the files being shared. However, if the same file is shared in a bundle together with other files, the hash code for that sharing will be completely different. For instance, the same movie shared by itself or together with a subtitles file (a very small text file, when compared to the size of the movie) will yield completely different hash codes.

[66] AudibleMagic reports that its database contains content from 20th Century Fox, EMI, NBC Universal, Sony BMG, Universal Music Group, V2, Viacom and Warner Music Group (http://www.audiblemagic.com/clients-partners/registration.asp).

Finally, even if it is possible to extract the features to perform a comparison and if the particular title is represented in the database of features, the possibility of such comparison failing to produce a positive identification cannot be ruled out, especially if the comparison is not a simple test of equality, but instead based on a likelihood threshold (as is likely the case in audio fingerprinting matches).

5.1.2 DPI's effectiveness in detecting P2P users who transfer copyrighted media

Despite the limitations in detection of copyrighted content discussed above, when given enough time, AM was able to detect most users that attempted to transfer copyrighted content out of those users detected performing P2P. This is shown in figure 10.a, which depicts the cumulative ratio of detected DATCoM users out of detected P2P users as a function of the number of hours given to monitoring in each period. Despite the fact that in the first hours of monitoring the percentage of detected P2P users observed in DATCoM was low, after a couple weeks of monitoring that percentage tended to stabilize around a fixed value corresponding to the majority of detected P2P users. Figure 10.b plots a cumulative evolution similar to that of figure 10.a, but for the percentage of users detected in events with metadata out of those detected performing P2P. Comparing the figures shows that, after the first two weeks, the percentage of P2P users detected performing DATCoMs is not far off from that of users detected in events with metadata, which further shows that DPI could detect attempts to transfer copyrighted content for most P2P users that it could detect transferring files at all.

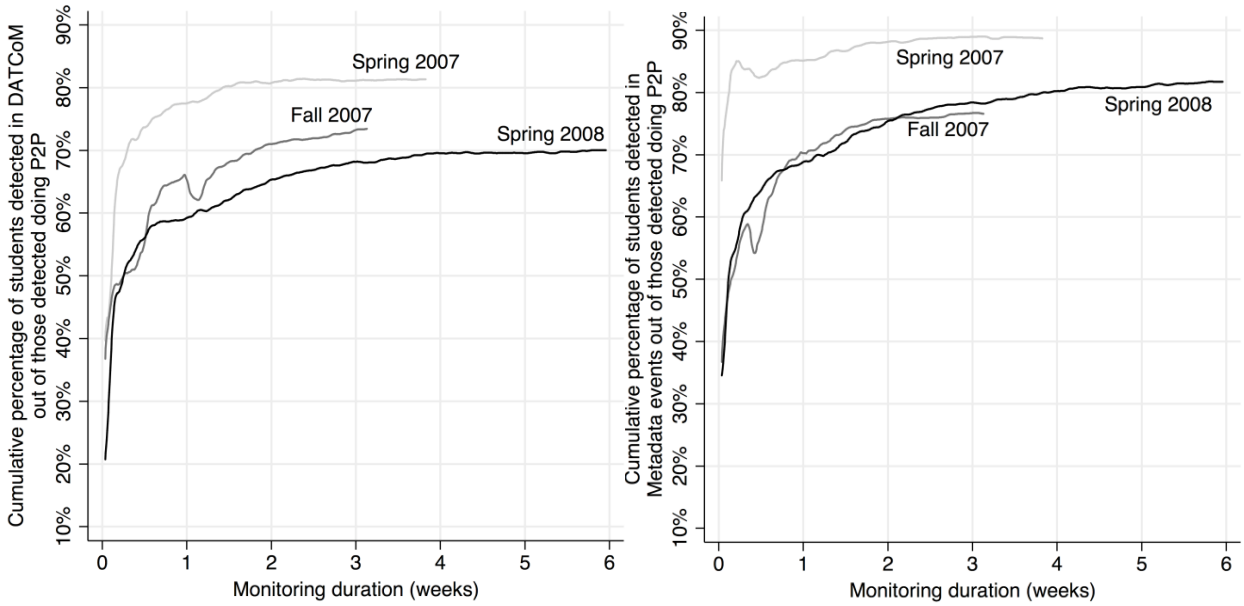


Figure 10. a) Cumulative ratio of detected DATCoM users out of detected P2P users as a function of the number of hours given to monitoring in each period. b) Cumulative ratio of users detected in events with Metadata out of detected P2P users as a function of the number of hours given to monitoring in each period.

In conclusion, AM is effective in detection of users that transfer copyrighted content out of those detected doing P2P if it is given enough time to monitor, which means that there are DPI-based products in the market suitable for the purpose of detecting users who eventually attempt to transfer copyrighted content using unencrypted P2P.

5.1.3 Impact of False Negatives

False Negatives are detection errors that occur when there is copyrighted content being transferred in a P2P communication session but the monitoring appliance fails to detect that content and classify that communication session as a DATCoM. We cannot estimate how many false negatives the monitoring appliance we used yields because we do not have data on what was actually transferred in each P2P communication session. However, by means of simulation, we find that eventual false negatives in detection of DATCoMs have low impact in the number of users detected transferring copyrighted content over unencrypted P2P. Impact is greater when we consider detection of which copyrighted titles were transferred overall, and even greater when detecting which users transferred which titles (user×title pairs).

We estimated the impact that false negatives have in the percentage of detected users, detected titles, or detected user×title pairs by re-sampling our dataset multiple times. In each re-sampling step, we removed a fixed percentage of detected DATCoMs selected at random, and recorded the resulting number of detected users, titles and user×title pairs found in the remaining DATCoMs. This was performed 1000 times for each 1% increase between 0% and 100%. Figure 11 portrays the results of that simulation and shows that not detecting some DATCoMs at random has little impact on the number of detected users, but a higher impact in the number of detected titles or user×title pairs.

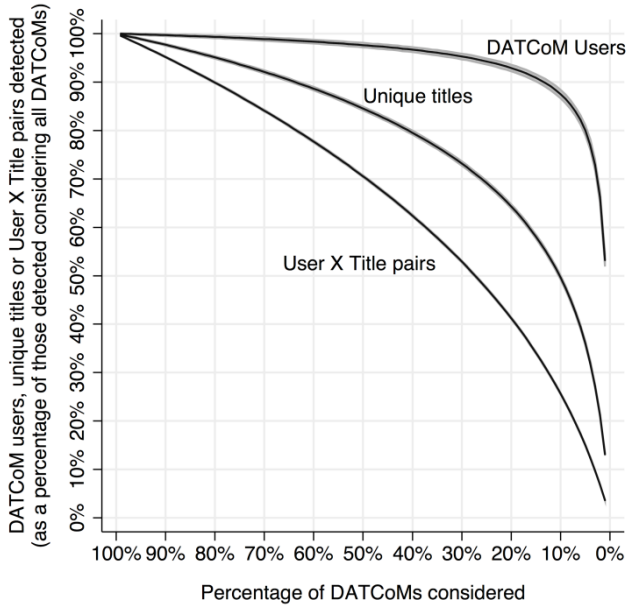


Figure 11. Percentage of DATCoM users, unique copyrighted titles and user×title pairs that would be detected under smaller percentages of detected DATCoMs (data resampled with each DATCoM having equal probability of being removed). Lines represent the mean percentages and shaded areas represent two standard deviations from the mean.

As the figure shows, it is possible to miss most DATCoMs and still meet the goal of detecting most users who transfer copyrighted content at some point. With as few as 15% of the DATCoMs detected in a month-long monitoring period, it is still possible to detect 90% of DATCoM users. In doing this analysis we assume that DATCoMs would be missed at random, which may not be the case. For instance, if certain users transfer mostly titles of genres rarely represented in the database of identifiable content, or take active measures to make detection of transfers harder, then monitoring is more likely to miss DATCoMs for those users. This becomes particularly relevant also when some types of content are more difficult to

detect than others (see section 5.2). In addition, the time given to monitoring is a very important factor, since it is expected that the longer the monitoring the more DATCoMs per user will be collected, and hence, the lower the probability of missing regular users of P2P. Despite these caveats, it is fair to conclude that eventual false negatives have lower impact in detection of users transferring copyrighted content over unencrypted P2P than in detection of individual copyrighted titles or of particular user×title pairs.

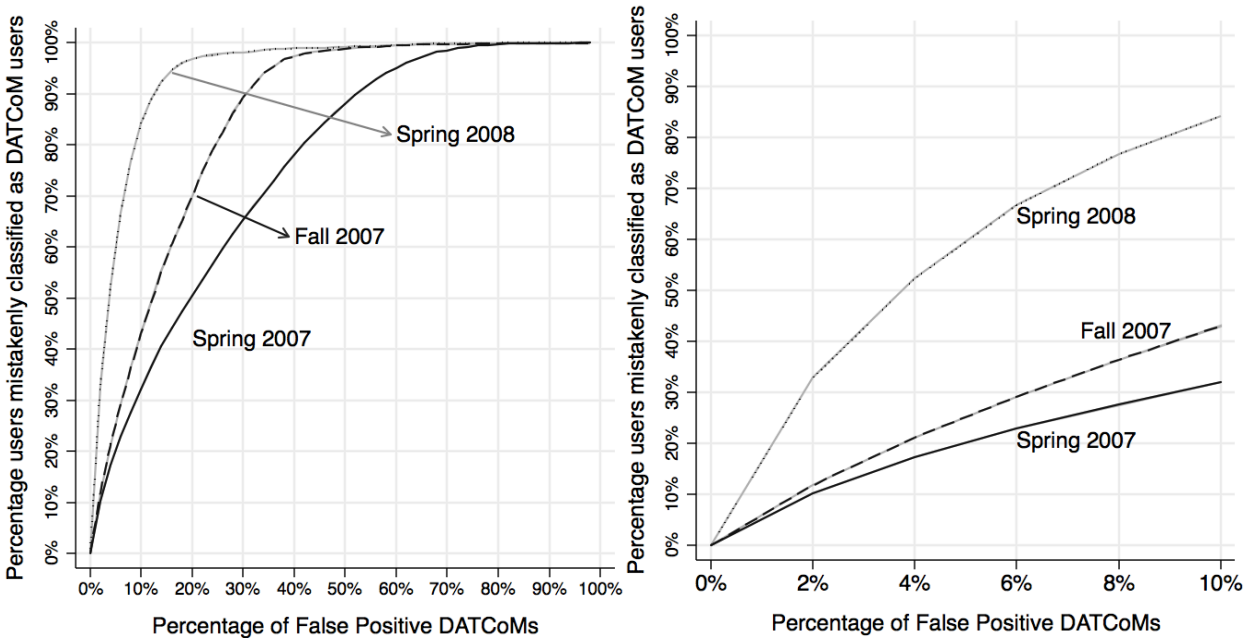
5.1.4 Impact of False Positives

False Positives are detection errors that occur when communication sessions are classified as a DATCoMs but the content transferred therein was being transferred lawfully. This type of error is particularly problematic if the results of detection are used to act upon the user performing the activity in question, since it means acting upon someone that is innocent. We cannot know the actual number of False Positives in the classification of communications on campus, but once again, using simulation we can parametrically estimate the effect that different percentages of false positives would have in the percentage of users detected transferring copyrighted content, or conversely, in the percentage of users that would have been falsely accused of transferring copyrighted content. We find that even small percentages of False Positive DATCoMs can lead to a considerable number of users being mistakenly classified as attempting to transfer copyrighted content. If DPI is actually employed for enforcement of copyright, then one possible way of dealing with this issue is by requiring each user to be detected in a number of DATCoMs before considering that the user is engaging in copyright infringement.

Estimation of the effect of False Positives is performed by means of a resampling process similar to the one described in the previous section False Negatives, with the difference that in this case the resampling process selects DATCoMs with equal likelihood from each detected user[67]. The results of such simulation

[67] This process assumes that each user has equal likelihood of having a false positive. The resampling scheme used to calculate effects of false negative rates assumed each DATCoM to be equally likely to be a false negative, thus making it more likely for users with more DATCoMs to have false negatives. This is clearly not the case when talking about false positives: the likelihood that a user detected in a single DATCoM is the victim of a false positive is much greater than that of a user detected in 1000 DATCoMs. In

are portrayed in figures 12.a and 12.b, which present the estimated percentage of users that would wrongly be classified as DATCoM users under different simulated percentages of false positives in the detection of DATCoMs.



a. False positive DATCoMs from 0% to 100%

b. False positive DATCoMs from 0% to 10%

Figure 12. Percentage of users that would be wrongly classified as DATCoM users in each period under percentages of false positives in detection of DATCoMs. DATCoMs resampled using equal likelihood of false positives for each user. Lines represent mean percentages and shaded areas represent two standard deviations from the mean. a) Variation of the percentage of false positives between 0% and 100% of DATCoMs. b) Zoom in to percentage of false positives between 0% and 10% of DATCoMs.

The figures show that even small percentages of false positives can lead to a considerable number of users being mistakenly classified as attempting to transfer copyrighted content. Figure 12.a shows the general picture of how big an impact false positive DATCoMs can have in detection of users, while figure 12.b zooms in to more plausible percentages of false positives to show how big an effect even small percentages can cause. The difference between monitoring periods is due to differences in the distribution of number of DATCoMs per detected user. Later periods show a heavier tail of users with few DATCoMs each, which makes it more likely for those users to have all their DATCoMs be false positives.

modeling effects of false positives we considered alternative resampling schemes, namely schemes that attributed likelihood of false positives inversely proportional to the number of DATCoMs detected per user, and others with likelihood measures taking into consideration also the average number of P2P bytes per DATCoM detected for each user. Those alternative models yielded results similar to the one with uniform likelihood throughout users, and for simplicity we decided to use the latter.

In practical terms, the actual percentage of users falsely accused of transferring copyrighted content due to false positives depends on the actual False Positive Rate (FPR) of the appliance used for monitoring. In our particular case, AM makes no information available regarding its average FPR. However, there are reasons to believe that it is low, given that this particular technology is implemented in various high-visibility outlets[68] where false positives would lead to great backlash, and such backlash has not been observed[69]. However, given that even very small percentages of false positives can potentially lead to considerable amounts of users being wrongly detected transferring copyrighted content, if such technology is applied for actual enforcement, then a minimum number of DATCoMs should be considered before any user is accused of copyright infringement, and it should be paired with some form of review/appeal mechanism to account for wrongly accused users.

In conclusion, we find that AM, one of the most cutting edge DPI appliances used for detection of transfers of copyrighted content, is more reliable in detecting users that transfer copyrighted content than in detecting individual copyrighted titles transferred by each user, and that given enough time it eventually detects most users with DATCoMs out of the P2P users that it detects. This makes it appropriate for the purpose of detecting users of unencrypted P2P that at some point transfer copyrighted content, but less appropriate if the goal is to detect how many, or which copyrighted titles each of those users transferred. Furthermore, there is the potential that an even small percentage of false positives will lead to many users being wrongly detected transferring copyrighted content. Hence, if DPI is actually employed for enforcement of copyright, then users should not be accused of copyright infringement before they are detected in a number of DATCoMs, and there should be a form of review/appeal mechanism to deal with wrongly accused users.

[68] AudibleMagic lists over 30 clients in their website, among which are YouTube, MySpace, Facebook, MTV or DailyMotion.

[69] There have been discussions going on recently about AM's content identification and the standards that YouTube uses to take down videos that contain copyrighted content (Smitelli 2009). However, most of the debate in this case has been around issues of fair use, and not around issues of whether or not the videos that were taken down contained copyrighted material.

5.2 Detection of Audio Vs. Video

This section characterizes the types of content that students living on campus transfer over P2P and assesses how well DPI can detect transfers of copyrighted content from each of those types. This assessment uses data collected by AM, which, despite being a specific implementation of DPI technology, is a leading product in detection of copyrighted content using this technology, which makes it a good proxy for what DPI can do more generally in this area. As discussed above, one of the main limitations of DPI technology in detection of copyrighted content is that it can only detect content from a predefined pool of titles, and AM in particular can only detect songs, movies and TV shows because the pool of detectable content contains features only for titles of these types. This leaves out other types of copyrighted content often found in P2P networks, such as software or adult content, a limitation that can be overcome by adding features (namely hash codes) for titles of those types of content to the database of detectable titles[70].

Through analysis of the media titles in DATCoMs and of the filenames in metadata, we find that students living on campus transfer a great amount of song and video files whose names indicate songs, movies and TV shows, and which are likely copyrighted, but that are not detected by DPI as being copyrighted. Furthermore, that close to 20% of the filenames found in metadata indicate types of content that AM could not detect because its content database does not include features for such content (software, adult content, music albums, etc.). Finally, we show that the appliance used for monitoring is better at detecting copyrighted audio than at detecting copyrighted video, which can be problematic given that the share of video content in P2P networks is increasing. As a consequence, BitTorrent users on campus were less likely to be detected transferring copyrighted content than Gnutella users, due to the fact that BitTorrent was used mostly to transfer video files while Gnutella was used mostly to transfer audio.

[70] Detection by hash code means that the hash code of the title being transferred has to precisely match the hash code in the database, that is, that the files from which both hash codes were calculated have to be equal bitwise. This can work well for software, where transferred files need to have the same bit-content, otherwise the software will not work. For music or movies, due to differences in bit-rates and encoding, hash code detection is likely to miss many versions of the same content.

5.2.1 Breakdown of detected Titles and Filenames

Information on types of content transferred by students living on campus was collected from media titles contained in DATCoMs and from metadata contained in communication sessions not classified as DATCoMs, which many times corresponds to the name of the file being transferred. Overall in the three monitoring periods, AM detected over 36 thousand distinct media titles in DATCoMs and over 100 thousand distinct filenames in metadata. DATCoMs were detected for an average of 74% of detected P2P users, and communication sessions with metadata from which filenames could be extracted were detected for an average of about 85% of detected P2P users.

To assess how well copyrighted content can be detected within each type of media, media titles detected in DATCoM and filenames detected in Metadata are broken down according to the type of content they advertise[71], and rates of detection using DATCoM are compared to rates using filenames for particular types of content. Filenames provide a control group against which detection of copyrighted titles in DATCoMs is tested. They provide a suitable control group because they can be collected independently of the type of content within the file. Therefore, the percentage of files for which filenames can be collected, out of all transferred files of a given content type, should be roughly the same for all content types. Filenames are equally collected for files containing copyrighted content and for files whose transfer using P2P is completely lawful. To distinguish these two cases, detected filenames are broken down in different categories of content and analysis focuses on those where the probability of a file being copyrighted is greater (filenames indicating known songs, music albums, movies and TV shows). Also, there is no guarantee that the filename actually represents what is contained within the file[72]. For instance, it is impossible to tell whether a file whose filename contains the title of a well known

[71] The classification process was done automatically for the most part of DATCoM titles, using information collected online from Amazon.com's media catalog. For metadata filenames, the filename extension was used to infer the content type, and further classification was performed using the same Amazon.com source, as well as other sources (catalogs of adult content studios, for instance, in the case of identification of adult content). The automatic process classified most of the nearly 140 thousand titles and filenames, but there were a few thousand titles and filenames that could not be automatically classified. These were handled manually using the authors' best judgment.

[72] There is evidence of the existence of files advertising different content than the one they actually contain in P2P networks. Such files are made available for many reasons, and constitute what is called "poisoning" in P2P networks (Christin, Weigend, and Chuang 2005).

copyrighted work actually contains that work. However, since today's most popular P2P networks sport some type of content rating system which allows for "fake" files to be flagged and consequently disregarded by users, and since searches in P2P are performed by matching the filenames of shared content to the search keywords, it is fair to consider such filenames as a good proxy for content transferred over P2P.

Titles detected in DATCoM were found to be songs, movies or TV shows, as expected. Table 8, which is a repetition of the first two columns in table 5 in section 4.3.3, presents this breakdown of detected titles and filenames by type of content. It shows that most filenames indicated songs, movies and TV shows, but close to 20% of them indicated types of content that AM could not detect, such as software, adult content or music albums. Clearly, there is potential for increased detection of transfers of copyrighted content if features from these types are included in the database of detectable copyrighted content.

Table 8. Percentage of copyrighted titles detected in DATCoM and of filenames detected in Metadata for each type of content (columns add up to 100%).

	Titles in DATCoM n = 36,313	**Filenames in metadata** n = 101,879
Unclassified		7.7%
Song	99.5%	66.0%
Album		2.3%
Movie	0.4%	3.9%
TV Show	0.2%	3.0%
Adult / Software / Books / Pictures		17.0%

5.2.2 Audio vs. Video

Focusing only on content types detected by both methods in table 8, we observe a much smaller percentage of movies and TV shows out of titles detected in DATCoMs than out of filenames. This difference becomes even more obvious when taking into account the percentage of P2P users detected transferring movies and TV shows in figure 13. If taking into account only movies or TV shows, then AM would detect at most 4% of all P2P users on campus transferring copyrighted content, but it detects over 25% of P2P users on campus transferring filenames that appear to be movies or TV shows.

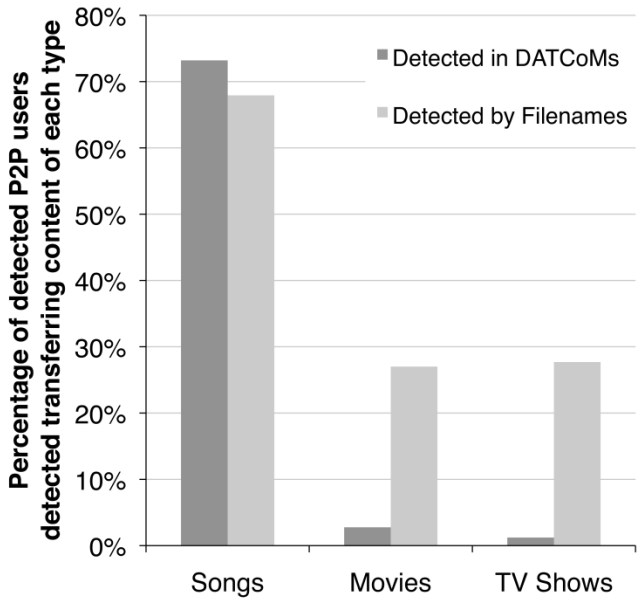


Figure 13. Average for the three monitoring periods of the percentage of detected P2P users detected transferring songs, movies or TV shows by means of DATCoMs and by means of filenames.

One possible explanation for the difference in figure 13 is that AM can better classify copyrighted songs as copyrighted than it can classify movies or TV shows. This can be because movies and TV shows are underrepresented in AM's content database, which is an hypothesis that cannot be tested with available data, or because, for titles equally represented in AM's content database, AM fails to detect copyrighted videos and TV shows more often than songs. This second hypothesis is suitable for testing using the available data, and it is corroborated by the fact that, for technical reasons, copyrighted video transferred over P2P can be harder to detect than copyrighted music using AM's classification technology.

Comparing DATCoM and filename detection rates for a set of songs and movies known to be present in AM's database does indeed support the hypothesis that AM fails to classify copyrighted video as copyrighted more often than it fails to classify copyrighted audio. To conduct such comparison, each of the top 100 copyrighted songs and movies detected in DATCoMs was matched to filenames that indicate the same content. Then, the number of users detected by means of DATCoMs, of filenames, and of both DATCoMs and filenames was collected for each of them. The results of this comparison rest on two main assumptions: first, that the number of students who transfer the files in question without getting detected by either DATCoM or metadata is the same for movies and for songs; and second, that the percentage of

filenames that truly advertise the actual content in the file is equal for movies and for songs. If both assumptions hold, then AM being able to detect a copyrighted video as being copyrighted as well as it is able to detect a song would imply that the percentage of people detected by DATCoMs, out of those detected by either DATCoMs or filenames, should be the same on average for movies and for songs.

However, figure 14 shows that this is not the case. The figure presents, in each monitoring period, the average percentage of P2P users detected transferring the top 100 song and movie titles by means of DATCoMs, out of all users detected transferring the titles (by DATCoMs or by filenames). It is clear that more people are detected through DATCoMs in the case of songs than in the case of movies in any of the periods. A formal test of the hypothesis that the percentage of people detected transferring each copyrighted song by DATCoMs is greater than the percentage detected transferring each copyrighted movie by DATCoMs, against the null hypothesis that they are equal, yields statistically significant differences in mean percentages for songs against movies in all periods, ranging from a low of 26% (95% CI: 15% to 36%) in Fall of 2007 to a high of 48% (95% CI: 36% to 59%) in Spring of 2007.

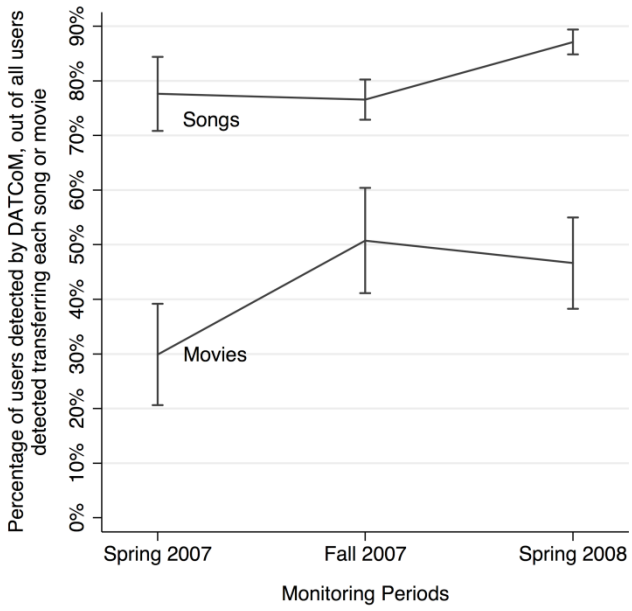


Figure 14. Average percentage of users detected by DATCoM transferring each song and each movie out of all users detected transferring each song or each movie (by DATCoM or by filename).

The logical conclusion is that AM fails to classify copyrighted video as copyrighted more often than it fails to classify copyrighted songs as copyrighted, and that its ability to classify video as copyrighted did not improve significantly in the 1-year period between Spring 2007 and Spring 2008. This implies that the percentage of users detected transferring copyrighted movies or TV shows is a much lower bound than the percentage of users detected transferring copyrighted songs. A broader implication is that one of the most cutting edge appliances in the market for this type of detection has a hard time detecting video transferred over P2P even when that video is present in its title database. Looking forward, unless video detection is improved, if people start transferring greater amounts of video content in P2P, then the percentage of transferred files DPI detects as being copyrighted is likely to decrease. Whether or not this leads to a smaller number of users detected transferring copyrighted content will depend on the mix of content types that such users transfer, particularly on whether people who use P2P to get copyrighted video content also use P2P for copyrighted music.

However, despite the difficulties in detection of copyrighted video and the fact that music albums transferred within archives could not be detected as copyrighted, AM was able to observe most users transferring copyrighted content out of those ever seen transferring audio or video (comprising songs, movies, TV shows and music albums). This is clear in figure 15, which also shows that over the 1-year period between Spring 2007 and Spring 2008 there was a decline in activity related to transfers of audio or video using unencrypted P2P on campus, made clear by the declining percentage of detected P2P users observed transferring audio or video by either DATCoM or metadata. This decrease is independent of whether or not AM's pool of detectable content was updated with latest most popular titles over time because detection by means of metadata is independent of the titles in that pool. Nevertheless, the figure also allows concluding that AM's pool of detectable content was timely updated, because otherwise we would observe a growing gap between the percentage of P2P users detected by DATCoMs and that of P2P users detected by filenames.

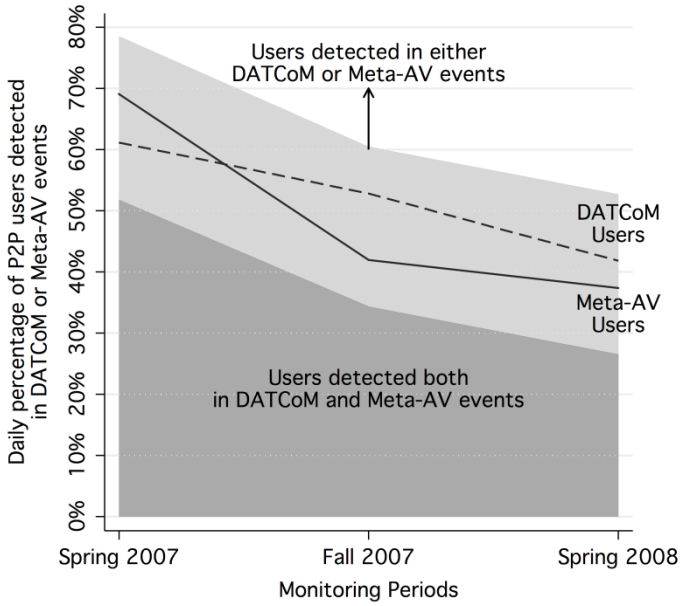


Figure 15. Average daily percentage, for each monitoring period, of users detected engaging in DATCoMs (DATCoM users), in activities containing Metadata whose filenames indicate songs, albums, movies or TV shows (Meta-AV users), in either of those two, or in both of them, out of P2P users detected in each day.

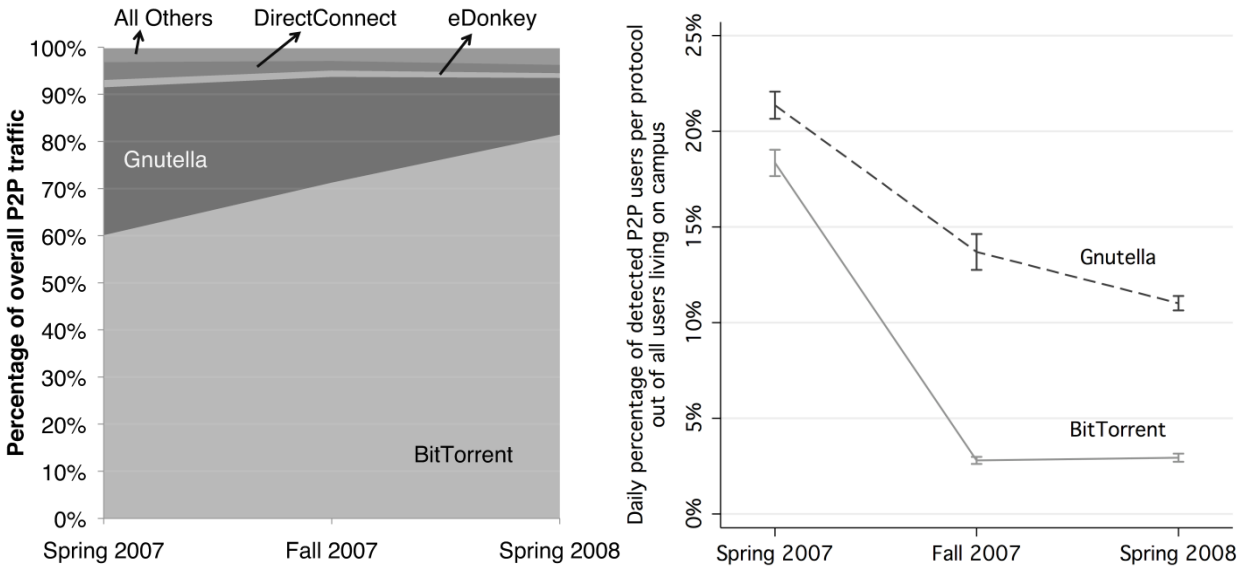
5.2.3 Consequence: Detection of Copyrighted Content in Different P2P Networks

There are multiple P2P networks in use nowadays, some of which cater to particular audiences and are used to transfer specific types of content. BitTorrent and Gnutella are the two main networks used in the ISU campus during our monitoring, and the fact that Gnutella is used mostly to transfer songs while BitTorrent is used mostly to transfer video and archives causes Gnutella users to be much easier to detect in DATCoMs than BitTorrent users.

Collected data shows that more than 20 P2P networks were used on campus in all monitoring periods[73]. However, most of those networks represent a residual percentage of traffic. The main P2P networks detected were BitTorrent, Gnutella, eDonkey and DirectConnect, altogether accounting for over 95% of all detected P2P traffic. Out of those, as figure 16.a shows, BitTorrent and Gnutella were clearly the dominant networks, accounting for over 90% of traffic, an observation that is consistent with reports on

[73] At some point in the monitoring periods, there was traffic detected for the following P2P networks: BitTorrent, Gnutella, EDonkey, DirectConnect, SoulSeek, Ares, Manolito, WinMX, IRC-DCC-Send, OpenFT, Twister, FastTrack, Soribada, Morpheus, Blubster, KaZaA, PeerEnabler, Hotline, Napster, EarthStationV, Furthurnet, Filetopia, Aimster, Audiogalaxy and Groove.

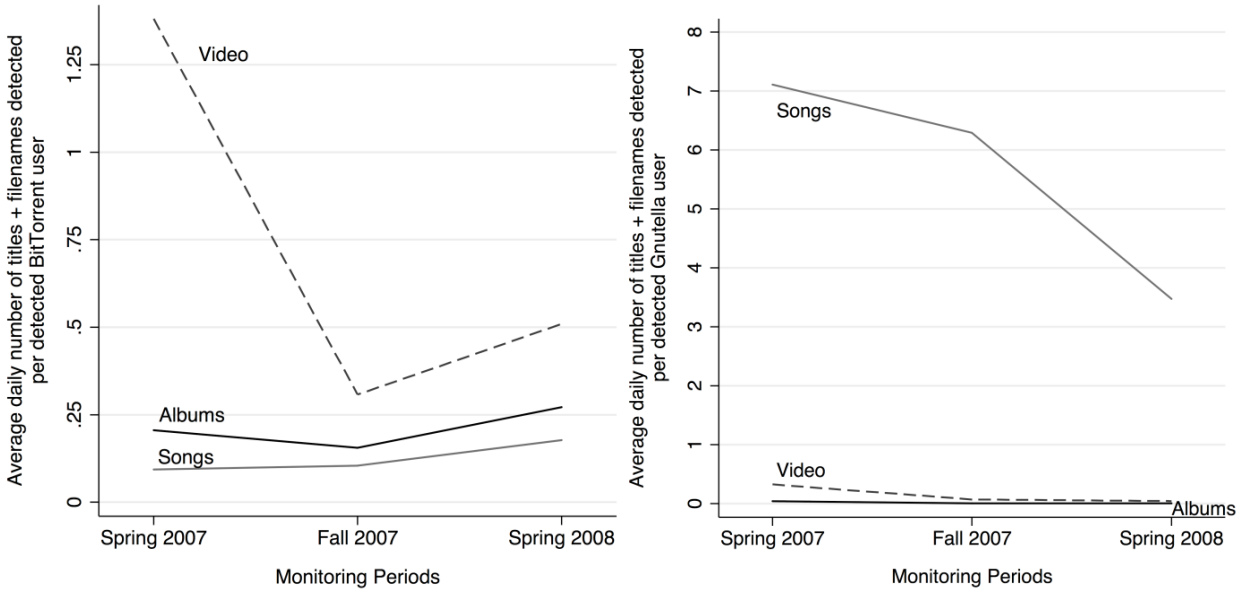
dominant P2P protocols overall in the Internet (Menta 2008). BitTorrent is the dominant network in terms of traffic, and its share out of overall P2P traffic is increasing over time. Gnutella, while showing a decreasing share out of all P2P traffic, is dominant in terms of detected users, as figure 16.b shows.



a. b.

Figure 16. (a) Break down of percentage of P2P traffic by P2P network over the three monitoring periods. (b) Evolution of the percentage of detected BitTorrent users and Gnutella users on campus over the three monitoring periods.

Since Gnutella has more users but less traffic than BitTorrent, then each Gnutella user must transfer on average less traffic than each BitTorrent user. The cause for this can be understood by looking at figure 17, which shows that BitTorrent users transfer mostly video files (movies and TV shows) and archives containing music albums, whereas Gnutella users transfer mostly individual songs. Considering that a typical video file contains about 100 times more bytes than a typical song file and that a music album contains typically about 10 individual songs, then it is clear why the traffic from Gnutella users adds up to a smaller amount of bytes than the traffic from BitTorrent users.



a. BitTorrent b. Gnutella

Figure 17. (a) Average daily number of titles and filenames detected being transferred per BitTorrent user, broken down by type of content, over the three monitoring periods. (b) Average daily number of titles and filenames detected being transferred per Gnutella user, broken down by type of content, over the three monitoring periods.

The percentage of users detected transferring copyrighted content is much lower among users detected using BitTorrent than among users detected using Gnutella. Figure 18 shows that in any of the monitoring periods, over half of detected Gnutella users are observed transferring copyrighted content versus only up to 10% of detected BitTorrent users. This difference can also be attributed to the different content transferred using each network, since the appliance used for detection of transfers of copyrighted content can detect copyrighted songs with greater success than it can detect copyrighted video, and it cannot detect copyrighted content transferred within archives at all, which makes it harder to detect the two principal types of content transferred by detected BitTorrent users.

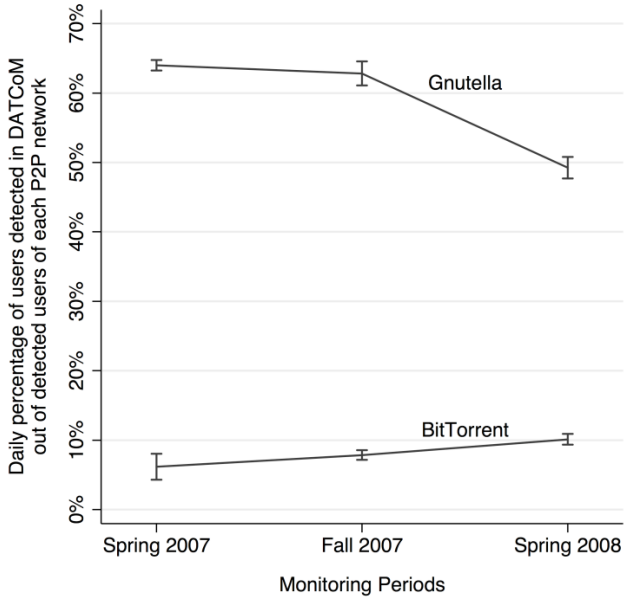


Figure 18. Average daily percentage of detected DATCoM users out of detected BitTorrent users and out of detected Gnutella users, for each monitoring period.

Hence, due to the type of content typically transferred in BitTorrent, the P2P network with the highest share of overall traffic on campus, and on the Internet by many accounts, is the one where DPI has greater difficulty in detecting copyright infringements.

5.3 Encrypted BitTorrent: Trends and Detection Methods

This section applies a variety of techniques to quantify the extent of both encrypted and unencrypted BitTorrent traffic on the ISU campus. This analysis serves two important purposes: (i) to understand how use of encrypted P2P, which is undetectable using prominent tools like DPI, might be changing over time, and (ii) to investigate the effectiveness of various technical tools at detecting P2P, even when some users take steps to evade detection. More specifically, we focus on BitTorrent as our representative of P2P, because it was the most popular P2P protocol at ISU, and overall on the Internet, and one of the two P2P protocols that support encryption.

Beginning with the first of these two purposes, a broad understanding of how copyrighted files are disseminated online must include some consideration of use of encrypted P2P. This was apparent in our

analysis of P2P traffic trends at ISU, as shown in chapter 4. Using conventional monitoring tools, it was possible to show that the amount of observable P2P was falling at ISU, but it was impossible to tell with such tools alone whether this was the result of decreasing P2P activity, increasing use of tools such as encryption that make P2P traffic unobservable, or some combination of the two. The real explanation has large policy and business implications because it may indicate whether dissemination of copyrighted material via P2P is actually increasing or decreasing, and whether users are already increasingly adopting measures that would lead to an underestimation in P2P activity, and would evade the technical tools that some are proposing as a means of copyright enforcement. In this section, by employing a wider range of detection tools, we show that there is reason to believe that use of encryption was actually growing on campus, and thus that at least part of the decrease in detected P2P activity was likely due to users escaping detection.

The second purpose of our analysis is to compare and assess the effectiveness of different approaches to detecting P2P traffic, including DPI as well as different kinds of behavioral classifiers, which are able to detect BitTorrent activity by observing only summaries of the traffic generated by hosts. This analysis seeks to understand the comparative advantages of the different detection methods. In particular, the analysis assesses the extent to which behavioral classifiers are effective in detection of P2P activity. If effective, they could be useful for a variety of purposes, including enforcement of copyright policies, education campaigns (where detected users would get warnings, for instance), assessment of the economic impact of P2P on copyright holders, and determining which users may be using counter-measures such as encryption to escape detection via DPI. Results from this comparison reveal some of the advantages and drawbacks of behavioral methods against DPI as a way of informing those seeking to deploy monitoring technology for detection of P2P activity.

Although DPI is the only approach that can consider the content transferred via P2P, including whether that content is copyrighted, other technologies exist to detect P2P activity that present some advantages over DPI. The fact that behavioral detection relies only on traffic summaries makes it less intrusive to the privacy of network users than DPI, which needs to inspect the content of packets to infer whether traffic is

generated by P2P applications. Using only traffic summaries also makes behavioral detection less expensive to deploy than DPI, because collecting traffic summaries can be done with most already existing network equipment. Another advantage is that behavioral detection can identify encrypted P2P activity, which makes it immune to P2P encryption, the most effective form of circumvention of DPI detection.

In order to fulfill the objectives stated above, we use data collected from ISU. This data was already classified into traffic classes using DPI, including a class for BitTorrent. In addition, we implement three behavioral classifiers that use Netflow-like summaries to identify BitTorrent, and we apply these classifiers to the ISU data. Using the results of both types of classification we compare DPI to behavioral detection. The comparisons are established in terms of the share of hosts that engaged in BitTorrent activity that could be identified by each detection method, and in terms of the percentage of hosts that did not engage in BitTorrent activity that are mistakenly flagged as using the protocol. To assess whether users on campus where shifting from unencrypted BitTorrent to encrypted BitTorrent, we compare the average hourly percentage of hosts detected in each activity between Fall 2007 and Spring 2008.

The remainder of the section is organized as follows. Next we describe the methodology we used, focusing on the implementation of the behavioral detection methods (section 5.3.1). We then present the results of detection, starting with a comparison between detection methods, followed by results pertaining to detection of encrypted BitTorrent and to the extent to which BitTorrent encryption was used on campus (section 5.3.2).

5.3.1 Methodology

We implemented three behavioral classifiers that can identify hosts that engaged in communications using BitTorrent. These classifiers use characteristics specific to BitTorrent traffic that can be observed in

traffic summaries, such as Netflow records[74], produced by common network hardware. This section describes the three behavioral classifiers that we developed.

The simplest and probably least expensive of the approaches to behavioral detection we implemented was the Port classifier, which uses port numbers to identify hosts running BitTorrent by detecting whether such hosts communicate using port 6881. In the earliest versions of the BitTorrent protocol, hosts were forced to listen for incoming connections on port 6881, which provided for an easy way to detect BitTorrent activity. This is no longer a requirement in BitTorrent, which decreased considerably the effectiveness of detection via port numbers. However, many BitTorrent clients still use port 6881 by default, and since hosts using BitTorrent typically communicate with a large number of other hosts, there is a non-negligible probability that some of those other hosts still use port 6881, which allows for the detection of the original host even it is not listening on port 6881. The main advantage of detection via port numbers is that it is inexpensive and can be performed quickly, since it needs only to observe the ports that each host uses to communicate. Our implementation of the Port classifier considers that a host used BitTorrent in a given hour if there is at least one communication session in which that host is involved where port 6881 was used (on either side of the communication).

Whereas the Port classifier is based on the presence of a single characteristic common to BitTorrent traffic, the second behavioral method we implemented, the LC classifier, relies on five different characteristics typical of BitTorrent traffic and uses a linear combination (LC) of those characteristics to determine whether a host is engaging in BitTorrent. This type of approach has received significant attention in the literature in recent years and has been proven effective in identifying unencrypted BitTorrent (Collins and Reiter 2006). We selected a mix of characteristics well documented in the literature (Karagiannis, Broido, et al. 2004; Karagiannis, Papagiannaki, and Faloutsos 2005; Constantinou and Mavrommatis 2006; Collins and Reiter 2006; Bartlett, Heidemann, and Papadopoulos 2007), which include high frequency of failed connections, existence of communication sessions with low bandwidth,

[74] For a description of the Netflow record format, please refer to (Claise 2004).

existence of communication sessions with high volume of traffic, communication with many hosts in short time periods, and communication to/from port 6881. Since our purpose is to detect encrypted BitTorrent as well as unencrypted, our selection of characteristics contains only those that are not altered by use of encryption. By relying on a set of different characteristics other than port numbers, this second method has the advantage of being able to detect hosts engaging in BitTorrent even when those hosts do not communicate using port 6881. The implementation of the LC classifier is presented in further detail in section 5.3.1.1.

The third behavioral method we implemented, the Tiered classifier, is motivated by the observation that port numbers are of particular value in identifying encrypted BitTorrent, but not given enough importance by the LC classifier. The Tiered classifier uses the same five characteristics of BitTorrent traffic that the LC classifier does, but in two different tiers: it first detects whether hosts communicate using port 6881, and if not, it then tests the remaining four characteristics. This implementation mixes the best of the Port and LC classifiers, taking advantage of the high detection rates achievable using port numbers, but still being able to detect hosts that engage in BitTorrent but do not use port 6881. In terms of implementation, the Tiered classifier simply aggregates classification results obtained by the Port and LC classifiers, considering that a host engaged in BitTorrent whenever one of the two classifiers detects that host engaging in BitTorrent.

5.3.1.1 Implementation of the LC Classifier

This section describes the five tests used by the LC classifier and how those tests are combined to decide whether or not a host engaged in BitTorrent. The characteristics of BitTorrent traffic used by the classifier are high frequency of failed connections, existence of communication sessions with low bandwidth, existence of communication sessions with high volume of traffic, communication with many hosts in short time periods, and communication to/from port 6881. Those characteristics are detected using a set of fields available in Netflow records describing communication sessions that occurred in the network. In particular, the classifier uses the start and end times of communication sessions, IP addresses of involved

hosts (anonymized), port numbers, TCP flags, layer 4 protocol, and amount of bytes and packets exchanged.

The first three tests, high frequency of failed connections, existence of communication sessions with low bandwidth and existence of communication sessions with high volume of traffic, were extracted from the work of Collins and Reiter (2006), which uses four tests to distinguish BitTorrent traffic from HTTP, SMTP and FTP traffic. Since our implementation must detect encrypted BitTorrent traffic, we exclude the fourth test developed by Collins and Reiter, which compared particular message profiles[75].

Next we describe each individual test that composes the LC classifier and how they are combined into a decision about whether a host engaged in BitTorrent (section 5.3.1.1.1), and provide a brief overview of the how the parameters of each test are chosen (section 5.3.1.1.2).

5.3.1.1.1 Individual Tests

Each of the five individual tests that compose the LC classifier targets a specific BitTorrent behavior. Three of the implemented tests are based on characteristics of *flowlogs*. Following Collins and Reiter (2006), a *flowlog* corresponds to the set of communication sessions between a pair of hosts within a given time interval. We break the data set into periods of 1 hour for analysis, hence we consider *flowlogs* spanning one hour. Each of the three tests based on characteristics of *flowlogs* is then aggregated across all *flowlogs* detected for each host inside the campus network (since monitoring data was collected at the edge of the network, there is one host inside campus and one outside campus in every communication session). The remaining two tests are performed directly at the host level. In the end, all five tests are aggregated into a single test at the host level that decides whether or not the host engaged in BitTorrent.

[75] The test used the fact that un-responded BitTorrent handshake messages are very common. If BitTorrent is not encrypted those messages are always 68 bytes long, which makes them statistically noticeable. However, if BitTorrent is encrypted such messages are padded with a random amount of random bytes and become 96 to 608 bytes long, which makes it harder to detect statistically. The test implemented by Collins and Reiter compares the distribution of message lengths exchanged between each pair of hosts to the typical distribution of message lengths for hosts engaging in BitTorrent, estimated from a set of communication sessions know to be BitTorrent. While we could estimate the empirical distribution of message lengths for unencrypted BitTorrent using the length of messages that DPI classified as BitTorrent, it is impossible to do so for encrypted BitTorrent because we do not have a readily available log of encrypted BitTorrent messages.

Figure 19 illustrates such hierarchy of tests and how they are aggregated. Each rectangle represents one test. Each *flowlog* test is aggregated at the host level into a corresponding host-level test, and then all the host-level tests are aggregated in the final BATP test. Each test has one parameter (operating characteristic – OC) that influences its behavior. OCs are represented in the figure under the rectangle representing each test.

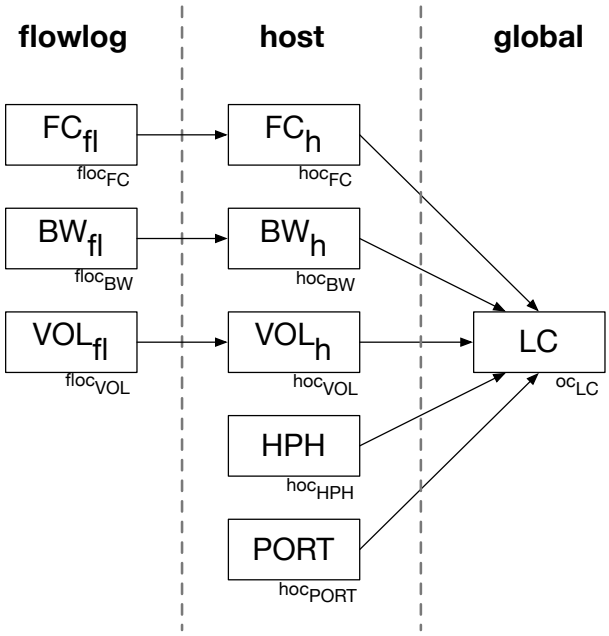


Figure 19. Hierarchy of tests used in the LC classifier. Tests are represented in rectangles and the parameter used in each test is presented under the rectangle. Each *flowlog* test is aggregated in the corresponding host test, and all host tests are aggregated in a final LC test.

The Failed Connections (FC) test takes advantage of the fact that BitTorrent clients often try to communicate with IP addresses that are not responsive to communication anymore, thus resulting in the failure to establish a TCP connection. We implement this test as described by Collins and Reiter (2006). For each communication session in a *flowlog*, we use the number of packets and the summary TCP flags to infer whether the communication session corresponds to a failed TCP connection. Sessions with 1 to 3 packets and only the SYN flag set are considered failed connections. We then calculate the percentage of failed connections out of all communication sessions in the *flowlog*. At the *flowlog* level, the classification parameter for the FC test, $floc_{FC}$, establishes a threshold for the percentage of failed connections in the *flowlog*. If the percentage of failed connections is greater or equal to $floc_{FC}$ then the FC test yields true for

that *flowlog*, otherwise it yields false. To aggregate at the host level we count the number of *flowlogs* for that host that tested positive on the FC test. The classification parameter, hoc_{FC}, is a threshold on this number. If the number of *flowlogs* whose result of the FC test is true is greater than or equal to hoc_{FC}, then the FC test yields true for the host in question, otherwise it yields false.

The Bandwidth (BW) test captures the fact that BitTorrent clients achieve high download speeds by aggregating various simultaneous downloads of different pieces of the content they seek from different hosts, thus resulting in communication sessions with relatively low bandwidth. Once again, we implement this test as described by Collins and Reiter (2006). We select the communication sessions in the flowlog that correspond to file transfers (those that either exchange more than 10 packets, or that exchange 4 to 10 packets but more than 2 kB), and estimate the bandwidth for each of them by multiplying 1500 bytes by the number of packets exchanged in the session, and dividing by the duration of the session. At the *flowlog* level, the classification parameter for the BW test, $floc_{BW}$, sets the threshold for the maximum bandwidth we consider individual BitTorrent file transfer sessions to achieve. If there is at least one file transfer session whose bandwidth is greater than $floc_{BW}$ then the BW test yields false for that *flowlog*, otherwise it yields true. To aggregate at the host level we count the number of *flowlogs* for that host that tested positive on the BW test. The classification parameter, hoc_{BW}, is a threshold on this number, establishing the number of flowlogs (which is equivalent to the number of other hosts) that the host in question needs to communicate with at low bandwidth for that host to be considered as doing BitTorrent. If the number of *flowlogs* whose result of the BW test is true is greater than or equal to hoc_{BW}, then the BW test yields true for the host in question, otherwise it yields false.

The Volume (VOL) test takes advantage of the fact that BitTorrent is typically used to transfer large files, which results in large volumes of information being transferred in some communication sessions between hosts using BitTorrent. This test was implemented as described by Collins and Reiter (2006). We measure of volume of information transferred by the number of packets exchanged in a communication session. At the *flowlog* level, the classification parameter for the VOL test, $floc_{VOL}$, sets the threshold for the number of packets we consider to be an indication of a high volume transfer. If there is at least one

communication session whose number of packets is greater than $floc_{VOL}$ then the VOL test yields true for that *flowlog*, otherwise it yields false. Once again, aggregation at the host level is achieved by counting the number of *flowlogs* for that host that test positive on the VOL test. The classification parameter, hoc_{VOL}, is a threshold on this number. If the number of *flowlogs* whose result of the VOL test is true is greater than or equal to hoc_{VOL}, then the VOL test yields true for that host, otherwise it yields false.

The Hosts Per Hour (HPH) test takes advantage of the fact that BitTorrent hosts typically communicate with a great number of other hosts in short time periods. For this test we count the number of distinct remote hosts that the host being tested communicates with over one hour. The classification parameter, hoc_{HPH}, is a threshold on the number of remote hosts: if this number is greater than or equal to hoc_{HPH}, then the HPH test yields true.

The Port 6881 (PORT) test relies on the fact that many BitTorrent hosts still listen for connections on the legacy port 6881. For each host inside campus, we count the number of remote hosts to which there are communication sessions that use port 6881 (either as the local port or the remote port). The parameter, hoc_{PORT}, is a threshold on the number of such remote hosts: if the number is greater than or equal to hoc_{PORT}, then the PORT test yields true for the local host.

The final LC test aggregates the five host-level tests described above. For each host, it counts the number of tests whose result was true, and if that number is greater than or equal to the value of the hoc_{LC} parameter, then the host in question is considered to engage in BitTorrent, otherwise the host is considered not to engage in BitTorrent.

5.3.1.1.2 Choosing Test Parameters

This section describes the process of choosing values for the parameters of the various tests used in the LC classifier. Our goal is to find the parameter set that yields the best average classifier performance in all hours in the data set. To find that parameter set, we ran a search algorithm for each individual hour to obtain a parameter set optimized for that hour. We then assessed how each such parameter set

performed on average in all hours in the data set and chose the one that presented the best average performance.

The performance of each parameter set is evaluated against the classification provided by DPI for the communication sessions that DPI could classify (i.e., excluding communication sessions tagged by DPI as *Unclassified*). Hence, we seek a parameter set that yields classification results with the maximum agreement with DPI's classification results. To that end, the performance metric we seek to maximize is the rate of true positive classifications against DPI (TPR_{DPI}), i.e., the number of hosts that were detected engaging in BitTorrent by the LC classifier using a given parameter set and also by DPI, divided by the total number of hosts that DPI detected engaging in BitTorrent. Such maximization is performed subject to the constraint that the rate of false positive classifications against DPI (FPR_{DPI}) remains under an acceptable threshold, in which FPR_{DPI} is the number of hosts that were identified engaging in BitTorrent by the LC classifier using a given parameter set but were not detected by DPI, divided by the total number of hosts that were not detected engaging in BitTorrent by DPI.

Using only data on protocols that were identified by DPI we ran a search for the parameters that maximize the TPR_{DPI} while maintaining the FPR_{DPI} under 0.05% for each individual hour of data. We used a search algorithm that greedily walks through the search space in the direction of increases in TPR_{DPI}. The initial state of the search is calculated by optimizing each parameter individually for each test. Chosen initial parameters are those that maximize the TPR_{DPI} for each individual test, while maintaining the FPR_{DPI} under the predefined threshold (or closest to the threshold if no individual test could satisfy the constraint on the FPR_{DPI}). Each state in the search tree can have at most 9 successors, one for each parameter in the parameter set. To generate each such successor, the value of the specific parameter is changed to the one yielding the greatest increase in TPR_{DPI} (while satisfying the constraint on the FPR_{DPI}). If none of the possible values in a parameter yields an increase in TPR_{DPI}, then no successor is generated for that parameter. The search algorithm keeps a list of unexplored states, from which it selects the state with highest TPR_{DPI} to explore next. The algorithm terminates when that list has no further states

to explore, or when 100 consecutive states are explored without yielding an increase in the TPR_{DPI}. The solution returned is the set of parameters that yield the highest TPR_{DPI} of all explored states.

Running the search algorithm for each individual hour yields a parameter set optimized for that hour. Of all of the resulting parameter sets, we then selected the one that performs best when applied to the whole data set. As figure 20 shows (for each monitoring period), despite the fact that in the training stage we required each parameter set to yield at most 0.05% FPR_{DPI} for the hour in which it was trained, when applied to all hours in the data set, the resulting average FPR_{DPI} is considerably higher. The parameter set selected for each monitoring period was the one that yielded the highest average TPR_{DPI} while maintaining the average FPR_{DPI} under 1%.

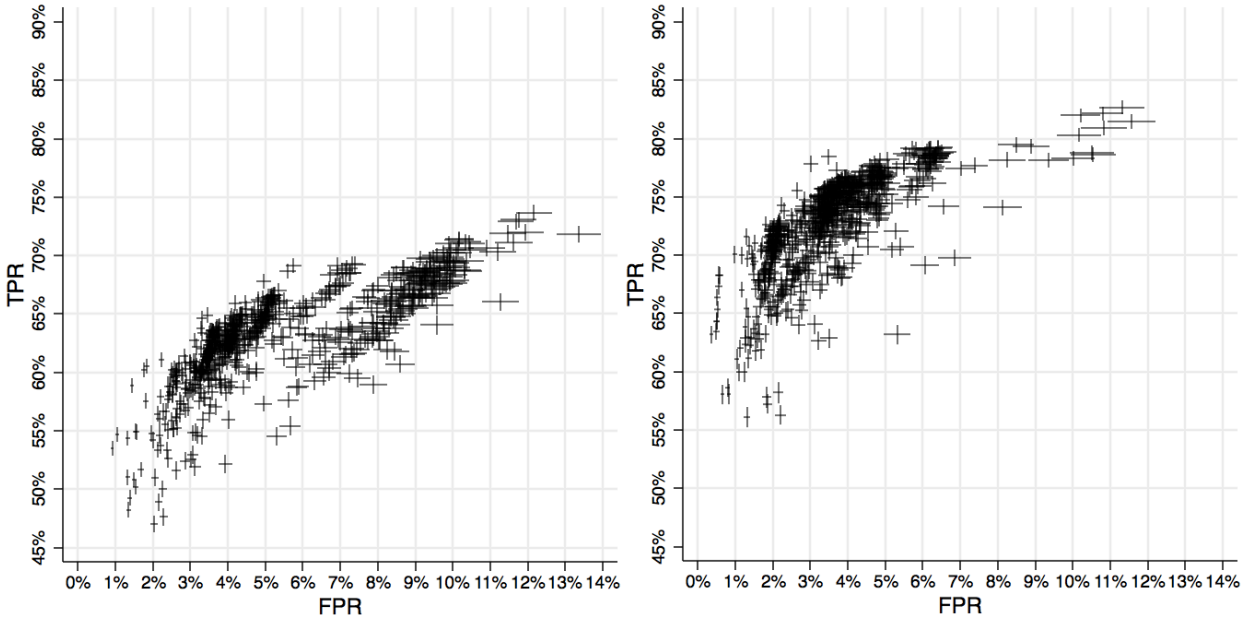


Figure 20. TPR_{DPI} and FPR_{DPI} obtained from classifying the entire data set with the parameter sets found for each individual hour. Horizontal lines represent 99% confidence intervals for the FPR_{DPI}, and vertical lines represent 99% confidence intervals for the TPR_{DPI}. Each intersection corresponds to the mean TPR_{DPI} and FPR_{DPI} found for each parameter set (calculated over all the hours in the dataset classified using that parameter set).

The greedy approach described above is not guaranteed to find the optimal parameter set. However, given the complexity of the calculations involved in evaluating the performance of each parameter set, and the number of possible parameter sets (around 28×10^{12}), an optimal solution is probably

unachievable in practice, and the greedy approach is likely to yield a good solution. The chosen parameter set defines the behavior of the LC classifier as follows.

The LC classifier considers that a host is engaging in BitTorrent in a given hour if at least **4** of the five following tests are true:

- The host communicates with at least **4** other hosts with whom the communication sessions consist of failed connections in at least **28%** of the cases;
- The host communicates with at least **5** other hosts with whom the communication sessions' bandwidth never exceeds **1,856** bytes per second;
- The host communicates with at least **1** other host with whom there is at least one communication session with at least **16** packets;
- The host communicates with at least **5** other hosts during the hour;
- The host communicates with at least **1** other host using port 6881.

5.3.2 Results

This section presents results regarding the performance of the behavioral classifiers we implemented in detection of hosts engaging in unencrypted BitTorrent as a way of evaluating whether or not behavioral detection technology is effective in detection of unencrypted BitTorrent activity (section 5.3.2.1). It also presents results regarding the use of encrypted BitTorrent on campus during the monitoring periods and whether or not there was a shift in usage from unencrypted to encrypted BitTorrent during the 2007-2008 academic year (section 5.3.2.2).

5.3.2.1 Performance of Behavioral Classifiers

In this section we evaluate the performance of the behavioral classifiers we implemented by comparing the behavioral classifiers' rates of detection of hosts that engage in unencrypted BitTorrent against the classification provided by DPI using traffic that DPI was able to classify.

Behavioral classifiers were applied to two data sets collected from the ISU network by Packeteer in Fall 2007 and Spring 2008 and containing, respectively, 715 and 765 hours of network monitoring data. Data used for classification consists of logs of communication sessions in NetFlow format, augmented with a traffic type field with information on the type of traffic that was exchanged in the session, as inferred by DPI. In the monitored periods, DPI could classify about 70% of the bytes exchanged on campus, and the remaining 30% that could not be identified resulted in communication sessions tagged as *Unclassified*. *Unclassified* traffic corresponds to protocols that the DPI appliance was not designed to identify, such as encrypted BitTorrent, as well as possible classification errors. Before being used in classification, data went through a cleanup process in which we removed traffic generated using dynamically leased IP addresses[76]. Different people use each dynamically leased IP address over time, and thus from the generated traffic we cannot capture the activity any single network user. Performing this cleanup did not greatly affect the data used for classification because most network users on campus connected to the network using static IP addresses.

To assess how well the behavioral classifiers detect unencrypted BitTorrent we applied each classifier to datasets containing only the communication sessions that DPI could classify in each period (i.e., not containing traffic marked by DPI as *Unclassified*). We compare the performance of the behavioral classifiers against the classification provided by DPI in the traffic that it can classify[77]. That is, we consider that a host engaged in unencrypted BitTorrent whenever DPI detects that host exchanging BitTorrent traffic, and consider that if DPI did not detect a host exchanging BitTorrent traffic that host did not engage in unencrypted BitTorrent.

[76] We filtered out IP addresses which performed less than 5MB of traffic overall in each entire monitoring period (700+ hours) and which were detected online in less than 2% of the hours in each monitoring period, because IP addresses with few exchanged bytes or that are online for little time are more likely to be temporary DHCP leases. As a result, the datasets we used for analysis contain a total of 17,501 distinct IP addresses in the Fall 2007 period and of 28,223 distinct IP addresses in the Spring 2008 period.

[77] Measuring the performance of behavioral classification against results obtained using deep packet inspection is common practice in the literature, given that, for unencrypted traffic, deep packet inspection typically provides high-accuracy classifications. Another possible approach would be to use a set of hosts, which were scripted to engage in predetermined online activities, and generate a data set of traffic information that could be compared against the effective ground truth. However, such an approach would have to be very carefully designed in order to produce a data set that mimicked actual activity from a network being used by real people. Due to practicality and data availability we used the first approach.

The performance of behavioral classifiers is evaluated using two metrics: the True Positive Rate against DPI classification (TPR_{DPI}) and the False Positive Rate against DPI classification (FPR_{DPI}). These metrics are equivalent to typical TPR and FPR, except that they are calculated against the classification provided by DPI instead of being calculated against the ground truth. TPR_{DPI} is given by $TPR_{DPI} = \frac{TP_{DPI}}{P_{DPI}}$, where TP_{DPI} is the number of hosts identified engaging in BitTorrent by both the behavioral classifier and DPI and P_{DPI} is the total number of hosts identified engaging in BitTorrent by DPI. It is a measure of the extent to which each behavioral classifier is able to detect hosts that engaged in unencrypted BitTorrent. FPR_{DPI} is given by $FPR_{DPI} = \frac{FP_{DPI}}{N_{DPI}}$, where FP_{DPI} is the number of hosts identified engaging in BitTorrent by the behavioral classifier but not by DPI and N_{DPI} is the total number of hosts DPI did not detect engaging in BitTorrent. It is a measure of the extent to which behavioral classifiers mistakenly classify hosts as engaging in unencrypted BitTorrent when DPI shows those hosts did not.

The behavioral methods we implemented were able to detect most hosts engaging in unencrypted BitTorrent in short 1-hour periods. On average per hour, behavioral classifiers detected as much as 70% of the hosts that engaged in unencrypted BitTorrent with a very low rate of false positive classifications (under 0.6%). This is observable in table 9, which presents the TPR_{DPI} and FPR_{DPI} rates for the three behavioral classifiers in Fall 2007 and Spring 2008, both as hourly average rates and as overall rates achieved in the entire periods.

Table 9. TPR_{DPI} and FPR_{DPI} for the Port, LC and Tiered classifiers in the Fall 2007 and Spring 2008 monitoring periods. Presented results correspond to hourly averages, i.e., individual hourly results averaged over all hours in the period, and to overall results taking into account the whole period.

	Hourly average				**Whole Period**			
	Fall 2007		**Spring 2008**		**Fall 2007**		**Spring 2008**	
	TPR_{DPI}	**FPR_{DPI}**	**TPR_{DPI}**	**FPR_{DPI}**	**TPR_{DPI}**	**FPR_{DPI}**	**TPR_{DPI}**	**FPR_{DPI}**
Port	61%	0.02%	65%	0.01%	82%	1.7%	70%	0.7%
LC	51%	0.50%	63%	0.33%	87%	23.0%	81%	34.0%
Tiered	64%	0.59%	70%	0.44%	91%	22.8%	85%	33.9%

If given longer periods to monitor (about a month), behavioral classifiers are able to detect nearly all hosts that engaged in unencrypted BitTorrent, but at the expense of a high percentage of hosts that are

mistakenly flagged as engaging in BitTorrent. As table 9 shows, the percentage of hosts detected in the entire monitoring periods increases to as much as 91% of the hosts that engaged in BitTorrent. However, while the Port classifier yields a low rate of false positives, in the case of the LC and Tiered classifiers, the increase in the rate of detected hosts when given more time to monitor happens at the expense of an increase in FPR_{DPI}, which goes up to as much as 34%. This is probably a consequence of optimizing parameters based on their performance with one hour of data, and applying those parameters over an entire month. Hosts mistakenly identified doing BitTorrent in a very small number of hours add up when considering larger periods and cause the high FPR_{DPI} yielded by the LC and Tiered classifiers. This happens because these classifiers require a conjunction of factors to occur at the same time in order to produce a positive classification (i.e., that a set of tests yield true at the same time), and this is more likely to happen for random reasons the longer the time period given for classification. It is possible to address this false positive problem simply by determining that a host is engaged in BitTorrent only if the amount of BitTorrent-like traffic observed is commensurate with the length of the monitoring period, for instance, by requiring positive classifications in a percentage of the monitored hours before considering that such host engaged in BitTorrent. However, classifiers optimized for each hour are useful for our immediate research objectives, and such complexity is outside the scope of this dissertation.

A comparison of detection figures for the three behavioral classifiers shows that port numbers still play a crucial role in detection of BitTorrent activity, since the Port classifier achieves detection results that are very close to the best results achieved by all behavioral classifiers. The fact that the Tiered classifier outperforms the Port classifier shows that there is clearly an advantage in adding tests for other characteristics of BitTorrent traffic. However, the poor detection figures achieved by the LC classifier indicate that the marginal benefit of using these additional tests in addition to port may be small, because users identified using LC are also generally identified by port number. In both cases, our implementations were first attempts that leave ample room for improvement, and further research on behavioral classifiers may yield better methods that achieve higher detection rates than those that we developed.

These figures show that it is possible to build a simple, and inexpensive method of identifying hosts that might be engaging in P2P using only traffic summaries, provided that false positives and false negatives are tolerable. For instance, a Port classifier may be sufficient if the purpose is merely to issue warnings to users who may be engaging in BitTorrent about copyright issues, but not to impose any penalties. Adding other tests besides port number to the classifier can further improve accuracy.

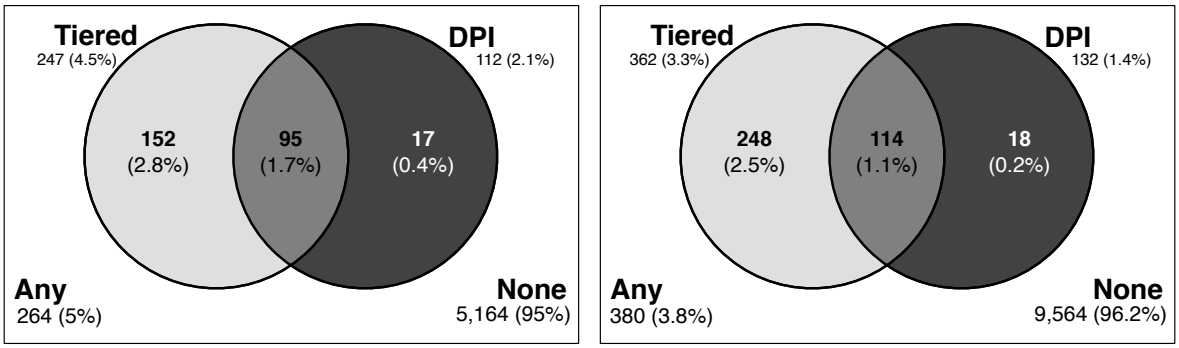
If behavioral classifiers are used in ways that users find problematic, it is possible that BitTorrent users, BitTorrent software developers, or both would adopt counter-measures to avoid detection. However, these actions are arguably more complicated than activating traffic encryption, which is currently available in most BitTorrent clients and prevents DPI from detecting BitTorrent traffic. Detection via port number, which has proven to be the most useful single indicator of BitTorrent traffic, is likely to diminish as time goes by and more BitTorrent clients are upgraded to versions that avoid port 6881 altogether (and thus decrease the probability that any BitTorrent hosts has of communicating to port 6881). The other tests used in our implementation of behavioral detection focus on characteristics of BitTorrent communications that may be difficult to conceal without affecting the performance of the protocol. For instance, if in order to avoid detection by the hosts per hour (HPH) test, BitTorrent clients were prevented from connecting to multiple hosts in short time periods, they would not be able to achieve the transfer rates they achieve nowadays by stacking up multiple connections.

5.3.2.2 Detection of Hosts Using Unencrypted and Encrypted BitTorrent

In this section we assess how well the behavioral classifiers perform when applied to all traffic exchanged on campus, and the extent to which they are able to detect hosts engaging in encrypted BitTorrent. We do so by comparing detection by DPI to detection using the best behavioral method we developed. In the previous section, by applying the behavioral classifiers only to traffic that DPI could classify we found that the Tiered classifier produced the best classification results. In this section we apply the Tiered classifier to the entire data sets for Fall 2007 and Spring 2008 (including DPI's *Unclassified* category) to detect all hosts using BitTorrent, i.e., those using unencrypted BitTorrent, those using encrypted and those that mix

both types. Since our Tiered classifier relies of characteristics of BitTorrent traffic that are not altered by the use of BitTorrent encryption, we assume that its performance in detection of hosts using encrypted BitTorrent is similar to the performance demonstrated in detection of users of unencrypted BitTorrent.

When applied to all traffic detected on campus, the Tiered classifier is able to detect the majority of hosts that DPI detected engaging in unencrypted BitTorrent activity during both the Fall 2007 and Spring 2008 periods. On average per hour, it detected about 85% of the hosts that DPI detected using BitTorrent, which is a higher percentage than that achieved when looking only at traffic that DPI could classify. This is visible in figure 21, which presents classification results obtained using the Tiered classifier in comparison to those obtained by DPI for all traffic detected on campus. The figure summarizes results for Fall 2007 (diagrams a. and c.) and Spring 2008 (diagrams b. and d.), both as hourly averages (diagrams a. and b.) and as figures for the entire monitoring period (diagrams c. and d.). In each Venn diagram, numbers inside each area represent the number of hosts detected (and percentage, out of total hosts seen on campus in parenthesis) by the individual method or combination of methods to which the area corresponds. Numbers under each detection method's name represent the number and percentage of hosts detected by that method alone. Numbers under "Any" represent the total number of hosts detected by the disjunction of all methods, and under "None" represent the number of hosts that were not flagged by any method. The overall count of hosts on campus (hourly average in diagrams a. and b., or total for the whole period in diagrams c. and d.) is presented in the lower right corner of each diagram.



a) Fall 2007 b) Spring 2008

Figure 21. Venn diagrams of classification of all detected traffic using DPI and the Tiered detection methods, displaying number of detected hosts, and percentage of detected hosts out of all hosts with network activity in each period (in parenthesis). Hourly averages for Fall 2007 (a) and Spring 2008 (b).

The Tiered classifier identified a significant number of hosts as engaging in BitTorrent that DPI missed. This presumably includes hosts exchanging encrypted BitTorrent traffic and no unencrypted BitTorrent. However, it also includes some hosts that were incorrectly identified by the Tiered classifier as doing BitTorrent. Figure 13 shows the percentage of hosts identified as engaging in BitTorrent by the behavioral classifier and not by DPI out of all hosts identified as engaging in P2P by any method[78], but with percentages adjusted to take into account the fact that the Tiered classifier yields false positives. This adjustment is based on the classifier's Positive Predictive Value against DPI classification (PPV_{DPI}) obtained from the evaluation against DPI classification of unencrypted BitTorrent (i.e., using only the traffic that DPI could classify, in section 5.3.2.1). The PPV_{DPI} ($PPV_{DPI} = \frac{TP_{DPI}}{TP_{DPI}+FP_{DPI}}$) is the percentage of true positives measured against DPI classification out of all positive classifications yielded by the classifier. The Tiered classifier yielded an average hourly PPV_{DPI} of 70% in both Fall 2007 and Spring 2008, which means that only 70% the hosts detected by the behavioral classifier engaged in BitTorrent. We assume that this PPV_{DPI} remains unchanged when the classifier is applied to all the traffic on campus, and thus assume that 70% of the positive classifications yielded by the behavioral classifier in figure 21 are true positives. Furthermore, we assume that the 30% of positives that are false positives are found among those hosts flagged only by the Tiered classifier and not by DPI[79]. By performing such adjustment we seek to guarantee that the difference between periods in percentage of hosts detected by each method is not an artifact of the classification errors yielded by behavioral classification.

Using the behavioral classifier, we find a significant share of users engaging in encrypted BitTorrent on campus in both monitoring periods, which DPI misses due to the fact that it is not able to detect encrypted BitTorrent traffic. According to figure 22, DPI was able to detect 59% of the hosts that engaged in

[78] We use percentages out of the overall number of hosts detected engaging in BitTorrent (by DPI or behavioral methods) for comparison. This percentage allows for a better comparison than the total number of hosts detected using BitTorrent or than the percentage of hosts detected using BitTorrent out of the total number of hosts on campus, because the total number of hosts with network activity on campus differs greatly between periods. Despite the fact that we attempted to disregard hosts likely to correspond to temporary DHCP leases, we cannot be sure to have identified all those hosts. Hence, we think using what both detection methods could detect in either period as the common ground to calculate the detection figures for each individual method provides for a safer means of comparison between periods.

[79] For instance, out of the 247 positive classifications yielded by the Tiered classifier in Fall 2007, 74 are likely false positives, and thus we account only for 78 hosts that are flagged by the Tiered classifier as true positives (out of the 152 positives that are flagged only by the Tiered classifier).

BitTorrent per hour on average in Fall 2007 and less than 49% in Spring 2008. This means that encryption not only has the potential to defeat detection by DPI, but that it was actually employed by BitTorrent users on campus and prevented DPI from detecting that those users were engaging in BitTorrent activity.

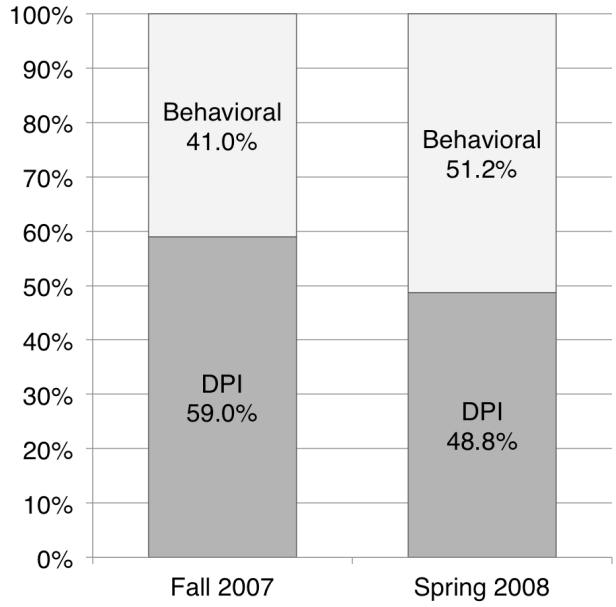


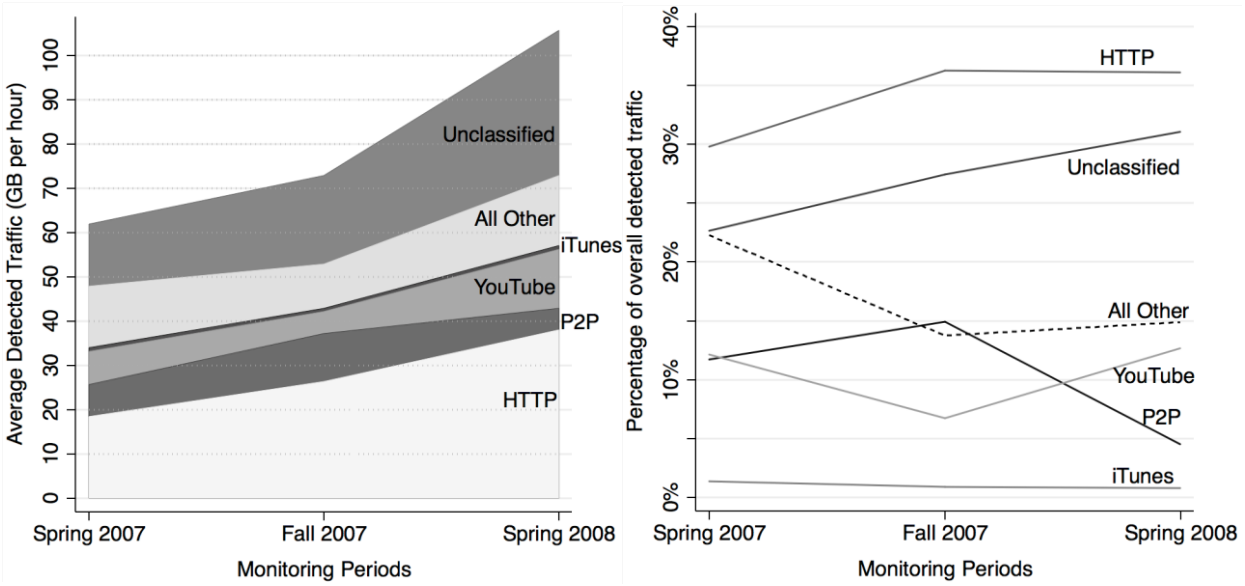
Figure 22. Breakdown of the hosts detected engaging in BitTorrent on campus in Fall 2007 and Spring 2008 between those that DPI could detect and those that DPI could not detect. The portion of the bar tagged "DPI" represents hosts that DPI detected engaging in BitTorrent (some of which were also detected by the behavioral classifier), and the portion tagged "Behavioral" represents hosts that DPI could not detect and that were detected by the behavioral classifier alone. The number of hosts detected by the behavioral classifier was adjusted down taking into account the classifier's Positive Predictive Value.

5.3.2.2.1 Were Users Turning to BitTorrent Encryption?

In this section we bring together evidence that supports the conclusion that P2P users at ISU were increasingly activating encryption. This explains at least in part why the number of users detected engaging in P2P activity decreased from Spring 2007 to Spring 2008, as shown in chapter 4.

As shown in figure 23, while the amount of detected P2P traffic fell from Spring 2007 to Spring 2008, unclassified traffic, which includes encrypted P2P, roughly tripled in volume. If a small fraction of this unclassified traffic is encrypted P2P, this could more than compensate for the decrease in detected P2P.

Although not conclusive, this is consistent with the hypothesis that the decrease in detected P2P users was due to those users adopting P2P encryption.



a. Hourly average exchanged traffic

b. Percentage of overall traffic

Figure 23. Breakdown of traffic exchanged on campus among main detected protocols. a) Average number of bytes exchanged per hour for each category in each period. b) Overall percentage of traffic detected for each category, out of all traffic exchanged in each period.

Behavioral analysis further corroborates this hypothesis, at least for BitTorrent, which is by far the most popular P2P protocol used on campus. As shown by figure 22, the percentage of hosts detected engaging in encrypted BitTorrent (i.e., the hosts detected by the behavioral classifier only) increased from 41% to 51.2% of all hosts detected engaging in BitTorrent on campus.

We conclude that more BitTorrent users were encrypting their traffic in Spring 2008 than in Fall 2007, and this certainly contributed to the decrease in detected P2P users observed in the residence halls (chapter 4). However, we cannot determine from available data whether or not the shift to encrypted BitTorrent would exactly compensate for that decrease in observed (and thus unencrypted) P2P. In part, this is because our estimated trends in use of encryption were based only on BitTorrent traffic, and for computers throughout campus, whereas our estimates of trends in P2P use were based on all P2P protocols (including but not limited to BitTorrent), and only for computers in residence halls. Since these

data sets are not exactly comparable, we cannot tell whether the total number of users that engaged in P2P (encrypted and/or unencrypted) increased or decreased on campus.

5.4 Conclusions and Policy Implications

This chapter analyzes data collected in a university campus network to examine how DPI performs as a tool for detection of P2P activity and transfers of copyrighted content using P2P. Other relevant issues in use of DPI, such as cost and privacy, are outside the scope of this research and were not explored in our analysis. We focus only on effectiveness at detecting P2P activity and/or transfers of copyrighted content using P2P. We find that current DPI technology is effective in detecting users of unencrypted P2P who transfer copyrighted content, but that the technology has several limitations that can diminish its usefulness in the future, should users start employing counter-measures to conceal their P2P activity. We also find that a growing number of users are already employing such concealment measures.

From analysis of data collected using two distinct DPI monitoring appliances in three different periods of about one month each between Spring 2007 and Spring 2008 we find that, after a couple of weeks of monitoring, current DPI technology was able to detect most of the users that attempted to transfer copyrighted content out of network users that could be detected engaging in P2P. Depending on the monitoring period, DPI detected between 40% and 50% of students living on campus engaging in P2P at some point. Out of those, between 75% and 90% were detected transferring files over P2P, and between 70% and 80% were detected transferring content that could be identified as being copyrighted. This means that, at least in the short term, DPI could be an effective tool to assess which network users transfer copyrighted content using P2P, provided that such users do not employ counter-measures, such as traffic encryption, that prevent DPI from observing the contents of P2P transfers. Detecting the majority of users involved in transfers of copyrighted content after some weeks of monitoring may be sufficient for such purposes as putting in place user education campaigns and/or for enforcement.

We found limitations in DPI technology that reduce its effectiveness for purposes other than detecting

users that transfer copyrighted content. Despite missing some individual transfers, DPI detected most users that transferred copyrighted content using unencrypted P2P over one month of monitoring because each user engaged in multiple P2P communication sessions over time and was detected sooner or later. However, the fact that DPI misses individual transfers means that it is not as effective for the purpose of detecting specific media titles transferred over P2P, or which users transferred each media title, because detecting these is much more sensitive to failures in detection of individual communication sessions. As a result, DPI will not be as effective for purposes that require knowing which copyrighted content is disseminated, such as to estimate impact on revenues of copyright holders or to implement a revenue sharing model that would split collected money among copyright holders. In the latter case, the problem would arise if DPI's detection failures were biased towards content from some copyright holders and not others (which is actually the case, given that DPI succeeds at detecting audio content more often than it succeeds at detecting video content).

Another limitation results from the fragmented nature of P2P transfers, which sometimes prevents DPI monitoring appliances from collecting a large enough piece of the content being transmitted in order to attempt a match to a copyrighted title. This is more likely to happen for byte-rich content such as video content. In fact, the DPI appliances used for monitoring were significantly more successful in detecting copyrighted audio than in detecting copyrighted video for titles present in the database of known content. Despite being specific implementations of DPI technology, those appliances were leading products in the market for this type of detection, which makes them good proxies for what DPI can generally achieve. Hence, if DPI is used to collect data on which media titles are transferred using P2P, then its accounts are likely to be significantly biased against video content. This becomes particularly relevant given the growing importance of video content in P2P, which accounts for about 17% of all copies of content transferred using BitTorrent, versus about 18% of copies of audio content, as we show in chapter 6. As a result, if DPI is deployed to detect users for the purpose of copyright enforcement, then such enforcement is likely to benefit the music industry more than the movie/TV industries. This will depend ultimately on the mix of content types that users transfer, particularly on whether people who use P2P to get copyrighted

video content also use P2P for copyrighted music. For instance, users that occasionally go to P2P networks to get video only will be at lower risk of being detected than users that occasionally go to P2P to get music.

If information produced by DPI is to be used for enforcement, then false positive errors need to be taken into account, i.e., errors resulting from communication sessions marked as carrying copyrighted content when that is not the case. False positives are relevant because they may lead to false accusations of copyright infringement. DPI appliances are typically designed to minimize this type of error, and while we could not assess the false positive rate of the appliances we used for monitoring, we could determine that even a small percentage of misclassified communication sessions can lead to a considerable number of users being mistakenly detected. In Spring 2007, if as little as 1% of the communication sessions detected carrying copyrighted content were false positives, then over 10% of users detected transferring copyrighted content would be mistakenly accused. This calls for an assessment of the actual false positive rate of DPI appliances before they are used for enforcement, and if that rate is non-negligible, then design of enforcement procedures needs to accommodate for such errors, for instance by requiring the detection of a minimum number of communication sessions containing copyrighted content before a user is accused of copyright infringement, and by giving users the possibility to appeal when they feel they were wrongly accused.

Finally, one of the main limitations of DPI is that traffic encryption prevents it from observing the content of communications. This means that DPI cannot detect P2P users if they encrypt their P2P traffic, and neither can it detect whether or not those users transfer copyrighted content. Given that the two dominant P2P networks (BitTorrent and Gnutella) support encryption, this limitation can significantly diminish the usefulness of DPI for copyright enforcement should users start to employ encryption at a large scale. Focusing on BitTorrent only, we find that users were already activating encryption on campus in Fall 2007, and that there was a shift in usage from unencrypted to encrypted BitTorrent during the 2007-2008 academic year: the average hourly percentage of BitTorrent users activating encryption out of all detected BitTorrent users increased in 10%, from 41% to 51.2%, between Fall 2007 and Spring 2008. This shift

could have been motivated by several factors. For instance, during the 2007-2008 academic year the most prominent lawsuits against P2P users detected transferring copyrighted content were given significant press coverage[80]. This could have raised users' awareness to the fact that their activity could somehow be under surveillance, and while some of them might have stopped using P2P for copyrighted content, others might have adopted measures they thought would decrease the likelihood of being detected. Encryption could be one of the measures users though to offer that protection, despite the fact that it doesn't actually prevent detection via Swarm Infiltration, the typical method employed to collect information for lawsuits. While we cannot be sure of the reasons that motivated the shift to encrypted BitTorrent on campus, the fact that some users started encrypting even though there was no active enforcement using DPI means that, in the long term, use of DPI for enforcement can drive even more P2P activity underground. We cannot know how many users would switch to encryption in face of enforcement using DPI, but such a switch could significantly hinder the purpose of using DPI in the first place.

The fact that DPI was missing a significant share of BitTorrent activity in the 2007-2008 academic year due to encryption, and that it will possibly miss even more activity as time goes by, implies that DPI alone may not be the most effective solution for online copyright enforcement. Use of DPI together with other methods not vulnerable to encryption, such as behavioral detection methods like the ones we implemented, can be useful to make sure that hosts who escape DPI detection are not completely left out of policies that try to decrease online copyright infringement. For example, detection via DPI could be paired with behavioral detection, so that users of encrypted P2P traffic whose transfers could not be checked for copyrighted content by DPI could at least be identified as users of P2P and, for instance, receive warnings that they might be liable for copyright infringement if they transfer copyrighted content in their encrypted P2P activity.

[80] The most prominent lawsuit for copyright infringement was Capitol v. Thomas, in which Jammie Thomas was condemned in October 2007 to $222,000 in statutory damages for transferring 26 copyrighted songs (http://arstechnica.com/tech-policy/news/2007/10/riaa-first-judgement.ars).

In fact, the behavioral classifiers of BitTorrent traffic that we implemented showed the ability to detect a significant amount of BitTorrent hosts on campus, which shows that they can be useful both as complements to DPI's detection as well as on their own. Our implementations of behavioral traffic classifiers, which were first attempts and thus leave great room for improvement, could detect up to 85% of the users of unencrypted BitTorrent that DPI detected on average per hour, and could also detect users of encrypted BitTorrent. These figures indicate that behavioral classifiers alone may be useful for some applications, for instance, if the purpose is merely to issue warnings to users who may be engaging in BitTorrent about copyright issues, but not to impose any penalties. Such methods present advantages when compared to DPI, such as being able to detect encrypted P2P, and using only traffic summary information, which makes them less expensive to deploy as well as less intrusive in terms of user privacy than DPI. As for counter-measures, it is certainly possible to evade detection by behavioral classifiers, but accomplishing this is arguably not as easy as evading DPI detection, since it may require obfuscating different characteristics of BitTorrent traffic, which may have an impact on the performance of transfers.

6 Assessing the Magnitude of Transfers of Copyrighted Content using BitTorrent

This chapter assesses what and how much content is transferred using BitTorrent, currently the most popular file sharing P2P protocol in use. The chapter has three objectives. The first is to provide a reasonable empirically derived lower bound for the number of copies of copyrighted titles transferred using BitTorrent. This important figure, which shows the extent to which BitTorrent is used for copyright infringement, has not been accurately quantified before.

The second objective is to break that lower bound down into categories depending on characteristics of content transferred. Thus, we can distinguish between transfers that would probably violate copyright law and those that probably would not, as well as estimate the number of copies transferred for distinct content types, such as songs, movies, and software. This makes it possible to compare the number of copies transferred illegally via P2P with the number purchased from a variety of legal outlets, including sale of physical goods (CDs, DVDs), downloads from legal sites, and theater ticket sales. These comparisons put our estimates in perspective, providing a better understanding of how serious an issue copyright violation using P2P still is and of how that varies by content segment. In particular, we find that for many content types, the number of copies transferred illegally via P2P exceeds the number of legal sales by an order of magnitude. Finally, we estimate the number of copies transferred by specific title, and differentiate the more popular titles from the less popular. Looking at the distribution of popularity of transferred media titles tells us the extent to which P2P users prefer content that is popular through legal outlets, and the extent to which they seek less popular titles that may not be widely available in legal outlets. These results can help copyright holders provide more compelling legal alternatives to P2P.

The third objective is to understand which content formats and technical characteristics of content (different methods of video digitalization, video resolutions and audio bit rates) users prefer. Such information can be useful to understand consumer demand, and inform those providing legal media

outlets. It can also help predict how well enforcement technology is likely to perform, since current techniques, like Deep Packet Inspection, are more effective at detecting some content types than others. The effectiveness of such technical mechanisms may influence policy decisions.

This chapter will present the most accurate measure to date regarding how much content is transferred using P2P. Our method estimates both the supply of content, i.e., how many BitTorrent swarms are available with content of different types, as well as the number of copies of that content that is actually transferred by peers connected to those swarms. Previous studies have attempted to quantify how much content is available in P2P (Envisional 2011; Layton and Watters 2010), but failed to estimate how much of that content is actually transferred by users, and the amount of copies of content transferred is the relevant metric when looking at copyright violations performed using BitTorrent. Other studies provide imprecise estimates of overall P2P based on traffic measurements (Labovitz et al. 2009; Sandvine 2009; Cisco 2010; Schulze and Mochalski 2009). Such estimates have two main limitations. First, it is difficult to tell how much P2P traffic there is, and any estimate is inherently dependent on the vantage point of the network from where data is collected. Second, looking at traffic cannot tell what type of content was transferred or whether it was copyrighted or not. This is especially relevant when not all bits of transferred content are valued equally and there are great differences between the sizes of different types of media. For instance, the economic value of a copyrighted movie transferred illegally is different from the value of the same number of bits of copyrighted songs or of proprietary software.

The remainder of this chapter is organized as follows. First we present our estimation methodology and data collection procedures in section 6.1. Next we present obtained results in section 6.2, starting with estimates of how much content is supplied in BitTorrent (section 6.2.1), how much content is transferred using BitTorrent (section 6.2.2), and how much of that content can be found in indexes of legal torrents (section 6.2.3). We then proceed to breaking down the number of transferred copies by different types of media and comparing those figures to sales from legal outlets (section 6.2.4). Next we examine the distribution of popularity of content transferred using BitTorrent (section 6.2.5), and end with the results

section focusing on technical characteristics of that content (section 6.2.6). The chapter concludes with a summary of findings and policy implications in section 6.3.

6.1 Methodology

In order to estimate the rate at which copies of various types of content are transferred using BitTorrent, we estimate the rate at which copies are transferred in each swarm, and then aggregate those estimates for swarms sharing each type of content. We consider the swarms being managed by the two largest public BitTorrent trackers: OpenBitTorrent and PublicBT. These trackers are likely not managing all existing BitTorrent swarms, but they account for a large share of existing swarms and thus allow us to estimate a reasonable lower bound for the amount of content being transferred using BitTorrent.

The transfer rate of each swarm is estimated using equation 1, where $Speed$ is the average transfer speed in bytes per unit of time achieved by each leecher in the swarm, L_{active} is the number of leechers actively downloading content in the swarm, and $Bytes$ is the number of bytes of content shared in the swarm. In this analysis, we assume that all file transfers eventually complete successfully, even if it takes multiple BitTorrent sessions to do so. This assumption generally holds, because BitTorrent clients are designed to automatically resume incomplete transfers upon launch. However there are some cases when leechers abort transfers before obtaining the entire content and do not come back for the rest. We believe such cases are sufficiently unusual that this assumption will not greatly affect estimations.

$$Copies = \frac{Speed \cdot L_{active}}{Bytes} \quad (1)$$

The next sections describe the data collection and estimation processes by which we obtained each of the inputs in equation 1: section 6.1.1 describes the estimation of $Speed$, section 6.1.2 describes the estimation of L_{active}, and section 6.1.3 describes the estimation of $Bytes$.

6.1.1 Estimating the average download rate achieved by a leecher in a swarm

In this section we estimate the average download rate achieved by a leecher in a swarm, which is one of the inputs to equation 1. The download rate achieved by a leecher in a swarm is expected to depend on two main factors. One factor is the speed of the leecher's Internet connection, which imposes a ceiling on the transfer speed that the leecher can achieve. The other factor is the number of peers connected to the swarm, which limits the number of peers from which the leecher can download content. The number of peers in the swarm is expected to influence transfer speed because BitTorrent leechers achieve high download speeds by stacking up several simultaneous downloads from several peers.

Through experimentation, we determine how transfer rates vary with the number of seeders and leechers in a swarm for different Internet connection technologies. We apply regression analysis to measurements of transfer speeds obtained using different technologies in swarms with different numbers of seeders and leechers. This allows us to estimate the average download speed that a leecher using a particular connection technology achieves in a swarm as a function of the number of seeders and leechers in that swarm. The number of copies of content transferred in each swarm can then be determined parametrically using different scenarios for the breakdown of Internet connection technologies among the leechers in the swarm.

Estimation was performed using a data set containing transfer speeds achieved using different Internet connection technologies in swarms with different sizes. We used a set of 20 swarms sharing content that can be legally transferred using BitTorrent, with different numbers of seeders (ranging from 0 to 269) and leechers (ranging from 0 to 67). For each swarm, we collected data from three types of connection technologies and five locations: from two Fiber/LAN connections in university campuses in the US and in Portugal, from two high-speed cable residential connections in the US and in Portugal and from a slower DSL residential connection in Portugal. We downloaded content for each swarm in each location every two hours over the course of a day and once per second recorded the number of seeders and leechers and download speed achieved in the session.

Our analysis shows that the download speed achieved in a swarm depends on the type of technology that the leecher is using to connect to the Internet. As figure 24 shows, the average transfer speed achieved differs between the different locations/technologies. Although there is not much difference between Fiber/LAN and high-speed cable modems in the US or Portugal, probably because they all have sufficiently high capacities that the Internet connection is rarely the bottleneck, it is clear that the use of a slower DSL connection greatly reduces the average download speed.

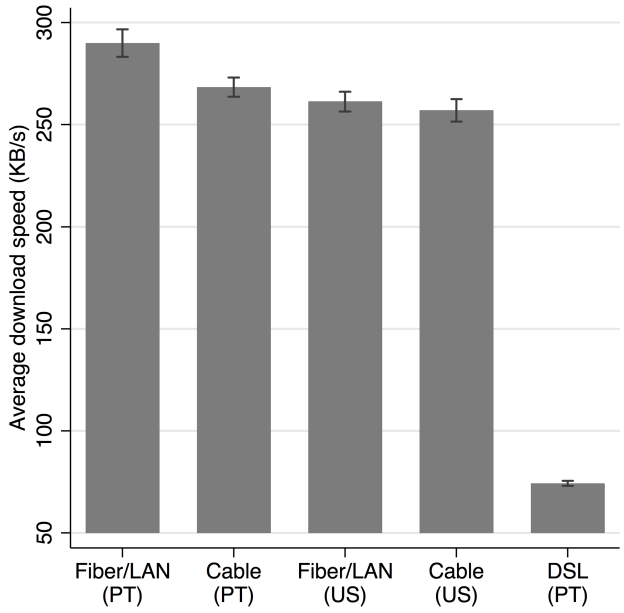


Figure 24. Average download speeds achieved using all monitored swarms for each location/technology monitored.

We also find that download speeds are higher in swarms with higher numbers of seeders and leechers. This is portrayed in table 10, which shows the positive correlation between the download speed achieved and the number of seeders and leechers in swarms. Furthermore, we expect decreasing returns to scale in download speed as swarms get larger, since BitTorrent clients typically define a ceiling for the number of peers from which they can download content at any moment, and that ceiling is independent of the size of the swarm. This means that download speed is expected to vary as a (strictly) concave function of the number of seeders and number of leechers.

Table 10. Correlation coefficients between transfer speeds and number of leechers and seeders in swarms and logarithms of number of seeders and leechers in swarms, for different Internet connection technologies monitored.

	Seeders	**Leechers**
Fiber/LAN	0.38	0.37
Cable	0.45	0.40
DSL	0.23	0.28

In light of the above findings, we use regression analysis to estimate the parameters for the model in equation 2, where download speed varies as a function of the logarithm of number of seeders and the logarithm of the number of leechers in the swarm. We perform separate estimations for each connection technology and obtain the parameter values presented in table 11. Such result parameters allow us to explain between 54% and 65% of the observed variance in transfer speeds, which we consider sufficient for the purpose of estimating of number of copies transferred in the swarms.

$$Speed(s,l) = \beta_s \log(s) + \beta_l \log(l) \qquad (2)$$

Table 11. Estimation results from fitting the model in equation 2 to the data collected using each individual connection type. Each row corresponds to one connection type and presents the number of observations used, the coefficients, significance levels (** means significance at the 1% level) and standard errors (in parenthesis) for each of the dependent variables, and the R^2 obtained for the regression. Estimations were performed with transfer speeds in Bytes/s.

	# obs.	β_s	β_l	R^2
Fiber/LAN	10,016	396,848** (12,036)	348,255** (24,230)	0.54
DSL	5,870	62,853** (1,534)	22,566** (3,299)	0.65
Cable	11,835	337,402** (6,915)	324,643** (14,805)	0.64

The parameters in table 11 are dependent on technology, that is, they allow estimating the average download speed that a leecher using each of the connection technologies will achieve in a swarm with *s* seeders and *l* leechers. In order to remove this dependence, we calculate the number of copies transferred in a swarm (equation 1) using five different scenarios for the breakdown of the swarm's leecher population by Internet connection technologies. Such scenarios are presented in table 12. The "All DSL" scenario assumes all leechers in a swarm have DSL connections and the "All Fiber" scenario assumes all the leechers in the swarm have Fiber/LAN connections. These are the boundary scenarios, which result in the lowest and highest transfer rates respectively. The "Portugal", "USA" and "OECD"

scenarios assume that the breakdown of connection technologies among leechers in a swarm is identical to the breakdown of fixed broadband connections that exists in Portugal, in the USA and in all OECD countries respectively[81]. These three scenarios result in intermediate values for the estimates of number of copies transferred in each swarm. The scenario that is likely to yield the most accurate estimates is the "OECD" scenario, given that it represents the breakdown of connection technologies of a wide range of countries with a high penetration of broadband Internet, which is likely to be more representative of the breakdown of connection technologies for BitTorrent users worldwide than the other scenarios.

Table 12. Scenarios used in estimation of the average transfer speed achieved by a leecher in a swarm. In each scenario, the swarm is assumed to have a breakdown of leechers for each connection technology according to the percentages indicated in each corresponding table cell.

Scenario	Breakdown of leechers in a swarm for each technology		
	Fiber / LAN	Cable	DSL
All DSL	0%	0%	100%
All Fiber	100%	0%	0%
Portugal	7%	41%	52%
USA	5%	55%	40%
OECD	12%	30%	58%

6.1.2 Estimating the number of leechers downloading content in each swarm

In this section we estimate the number of leechers actively downloading content in each swarm, L_{active}, which is one of the inputs to equation 1. BitTorrent trackers report the number of leechers that they know could be connected to each swarm at any given moment. Each of those leechers can be in one of three states. It can be actively downloading content from the swarm, it can be waiting for the desired content to become available for download, or it may have disconnected from the swarm without informing the tracker. We use information collected from trackers to estimate the number of leechers that are effectively transferring content in the swarm, out of those reported by the tracker.

It is possible for a leecher to be waiting for the desired content to become available, e.g. if there are not enough peers sharing the content to satisfy the demand from all the leechers in the swarm, but this is

[81] Data on breakdown of Fixed Broadband connections collected from OECD's "Broadband Subscribers by 100 inhabitants" statistics available at http://www.oecd.org/dataoecd/21/35/39574709.xls

unusual. BitTorrent's *Rarest First* scheduling algorithm prevents this from happening in swarms that have passed the initial ramp-up phase when only the original seeder holds all the pieces of the content (Legout, Urvoy-Keller, and Michiardi 2006). Since this ramp-up state is transient and its duration is typically much smaller than the lifespan of the swarm, for the purpose of calculating how many leechers are actively downloading content, we assume that the number of leechers waiting for content is negligible.

In contrast, the other situation in which a leecher is counted among the swarm but is not downloading, i.e. when the leecher has failed, is too common to ignore. We define failed leechers as those who disconnect from swarms without warning the tracker. This can happen for several reasons, such as users quitting their BitTorrent clients without stopping active downloads, client application crashes, or loss of Internet connectivity. In such cases the tracker takes some time to notice that the leecher has departed[82], and during that period it still accounts for the leecher in the reported counts. We take failed leechers into account and estimate the number of leechers actively downloading content in each swarm, L_{active}, using equation 3, where L_{all} is mean value of the total number of leechers reported by the tracker and L_{failed} is the mean value of the number of leechers that have failed but are still reported.

$$L_{active} = L_{all} - L_{failed} \qquad (3)$$

We obtain L_{all} for each swarm managed by the PublicBT and OpenBitTorrent[83] trackers using the BitTorrent tracker scraping mechanism[84]. At specific time intervals, we requested the list of swarms managed by the tracker and the counts of seeders and leechers for each of those swarms. Such data

[82] When peers depart in a graceful manner the tracker immediately updates its seeder or leecher list (peers depart gracefully when they contact the tracker with a "stopped" announce – as happens when users pause/stop a download in their BitTorrent clients (Cohen 2008))

[83] OpenBitTorrent was the largest public BitTorrent tracker in operation in Summer 2010, managing over 2 million swarms. In September, after an outage of OpenBitTorrent, PublicBT became the most popular tracker. Most of our data was collected from PublicBT.

[84] BitTorrent trackers make available the counts of known seeders and leechers connected to each swarm, which can be easily accessed via an HTTP request to the tracker. This information is used mostly by index websites to compile statistics on how active each swarm is. It is possible to request seeder and leecher counts for each specific swarm, by providing the info-hash that identifies the swarm in the HTTP request, or to request information for all swarms managed by the tracker, i.e., the tracker's "scrape file". We used the last method.

was collected from OpenBitTorrent at 1-hour and 2-hour intervals between August 6 and September 23, 2010, and from PublicBT at 10-minute intervals between November 23, 2010, and February 4, 2011. We switched collection from OpenBitTorrent to PublicBT because OpenBitTorrent phased out its support for the tracker protocol over HTTP in favor of UDP. Since some BitTorrent clients do not support tracker protocol over UDP yet, this made the popularity of OpenBitTorrent decline, and PublicBT took its place as the most popular public BitTorrent tracker[85]. We detected an average of 2.6 million swarms being managed by OpenBitTorrent and an average of 2.7 million swarms being managed by PublicBT at any moment. Overall, we detected close to10 million swarms over the entire data collection.

The main challenge in estimating the number of leechers actively downloading content from each swarm is to estimate L_{failed}, the number of leechers reported by the tracker that have failed. The next section describes how we estimated L_{failed}.

6.1.2.1 Estimating the number of failed leechers

This section details the process of estimating L_{failed}, the number of leechers that are no longer connected to a swarm but that are still reported in the leecher counts obtained from the tracker, which is one of the inputs to equation 3 and allows us to estimate the number of leechers actively downloading content from the swarm. We perform such estimation using a novel method that takes advantage of the fact that the actual removal of failed leechers from the tracker lists happens in bursts at regular time intervals. By observing short-interval variations in the number of leechers reported by the tracker it is possible to estimate the percentage of leechers that already failed but that are still accounted in the tracker counts, which we observed to be relatively constant for different swarms and at different monitoring points. While we detail the estimation in terms of leechers, the same process happens for seeders reported by the tracker, and we present results for seeders obtained using a similar estimation process.

[85] OpenBitTorrent and PublicBT are “twin” trackers that use the same tracker software and present a similar way of operation and even similar websites, so we expect similar behavior concerning peer management from both trackers.

BitTorrent trackers use the following timeout mechanism to detect peers that have failed. Peers need to contact the tracker at least once per pre-defined time interval (t_a, the announce interval) so the tracker knows that they are still operating. To remove failed peers, the tracker performs a cleanup of each swarm's seeder and leecher lists at regular intervals of t_c time units. In each cleanup, a peer is considered to have failed and is removed from the respective list of leechers or seeders if that peer has not communicated with the tracker for a predefined timeout period of t_{to} time units ($t_{to} > t_a$). Hence, every t_c time units, the tracker removes from the seeder and leecher lists for each swarm all the peers that have not contacted in the last t_{to} time units.

We assume that new peers arrive according to a Poisson process. Let λ be the average leecher arrival rate and f represent the probability that a leecher will fail. Let t_f be the average time that a leecher remains in the tracker lists after failure. Assuming these variables are independent, the average number of leechers that have failed but are still accounted by the tracker at any moment, L_{failed}, is given by $L_{failed} = \lambda \cdot f \cdot t_f$. Under these same assumptions, the average number of leechers removed in each cleanup process, L_r, is given by $L_r = \lambda \cdot f \cdot t_c$. Solving these two equations yields that the average number of failed leechers is given by equation 4.

$$L_{failed} = \frac{L_r}{t_c} \cdot t_f \qquad (4)$$

Dividing both sides of the above equation by L_{all}, the average number of leechers in the swarm as reported by the tracker, we obtain equation 5, which allows us to estimate the ratio of the average number of failed leechers to the average number of leechers reported by the tracker (L_{failed}/L_{all}) as a function of the ratio of the average number of leechers removed in cleanup processes to the average number of leechers reported by the tracker (L_r/L_{all}), the time between cleanups (t_c), and the average time that a leecher remains in the tracker lists after failure (t_f). Next, we detail the estimation of these three inputs to equation 5.

$$\frac{L_{failed}}{L_{all}} = \frac{L_r}{L_{all}} \cdot \frac{t_f}{t_c} \quad (5)$$

6.1.2.1.1 Estimating t_c and L_r/L_{all}

We estimate the time between cleanups of the tracker peer lists (t_c) and the ratio of the average number of leechers removed in cleanup processes to the average number of leechers reported by the tracker (L_r/L_{all}) by observing the dynamics of the tracker's peer and seeder counts at short time intervals.

For a diverse set of 500 swarms[86], we queried the PublicBT tracker for the number of leechers and number of seeders for each swarm at time intervals less than 1 second apart[87] during a period of about 24 hours. This yielded a data set with about 18 million observations, at a median rate of one observation every 0.7 seconds for each swarm. The variation over time in the number of leechers reported by the tracker for each swarm indicates that $t_c = 60\ s$, i.e., that the tracker removes peers that failed every 60 seconds. Figures 25.a and 25.b show this by illustrating the leecher dynamics observed for monitored swarms. Figure 25.a portrays the number of leechers in one swarm during an interval of 10 minutes, and makes it is clear that leechers are removed in bursts every 60 seconds. Figure 25.b further corroborates this fact by portraying the distribution of number of seconds between decreases in the number of leechers reported by the tracker, which has a clear peak at 60 seconds corresponding to the burst removals of failed leechers[88].

[86] Selected swarms have a wide range of sizes. Number of seeders ranges from 0 to 16,392 with a median of 44 and mean of 510, and number of leechers ranges from 0 to 8,981 with a median of 27 and a mean of 432.

[87] Since we have no control over the time it takes to transmit our requests to the tracker, for the tracker to respond, and for the response to return to us, we cannot guarantee uniform sampling intervals. However, in most cases we obtained responses that were under 1 second apart.

[88] The graph zooms in to seconds 56 to 64. The PDF was greater than zero for seconds lower than 56, which are not displayed in the graph. These correspond to cases in which the number of leechers that departed gracefully (and that were removed immediately from the list) was greater than the number of new leecher arrivals (thus yielding a negative variation in the overall number of leechers). Higher frequencies around the 60-second mark can be explained by rounding of data collection times.

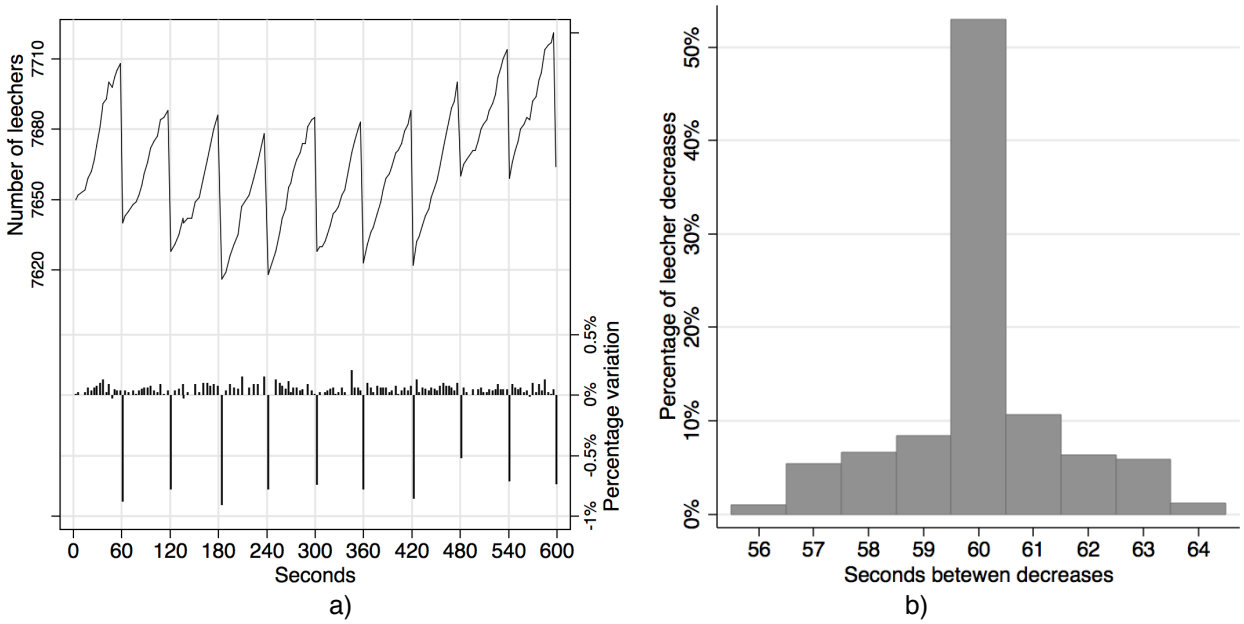


Figure 25. Dynamics of removal of failed peers by the tracker. a) Snapshot of the evolution of the number leechers reported by the tracker for a swarm. The lines on the top of the graph represent absolute numbers reported by the tracker and the bars on the bottom of the graph represent the percentage variation in reported number of leechers from the previous observation. b) PDF of number of seconds between decreases in number of leechers reported by the tracker, detail of seconds 56 to 64.

We calculate the average number of leechers (L_{all}) by averaging across all samples, and we calculate the average number of leechers removed (L_r) by averaging the decreases in number of leechers every 60 seconds across all observed cleanups. The ratio L_r/L_{all} we seek to estimate is fairly constant across swarms of very different sizes, as shown by the narrow confidence interval in table 13, which presents L_r/L_{all} averaged over the 500 swarms for which we collected data.

Table 13. Ratio of the average number of leechers removed in cleanup processes to the average number of leechers reported by the tracker (L_r/L_{all}), averaged across the 500 monitored swarms (95% CI in parenthesis).

	Leechers	**Seeders**
L_r/L_{all}	**0.0105** (0.0102 – 0.0109)	**0.0103** (0.0099 – 0.0107)

6.1.2.1.2 Estimating t_f

In this section we estimate t_f, the mean time between the failure of a leecher and its removal from the tracker lists, the final input needed for equation 5. To do so we use a probability model that incorporates

information about the tracker's timeout mechanism and that assumes that the time until a leecher fails is distributed exponentially.

The two main parameters that influence the tracker's timeout mechanism are the tracker's timeout period, t_{to}, which is the time between a peer's last contact and it's removal from the tracker lists; and the peer's announce time, t_a, which is the maximum time allowed between successive contacts to the tracker from each peer. Both such parameters can be estimated by observing tracker behavior. The announce time, t_a, is set by the tracker, and is communicated to the leecher in the response to every interaction with the tracker. To estimate t_a, we forged announce requests to PublicBT and collected the responses. The resulting t_a were uniformly distributed between 1620 and 1980 seconds (27 to 33 minutes). To estimate the timeout time, t_{to}, we created a swarm sharing a file with random bytes and registered that swarm in PublicBT. We then consecutively collected seeder and leecher counts at short time intervals while sending forged announces for new peers in that swarm, which we would then let timeout. We collected the time difference between the last announce sent by each peer and the moment that peer stopped being counted by the tracker, which yielded an estimate for t_{to} of 45 minutes[89].

Let F be the distribution of time until a leecher fails. We assume that F is exponential with parameter γ, the average leecher failure rate, and the probability density function, $f(x)$, in equation 6.

$$f(x) = \gamma e^{-\gamma x}, \qquad x \geq 0 \tag{6}$$

Let G be the distribution of time between the failure of a leecher and its removal from the tracker list, with probability density function $g(x)$. Clearly, $g(x) = 0$ in its entire domain, except when $x \in [t_{to} - t_a, t_{to}]$, or since $t_{to} = 45$, except when $x \in [45 - t_a, 45]$. For a particular value of the announce time, t_a, the density function, $g(x|t_a)$, is the one in equation 7.

[89] From a sample of 50 observations ranging from 44:02 minutes to 45:00 minutes (mean of 44:44 minutes and median of 44:55 minutes).

$$g(x|t_a) = \begin{cases} \dfrac{f(x-(45-t_a))}{\int_0^{t_a} f(x)\,dx}, & 45-t_a < x < 45 \\ 0, & \text{otherwise} \end{cases} \quad (7)$$

As we determined empirically, t_a is a uniformly distributed random variable that ranges from 27 to 33 minutes. Let $h(x)$ be the density function of t_a, defined according to equation 8 if we consider minutes as the time unit.

$$h(x) = \begin{cases} \dfrac{1}{6}, & 27 \le x \le 33 \\ 0, & \text{otherwise} \end{cases} \quad (8)$$

Assuming that the time until a failed leecher is removed is independent from the announce time, i.e., that G and t_a are independent, then $g(x)$ can be defined in terms of $g(x|t)$ and $h(t)$ according to equation 9.

$$g(x) = \int_{-\infty}^{+\infty} g(x|t)\cdot h(t)\,dt = \begin{cases} \dfrac{1}{6}e^{(45-x)\gamma}(\ln(1-e^{33\gamma}) - \ln(1-e^{27\gamma}) - 6\gamma), & 18 < x < 45 \\ -\dfrac{1}{6}e^{(45-x)\gamma}\left(\ln\left(1-e^{(45-x)\gamma}\right) - \ln(1-e^{33\gamma}) - 12\gamma + x\gamma\right), & 12 < x < 18 \\ 0, & \text{otherwise} \end{cases} \quad (9)$$

Given the above, the mean time that a leecher remains in the tracker lists after failure, which is the parameter we want to estimate, t_f, is simply the mean of the G distribution and can be calculated using equation 10.

$$t_f = \int_{-\infty}^{+\infty} x\cdot g(x)\,dx \quad (10)$$

In order to calculate t_f we need to know the value of the parameter γ, the average leecher failure rate. We calculate that parameter as the average of the ratio between the number of peers removed at cleanup

time and the number of peers in the swarm right before cleanup, calculated over all observed cleanup times for the 500 swarms we monitored at short time intervals (see section 6.1.2.1.1). Estimates for this average are presented in the first column of table 14. Using these estimates as the average leecher failure rate (in fraction of leechers per minute), we obtain the estimates for t_f presented in table 14. The estimates indicate that both leechers and seeders will remain in the tracker lists on average about 29 minutes and 13 seconds after they have failed.

Table 14. Estimates for the average leecher/seeder failure rate and for the mean time that leechers/seeders remain in the respective tracker list after failure.

	γ	t_f
Leechers	0.0106 (0.0102 – 0.0109)	29.21 minutes
Seeders	0.0103 (0.0100 – 0.0107)	29.22 minutes

6.1.2.1.3 Putting it all together

In section 6.1.2 we estimate the number of leechers actively downloading content from swarms. For that, we estimate the fraction of leechers reported by the tracker as being in the swarm but that have already failed using equation 5. The three inputs to that equation, as well as the estimated fraction of failed leechers (and seeders) are presented in table 15. We find that roughly 30% of leechers (and seeders) that the tracker reports as being active in the swarm at any given moment have already failed, and that this percentage is fairly constant across all swarm sizes. It is therefore important, for accuracy of our estimates of the number of copies of content transferred by the leechers in a swarm, to take failed peers into account. Estimates of the actual number of failed leechers for each specific swarm can be obtained by multiplying L_{failed}/L_{all} by the total number of leechers reported by the tracker for that swarm.

Table 15. Estimates for L_{failed}/L_{all}, the ratio of the average number of failed leechers to the average number of leechers reported by the tracker.

	L_r/L_{all}	t_c	t_f	L_{failed}/L_{all}
Leechers	0.0105	1 minute	29.21 minutes	**0.309**
Seeders	0.0103	1 minute	29.22 minutes	**0.302**

6.1.3 Estimating number of bytes of content shared in each swarm and categorizing swarms by type of content shared

This section describes how we estimate the number of bytes of content that is shared in each swarm, which is the final term in the calculation of the number of copies transferred per swarm in equation 1. The section also describes how swarms are categorized by type of content, which allows them to be aggregated in order to calculate how many copies of content of each type are transferred. Both number of bytes and the information necessary for categorization of swarms are obtained from each swarm's ".torrent" file. Next we describe how those data were collected and processed for each swarm.

The qualitative description of the content shared in each swarm is stored in ".torrent" files. For trackers, each BitTorrent swarm is identified by one info-hash, which is a unique digest of the content shared in the swarm that does not describe that content qualitatively. Users of BitTorrent find the content they wish to download by searching for ".torrent" files, which are the key to make the connection between a swarm's info-hash and the qualitative description of the content being shared in that swarm, and which allow for BitTorrent client software to contact the tracker start participating in the swarm.

We obtained ".torrent" files for the swarms whose information we collected from PublicBT by searching multiple torrent index sites[90] using the swarms' info-hashes. Obtained ".torrent" files were parsed in order to extract the relevant information: title of the torrent, number of files and total bytes shared, swarm creation date, and the name, number of bytes and extension of the largest file shared in the swarm[91]. This allowed us to obtain the number of bytes of content shared in each swarm.

Swarms for which it was possible to obtain torrent files are categorized by type and other characteristics of content in a second stage of processing. In this stage, we parse the title of the torrent to extract content

[90] The BitTorrent index sites that we searched were: Zoinc.com, Torrage.com, Torcache.com, IsoHunt.com and Torrentz.eu (this last one is a meta-index that aggregates information from over 30 BitTorrent indexes – http://torrentz.eu/help)

[91] Most torrents are composed by one or a few large files with the actual content and then a few smaller files containing descriptions or auxiliary data. The largest file shared in a swarm is likely the content that users are actually seeking to download. Therefore, the name of that file is typically a good description of the actual content users seek.

characteristics such as the actual title of the content (the title of the movie, for instance) and keywords typically included in torrent titles that indicate technical characteristics of the content, such as the type of content (song, movie, TV show, software, adult content, book, etc.), encoding (mp3, aac, divx, ogg, mkv, etc.) or quality (128kbps, 256kbps, 480p, 720p, 1080p, etc.). This is done via a semi-automated process that we briefly describe next.

In a pre-processing stage, we compiled a list of keywords that appear in the torrent titles and assessed the specific meaning of each of them by manually categorizing it concerning what characteristic of the content it refers to and whether or not it indicates a specific type of content[92]. This list of keywords was used in the categorization process, which ran in three phases. In phase 1 it categorized torrents whose titles have very specific structures, such as certain movies and TV shows[93]. This phase managed to categorize about 20% of swarms. For remaining swarms, phase 2 used a voting system to infer the type of content in the swarm from the keywords present in the title of the torrent. Each keyword found in the torrent title that indicated a type of content was accounted and the swarm was categorized under the type of content with most keywords. This phase managed to categorize about 25% of the swarms, but it clearly could not categorize swarms whose titles did not contain keywords or whose keywords were ambiguous (for instance, “dvdrip” can indicate both a movie or a TV show that was obtained by ripping a DVD). Remaining swarms followed to phase 3, which characterized content by analogy to titles categorized in previous phases. This was done by matching the content titles identified in the previous phases to torrent titles not categorized yet. For instance, the torrent title “How I met your mother.rar” is easily identified as long as there is one episode of that TV show that can be identified in the previous phases. This phase allowed for categorization of about 10% of swarms.

[92] For instance, the keyword “camrip” refers to the method by which the content was captured and digitalized (using a camera in a movie theatre) and indicates that the content is a movie, the keyword “128kbps” refers to the bitrate at which the content was encoded and indicates audio content.

[93] For instance, one typical structure for formatting titles of TV show torrents is “Title.SXXEXX.Episode.title.Keywords.Extension” (an example is “Entourage.S07E05.Bottoms.Up.HDTV.XviD-FQM.avi”, where “Entourage” is the TV show title, the particular episode being shared in episode 5 of season 7, the episode title is “Bottoms Up”, it is an episode collected from ripping “HDTV” content, encoded in “xvid” format, shared by the “FQM” team, and packed in an “avi” container).

Using the categorization process described above we were able to infer the type of media of slightly over 50% of the swarms for which it was possible to collect torrent information. For the remaining swarms the only information available is given by the file extension, which indicates very generally the type of content. For instance, while it is safe to assume that “mp3” or “aac” files are found in swarms sharing music content, “avi” or “mkv” files indicate video files that can correspond to movies, TV shows, adult content or other types of video.

6.2 Results

This section presents our estimates of the amount of content made available and transferred using BitTorrent and characterizes various aspects of that content. We start by estimating the amount of content made available in BitTorrent broken down by types of media as a way of assessing content supply in BitTorrent (section 6.2.1). Next we estimate how many copies of content are effectively transferred per day, a figure that had not been well characterized before, and that is relevant when considering BitTorrent from a copyright infringement perspective (section 6.2.2). In section 6.2.3 we estimate the amount of content transferred in BitTorrent that would not result in copyright violations, and in section 6.2.4 we compare our estimates of number of copies of copyrighted content transferred using BitTorrent to legal sales figures for music and movies to put the amount of copyright infringement in BitTorrent into perspective. In section 6.2.5 we look at the relative distribution of popularity of content transferred using BitTorrent as a way to understand whether users seek popular content or less mainstream media. Finally, section 6.2.6 examines what characteristics of content BitTorrent users prefer, which can be useful for those seeking to provide legal alternatives to P2P and can also influence the performance of technology for detection of transfers of copyrighted content.

6.2.1 Content Supplied in BitTorrent

In this section we characterize the supply of different types of content in BitTorrent, measured as the number of swarms detected sharing content. We compare supply of different types of media by breaking down the number of detected swarms by the type of media shared in each of them, which tells us how many bundles of content are shared for each type of media.

In the data we collected from the largest public BitTorrent trackers during 115 days between August 2010 and February 2011, we found an average of 2.6 million swarms offering content at any moment, which added up to a total of close to 10 million swarms offering content at some point in the period. These are lower bounds on the number of existing BitTorrent swarms, which means that there was a supply of at least 2.6 million bundles of content for download in BitTorrent at any moment.

To understand which types of media are most supplied in BitTorrent we aggregate swarms by the type of content shared. Such aggregation is performed using both the file type of the largest file shared in the swarm (audio, video, software, archive, etc.) and the actual type of media shared in the swarm obtained using the categorization process described in section 6.1.3 (movie, tv show, song, music album, etc.). File types provide a general idea of the content in each swarm, but fail to provide important distinctions, such as between different types of video content (movies, TV shows, music videos) or between different media types shared within archives. We present breakdowns by both type of file and type of media because it was possible to gather information on file types for 74% of detected swarms, whereas we could only obtain specific type of media for 52% of those (corresponding to 39% of detected swarms).

Focusing on file types, for which the breakdown of swarms is presented figure 26.a, we observe that the type with highest supply in BitTorrent is by far video (45% of swarms), with a greater percentage of swarms than that of the second and third most supplied types of files together, which are audio and archives. Focusing on actual types of media, figure 26.b, shows that movies have the highest supply in BitTorrent (39% of swarms), followed by music albums, TV show episodes and then by software. When compared to previous estimates of supply of content in BitTorrent (Envisional 2011), we find similar

percentages of movie and TV show swarms (previous estimates report 32% of Films and 13% of Television), but we find much lower percentages of adult content swarms (previous estimates report 36% of swarms sharing Pornography) and much higher percentages of music and software swarms (previous estimates report 3% of Music and 4% of Software swarms). Nevertheless, our results qualitatively confirm previous estimates that indicated video as the most supplied type of content in BitTorrent.

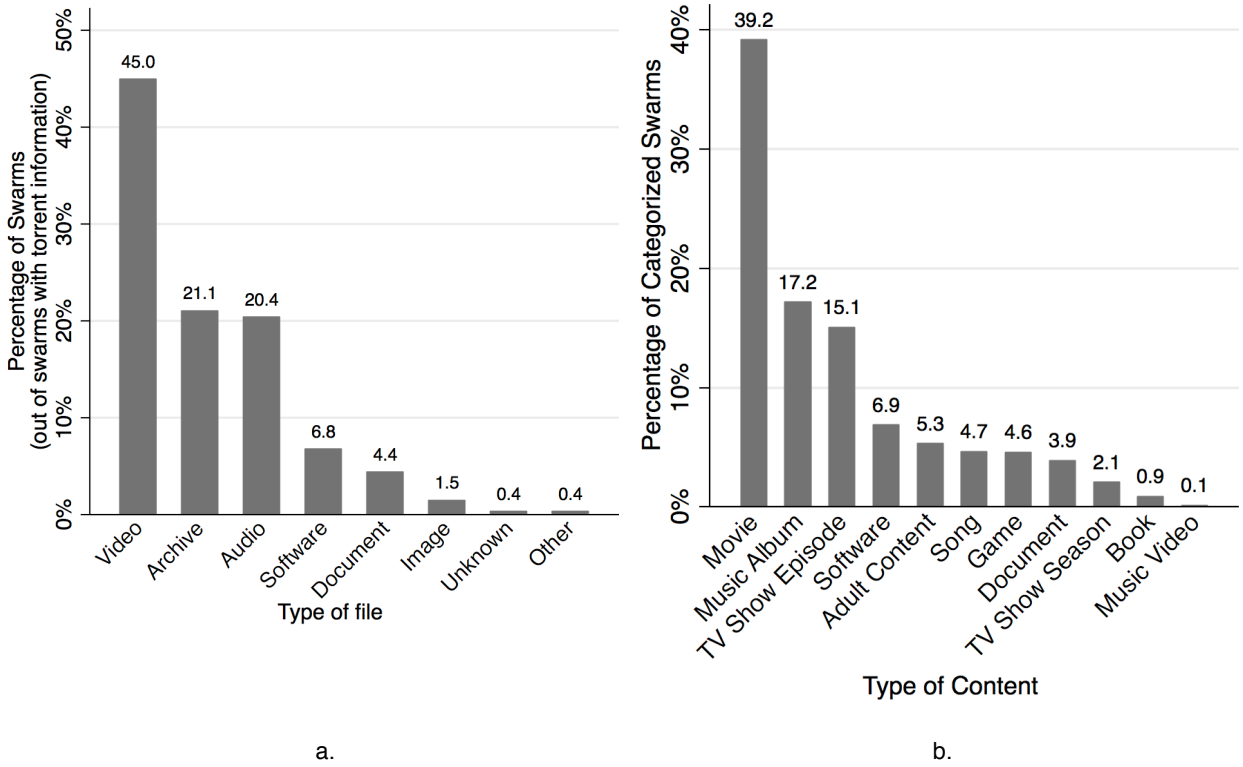


Figure 26. Breakdown of supply of content in BitTorrent by percentage of swarms sharing content of different types. a) Breakdown by file type. b) Breakdown by media type.

6.2.2 Content Transferred using BitTorrent

In this section we estimate the number of copies of content transferred using BitTorrent and break that number down by the type of media transferred. Such figures are important because it is the number of copies of copyrighted content transferred without permission from copyright holders that matters when considering the amount of copyright violations that occur using BitTorrent. While previous studies have

estimated the supply of different types of content in BitTorrent, as we estimated in the previous section, figures for actual number of transferred copies have not been well characterized before.

To calculate the overall number of copies of content transferred in each swarm during each day we estimate the instantaneous number of copies transferred per unit of time at each monitoring point according to the methodology described in section 6.1. We then interpolate such figures for each swarm over the several monitoring points during the day.

Estimates of the average number of copies transferred in all swarms with torrent files (74% of all monitored swarms) per day are presented in figure 27 for the different scenarios of leecher connection technology. The different scenarios result in a wide range of estimates, with the estimate for “All Fiber” being over 8 times higher than the estimate for “All DSL”. These two scenarios are not realistic, but are useful as boundaries for our estimates. Intermediate scenarios yield estimates that are much closer together and are more realistic since they were obtained from actual breakdowns of connection technologies in different geographical areas. We consider the OECD scenario as the one likely to provide the most accurate estimates because it represents the breakdown of connection technologies of a wide range of countries with a high penetration of broadband Internet, which is likely to be more representative of the breakdown of connection technologies for BitTorrent users worldwide than the other scenarios. We will use OECD estimates in the remainder of this chapter except when stated otherwise. By that account, the swarms with torrent information that we monitored transferred over 800 million copies of content on average per day.

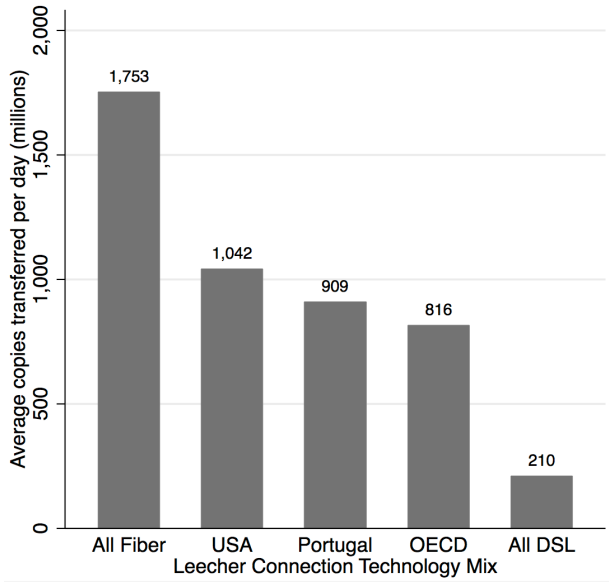
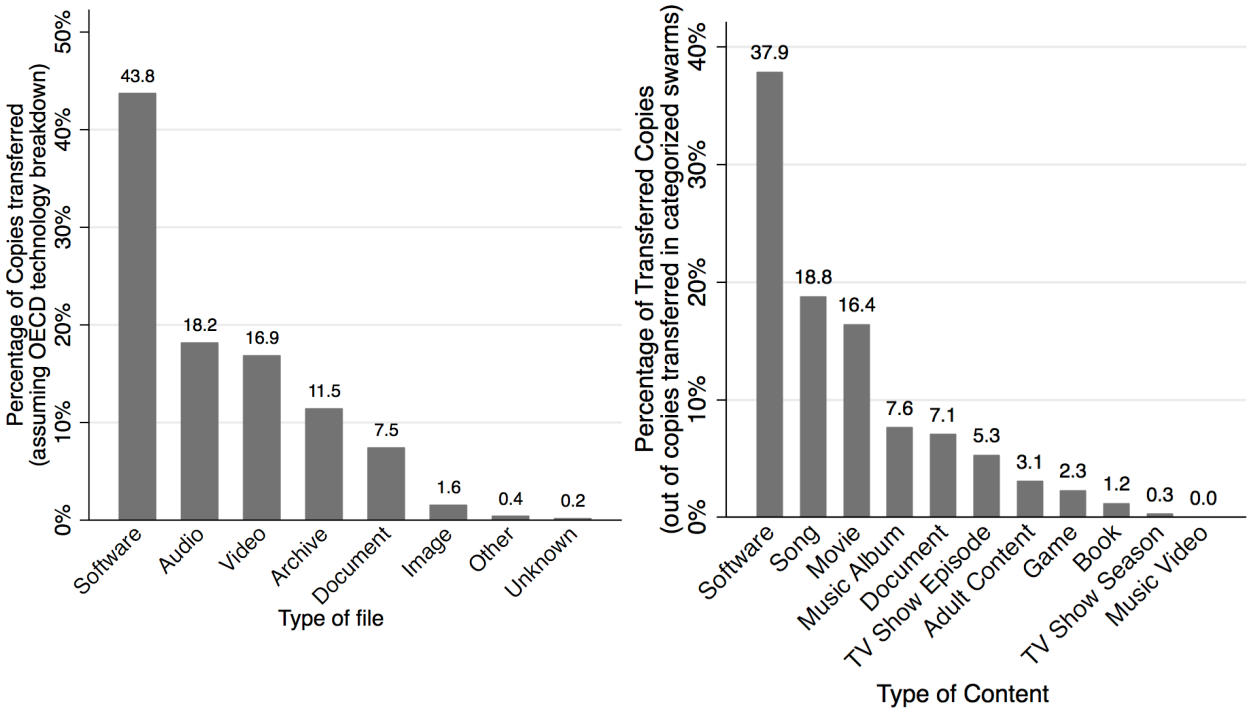


Figure 27. Estimates of overall number of copies of content transferred per day by all monitored swarms with torrent information using the different scenarios of leecher connection technology mixes.

We break down the number of copies transferred by the type of file transferred in each swarm for the 74% of swarms with torrent information, and by specific type of media transferred in each swarm for the 52% of those swarms for which we could collect media type information. Such breakdowns are presented in figure 28. Figure 28.a shows that the file types for which more copies are transferred correspond to software, accounting for close to half of all copies of content transferred, followed by audio and then closely by video. Looking at actual media types in figure 28.b shows that software is the type of media with more transferred copies, accounting to close to 38% of all transferred copies, followed by individual songs, movies and music albums.



a. b.

Figure 28. Breakdown of percentage of copies transferred using BitTorrent by type of content transferred. a) Breakdown by file type. b) Breakdown by media type.

A comparison of the breakdowns of content supplied in BitTorrent (figure 26 in the previous section) with breakdown by actual number of copies transferred (figure 28 above) shows that, despite being only the fourth most supplied type of content, software is actually the type of content most downloaded in BitTorrent. Furthermore, movies and TV shows are not the types of content for which most copies are transferred using BitTorrent, despite being the principal types of content supplied and being considered by many accounts the leading types of content shared in BitTorrent. Given their byte content, movies and TV show swarms are likely to dwarf other types of media in terms of traffic consumed by transfers, but in terms of copyright infringement, the number of copyright violations performed using BitTorrent is likely higher for other types of content (software and music, for instance) than for movies and TV shows.

The difference presented above between the most popular types of files in terms of supply and in terms of actual copies transferred demonstrates the main shortcoming in previous studies (Envisional 2011;

Layton and Watters 2010) that estimated BitTorrent activity by looking only at breakdown of swarms or number of peers connected to each swarm. Such are estimates of supply of content in BitTorrent, which, as demonstrated above, are not accurate when describing the actual transfers of content that occur. The difference between supply and demand figures is due to the scaling role performed by the number of bytes of content shared in each swarm. As table 16 shows, each audio, video and archive swarm shares respectively on average 2, 8 and 6 times more bytes than a software swarm, meaning that for similar numbers of peers connected to swarms, copies of software would be transferred on average twice, 8 and 6 times faster than copies of audio, video and archives.

Table 16. Average and median number of bytes shared in swarms containing files of each different type.

	MB per swarm	
	Median	**Average**
Software	11	210
Audio	131	470
Video	701	1769
Archive	357	1341
Document	11	158
Image	30	195
Other	226	908
Unknown	303	1804

6.2.3 Content that can be Legally Transferred using BitTorrent

In this section we estimate the number of swarms sharing content that can be legally transferred using BitTorrent and the number of copies of that content transferred on average per day. We identify whether each of the 10 million swarms we detect contains content that can be legally shared in BitTorrent by searching for the swarm's info-hash in the most popular BitTorrent index websites specialized in hosting torrents for legal swarms: mininova.org, legittorrents.com, youtorrent.com, linuxtracker.com and clearbits.com.

The index websites that we searched publicly declare to actively filter out content that cannot be legally transferred using BitTorrent. Mininova.org[94], legittorrents.com and youtorrent.com are general-purpose

[94] Mininova.org was in the past one the largest BitTorrent index websites, hosting torrents for copyrighted content as well as content that could legally be transferred using BitTorrent. In late 2009, after a court order, Mininova started to actively filter out torrents "if

index websites that filter out copyrighted content, Linuxtracker.com specializes in indexing torrents for swarms containing the Linux OS, and Clearbits.com specializes in hosting and distribution of open licensed media. While there are likely more swarms containing content that can be legally transferred using BitTorrent than those found in these index websites, these websites are the only indexing resources available for BitTorrent users who want to be sure they are legally downloading content from BitTorrent. Other (more popular) index websites contain torrents for copyrighted content alongside possibly torrents for content that can be legally transferred, so users that get torrents from those websites cannot be sure if their transfers are going to be legal or illegal.

Out of the close to 10 million swarms detected in our monitoring, we found 13,231 swarms whose torrents were indexed in the websites mentioned above. As table 17 shows, such swarms correspond to under 0.2% of all detected swarms and the number of copies of content transferred represents about 0.02% of overall transferred copies of content (both significantly different from 0 at the 5% level). Hence, despite the effort from these index websites to promote legal transfers in BitTorrent, the number of copies of those legal titles transferred is close to insignificant in practical terms when compared to the number of transferred copies of titles indexed by the more well known general-purpose indexes, many of which are likely copyrighted content whose transfer using BitTorrent is unlawful.

Table 17. Comparison of percentage of swarms and number of transferred copies between swarms sharing content that can be legally transferred using BitTorrent (found in mininova.org, legittorrents.com, youtorrent.com, linuxtracker.com or clearbits.com) and other swarms for which it was possible to find torrent files (which likely contain copyrighted content that cannot be transferred legally using BitTorrent).

	Content found in indexes	**Content found in other indexes**
Percentage of swarms	0.184%	99.816%
Percentage of transferred copies	.023%	99.977%

Concerning specific types of content shared, the breakdown of supply and of transferred copies in swarms found in the indexes of legal content that we considered presented in figure 29 shows that, despite the effort from these websites to promote legal video and audio, it is clear that BitTorrent users

there is reasonable doubt that the actual content contains copyrighted works" (http://blog.mininova.org/articles/2010/12/10/brein-mininova-settlement-reached-lawsuit-ended/)

are not interested in obtaining those legal audio and video titles. When it comes to legal content, BitTorrent users transfer mostly documents and software. This is unlike what we observe when considering all monitored swarms, where the great majority of shared titles are likely copyrighted content whose transfer using BitTorrent constitutes copyright infringement. In this case audio and video are among the types of content with the greatest supply (figure 26.a), and BitTorrent users are interested in obtaining those audio and video titles, as shown by the fact that audio and video are also among the types of content with the highest number of transferred copies (figure 28.a).

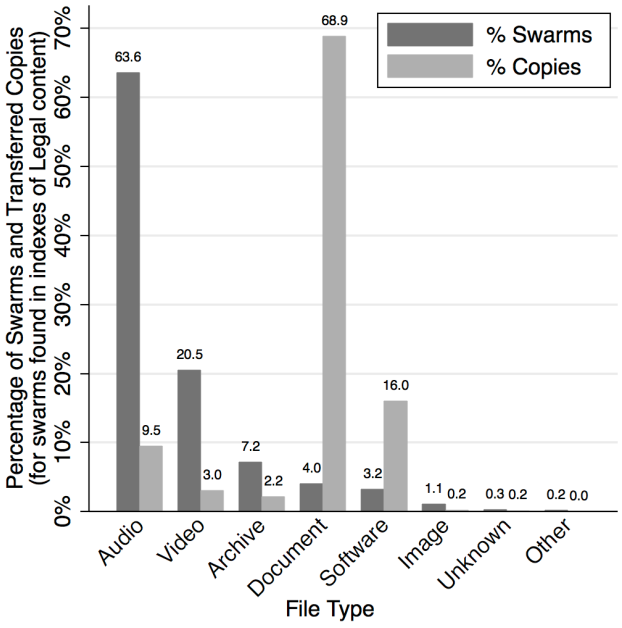


Figure 29. Breakdown by type of file of supply and number of copies transferred from swarms detected sharing content that can be legally transferred using BitTorrent.

6.2.4 Comparison of Transfers of Copyrighted Content to Legal Sales

This section compares estimates of the number of copies of copyrighted content transferred per day using BitTorrent to legal sales of content. This allows us to put those estimates in perspective and understand how the BitTorrent "market" for content compares to the legal market. We compare overall number of transferred movies, songs and albums to worldwide sales of corresponding media types. We also compare worldwide legal sales for each title in the top 10 bestselling songs, albums and movies in

theatres worldwide and DVDs in the U.S. to copies transferred in the swarms sharing each of those titles in BitTorrent.

We calculate the average daily number of copies of copyrighted movies, songs and music albums transferred using BitTorrent by adding up the estimated number of copies per swarm for all swarms categorized as sharing content of each of those types. This means that our estimates are lower bounds, since it was not possible to find the type of content shared in all detected swarms. However, swarms that could be identified as sharing movies, songs and albums are very likely sharing copyrighted content because the process used to identify the type of content relied heavily on the use of keywords that either indicate content that was obtained by illegal methods (e.g., “camrip”, “dvdscreener”) or by copying it from purchased media (e.g., “dvd rip”, “cd rip”).

We compare number of transferred copies per day to daily averages for 2010 of movie tickets sold in the USA and worldwide, of sales of DVDs and Blu-Rays in the USA, and of sales of digital singles and sales of music albums in the USA and Canada, and worldwide. We obtained figures for theatre tickets, digital singles and music albums sold in the USA and Canada from Nielsen Soundscan[95], and figures for DVD and Blu-Ray sales were obtained from The Digital Entertainment Group (Schaefer 2011). The number of units sold worldwide was calculated from reported revenue of the respective industries, and correspond to estimates with some margin of error. In the case of movie tickets, worldwide box-office receipts for 2010 reached $31.8 billion (MPAA 2010), which, assuming ticket prices at the 2010 US average of $7.89 (MPAA 2010), yields an estimate of 11 million tickets sold worldwide per day. For music, the IFPI reports that the digital music market has a trade value of $4.6 billion and represents 29% of the industry’s revenue (Moore 2011), which means that the revenue of the “physical” part of the market was about $11.3 billion. Assuming all revenue comes from music sales, that a digital song costs on average $1.2, a

[95] Unit sales for digital singles and music albums in the USA and Canada were obtained from Nielsen Soundscan (http://www.businesswire.com/news/home/20110106006565/en/Nielsen-Company-Billboard%E2%80%99s-2010-Music-Industry-Report)
Number of movie theatre tickets sold in the USA was obtained from http://www.the-numbers.com/market/2010.php.

digital album costs $10 and a physical album costs $15[96], and that albums contain on average 10 songs, we estimate sales of 3.3 million albums per day (physical + digital), which correspond to 33.4 million songs per day (physical + digital).

For movies as well as for music, the number of copies transferred using BitTorrent is many times greater than legal sales. The comparisons for the different types of media are illustrated in table 18. In the case of music, the number of single songs (i.e., songs transferred from swarms that contain a single song) transferred using BitTorrent worldwide was many times greater than the number of digital music singles sold in the USA and Canada, and the same happened for transfers of music albums when compared to album sales in the USA and Canada. However, since P2P is used worldwide, transferred copies should be compared to worldwide sales. Assuming that each music album transferred using BitTorrent contains on average 10 songs, the overall number of songs transferred using BitTorrent (either individually or as part of an album) was 13.1 times greater than estimated worldwide sales of songs (either digital or as part of physical media).

In the movie market, the number of movies transferred using BitTorrent was also many times greater than the number of movie theatre tickets or the number of DVDs and Blu-ray disks sold in the USA. Once again, worldwide sales provide a more relevant comparison. A comparison of BitTorrent movie transfers to box office sales shows that the number of movies transferred using BitTorrent was 6.8 times greater than worldwide box-office sales.

Focusing on the most popular titles in terms of legal sales during the period we monitored, we are able to establish a title to title comparison of worldwide sales to copies transferred using BitTorrent for the worldwide top 10 selling music singles[97], music albums[98], and top 10 box-office grossing movies[99], and for

[96] Average prices for digital songs and albums, and physical albums calculated using RIAA 2010 year-end sales figures for the U.S. market available at http://www.riaa.com/keystatistics.php?content_selector=2008-2009-U.S-Shipment-Numbers

[97] Top 10 music singles list compiled using weekly data available at http://www.mediatraffic.de

[98] Top 10 music albums list compiled using weekly data available at http://www.mediatraffic.de

[99] Top 10 grossing movies list compiled using weekly data available at http://www.the-numbers.com/movies/international/weekly.php

the U.S. top 10 selling DVDs[100]. Figures for BitTorrent transfers for each song, album or movie were obtained by adding up the number of transferred copies in all swarms whose torrent name or file names match that title. Figures for sales were obtained by merging weekly sales data for each type of media in the same weeks for which we collected BitTorrent data. This comparison is presented in tables 19 through 22, which show that BitTorrent transfers greatly exceed legal sales for the vast majority of the top-10 titles in each of the media types considered. The tables show that sales ranks are typically higher than BitTorrent ranks for the top 10 sales titles. This means that choosing the top sales titles to compare to BitTorrent transfers will yield smaller transfers to sales ratios than those that would be obtained if the comparison were done for the top transferred titles.

Table 18. Comparison between estimated daily number of copies of content transferred using BitTorrent for the swarms whose content could be categorized and sales figures for equivalent content types.

	BitTorrent transferred copies (M = millions)	USA Market		Worldwide Market	
		Sales (M = millions)	Transfers to Sales Ratio	Estimated Sales (M = millions)	Transfers to Sales Ratio
BitTorrent Movies vs. Box Office	**75.4 M** movies	**3.6 M** tickets	20.8	**11 M** tickets	6.8
BitTorrent Movies vs. DVD and Blu-ray	**75.4 M** movies	**4.6 M** DVD + Blu-Ray discs	16.4		
Music Singles	**86.3 M** single songs	**3.2 M** digital singles*	26.9		
Music Albums	**35.1 M** album bundles	**0.89 M** dig. + phys. albums*	39.3	**3.3 M** dig. + phys. albums	10.5
Songs	**437.3 M** single + bundled songs	**12.1 M** dig. + phys. songs*	36.0	**33.4 M** dig. + phys. songs	13.1

* Figures for singles, albums and songs include the USA and Canada markets.

Considering music titles, both singles in table 19 and albums in table 20, we can see that BitTorrent transfers exceed sales by over an order of magnitude for most titles. In the particular case of music albums, we observe a large variation of the transfers to sales ratio between titles. One possible explanation for this variation comes from the nature of the media transferred and the demographics it typically appeals to. Clearly, the transfers to sales ratio is greater for music albums of pop artists whose music caters to a teenager and young adult audience that is typically Internet-savvy as well. In comparison, albums that perform well in sales but not so well in BitTorrent are those that typically cater to

[100] U.S. top 10 DVD list compiled using weekly data available at http://www.the-numbers.com/dvd/charts/weekly/thisweek.php

an older audience, who may not know how to transfer content from P2P, may not be willing to do it because they know it is illegal, or may have higher willingness to pay for legal content. This hypothesis is also corroborated by the figures comparing DVD sales to movie transfers in table 22. In this case, the titles that perform worse in BitTorrent when compared to legal sales are mostly content destined for children, whose parents likely belong to the older audience that prefers to purchase content instead of transferring it from P2P.

Relevant implications for business and enforcement can be drawn if the difference in ratios of BitTorrent transfers to sales is indeed the result of different demographics having different propensity to transfer content from BitTorrent. It may be possible to predict which titles in copyright holders' catalogs are more likely targets of illegal sharing, and thus estimate the extent to which sales of those titles will be affected by online copyright violations. Titles with higher transfers to sales ratios are those likely to appeal to teenagers and young adults, an important segment of the population whose members are typically avid consumers of media, but who, at the same time, may have less willingness to pay or disposable income to purchase such media. This segment of the population has in P2P a free, yet illegal, alternative, which they seem to be taking advantage of. Copyright holders can use this information to try to drive those consumers away from P2P, either by deploying selective enforcement focusing on the titles that typically appeal to those demographics, or by further investigating which factors drive such consumers away from purchasing content in order to devise more compelling legal alternatives.

Table 19. Comparison of worldwide sales of music singles to number of copies transferred using BitTorrent for the top 10 most sold music singles during the monitoring period (sales and transfers in thousands).

Artist	Title	Sales		BitTorrent Transfers		Ratio of transfers to sales
		Rank	Average daily	Rank	Average daily	
Eminem feat. Rihanna	**Love The Way You Lie**	1	39.3	13	768.7	19.5
Bruno Mars	**Just The Way You Are**	2	35.3	17	616.7	17.5
Taio Cruz	**Dynamite**	3	34.9	22	429.8	12.3
Rihanna	**Only Girl (In The World)**	4	32.4	15	673.5	20.8
Katy Perry	**Teenage Dream**	5	31.3	34	261.1	8.3
Usher feat. Pitbull	**DJ Got Us Fallin' In Love**	6	28.8	29	355.7	12.4
Flo Rida feat. David Guetta	**Club Can't Handle Me**	7	27.4	9	965.0	35.3
Katy Perry feat. Snoop Dogg	**California Gurls**	8	24.6	1	4159.3	169.1
Nelly	**Just A Dream**	9	22.5	39	205.7	9.2
Katy Perry	**Firework**	10	21.7	68	126.2	5.8

Table 20. Comparison of worldwide sales of music albums to number of copies transferred using BitTorrent for the top 10 most sold music albums during the monitoring period (sales and transfers in thousands).

Artist	Title	Sales		BitTorrent Transfers		Ratio of transfers to sales
		Rank	Average daily	Rank	Average daily	
Eminem	**Recovery**	1	20.1	1	552.7	27.4
Susan Boyle	**The Gift**	2	17.2	111	36.7	2.1
Taylor Swift	**Speak Now**	3	14.8	24	133.4	9.0
Rihanna	**Loud**	4	14.6	2	484.9	33.2
Katy Perry	**Teenage Dream**	5	13.6	7	361.5	26.5
Take That	**Progress**	6	13.0	74	60.0	4.6
Justin Bieber	**My Worlds**	7	12.4	6	364.2	29.3
Bon Jovi	**Greatest Hits**	8	10.5	76	59.0	5.6
Kings Of Leon	**Come Around Sundown**	9	10.5	65	64.7	6.2
Lady GaGa	**The Fame Monster**	10	9.5	3	403.5	42.4

Focusing on movies, despite the fact that BitTorrent transfers exceed sales for the vast majority of titles, we observe much smaller transfers to sales ratios than for music in the case of the comparison to U.S. DVD sales (table 22) and especially in the case of the comparison to worldwide box-office ticket sales (table 21). One possible explanation for the low ratios of BitTorrent transfers to box-office sales is the fact that we are comparing sales of the most popular movies at the moment (since top selling movies in theatres are typically recent movies whose popularity peaks in the first couple weeks after release) to transfers of lower quality copies of those movies obtained by filming movie theatre screenings (since we monitored a relatively small period during which the majority of those movies were not yet released for sale in DVD format, and thus the versions available in BitTorrent were typically cam rips[101]). BitTorrent users do not seem to be very interested in obtaining the low quality copy, and in fact, by looking at the ratios and DVD release dates of the titles in table 21, we see that movies whose DVD release was within our monitoring period (and for which there were DVD-rip high-quality copies available in BitTorrent) typically attained higher transfers to sales ratio than those that were only in theatres during that period.

[101] See Appendix B.

Table 21. Comparison of estimated worldwide box-office ticket sales to number of copies transferred using BitTorrent for the top 10 box-office movies during the monitoring period (sales and transfers in thousands).

Title	Sales		BitTorrent Transfers		Ratio of transfers to sales	Release Date	
	Rank	Average daily	Rank	Average daily		Theatrical	DVD
Harry Potter and the Deathly Hallows	1	658.0	5	641.6	**0.98**	11-19-10	04-15-11
Inception	2	352.2	1	1007.7	2.86	07-13-10	**12-07-10**
Tangled	3	296.8	20	436.4	1.47	11-24-10	03-29-11
Tron: Legacy	4	242.2	10	530.2	2.19	12-17-10	04-05-11
Despicable Me	5	238.7	28	329.3	1.38	06-27-10	**12-14-10**
Megamind	6	199.1	4	668.3	3.36	10-30-10	02-25-11
Little Fockers	7	192.7	22	383.6	1.99	12-22-10	04-05-11
Toy Story 3	8	171.6	24	375.9	2.19	06-17-10	**11-02-10**
Narnia: The Voyage of the Dawn Treader	9	170.5	127	38.0	**0.22**	12-10-10	04-08-11
Resident Evil: Afterlife	10	161.9	17	476.4	2.94	09-10-10	**12-28-10**

Table 22. Comparison of U.S. sales to number of copies transferred using BitTorrent for the top 10 most sold DVDs during the monitoring period (sales and transfers in thousands).

Title	Sales		BitTorrent Transfers		Ratio of transfers to sales	DVD Release Date
	Rank	Average daily	Rank	Average daily		
Toy Story 3	1	51.6	22	383.6	7.4	11-02-10
The Twilight Saga: Eclipse	2	41.1	13	506.7	12.3	12-04-10
Despicable Me	3	32.6	26	357.5	11.0	12-14-10
How to Train Your Dragon	4	29.8	73	111.1	3.7	10-15-10
Iron Man 2	5	27.3	14	480.8	17.6	09-28-10
Inception	6	18.4	1	1007.7	54.7	12-07-10
Shrek Forever After	7	15.3	74	109.4	7.2	12-07-10
The Karate Kid	8	14.9	50	216.4	14.6	10-05-10
The Expendables	9	12.9	2	885.5	68.8	11-23-10
Tinker Bell and the Great Fairy Rescue	10	12.7	155	27.1	2.1	08-21-10

6.2.5 Distribution of Popularity of Transferred Content

In this section we estimate the distribution of popularity of top titles from different types of media transferred using BitTorrent and compare it to that of content sold in legal outlets, to better understand the preferences of users and how these preferences differ between what they can obtain for free and what they pay for. We estimate the popularity of the top 100 titles of Songs, Movies, Music Albums, TV Show seasons and TV Show episodes transferred in BitTorrent, where popularity is defined as the share of transferred copies of each title out of all transferred copies in all swarms sharing the respective type of media. The number of transferred copies of each title is obtained by summing the number of transferred copies in all the swarms that share that title.

We find that most BitTorrent transfers of media concentrate in a small number of very popular titles, especially in the case of movies and songs. Figure 30 presents the cumulative distribution of popularity of

the top 100 titles transferred using BitTorrent for different types of media, and shows that the 100 most popular titles in BitTorrent account for a large percentage of the copies transferred of each media type, especially in the case of individual songs and movies, for which the top titles account respectively for 57% and 59% of all transferred copies. These top 100 titles actually account for a small percentage out of all titles available in BitTorrent for each media type, which means that a small percentage titles accounts for a large share of transfers. The percentage of titles that the top 100 represents, out of all titles for a given media type, is upper bounded by the percentage of swarms that share those top 100 titles, out of existing swarms for each media type, which is presented in table 23 (this happens due to the fact that the number of swarms sharing each title is negatively correlated to the rank of the title - correlation coefficients between the number of swarms sharing each title and the rank of the title are also presented in table 23).

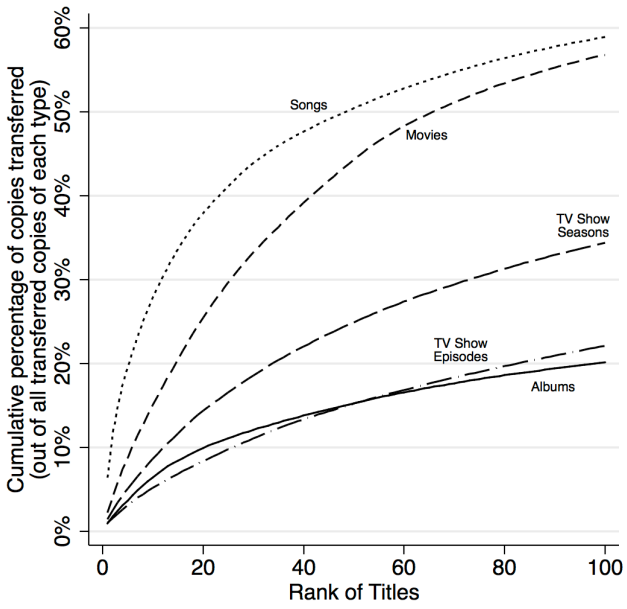


Figure 30. Cumulative distribution of the percentage of copies transferred of the top 100 titles of Songs, Music Albums, Movies, TV Show Episodes and TV Show Seasons found in BitTorrent, out of all transferred copies of each type.

The above figure shows that in BitTorrent, a lot of the copies transferred (and consequently copyright violations, and possibly impact on revenues of copyright holders) come from a very small number of titles. In fact transfers of the most popular titles account for a much larger share of all BitTorrent transfers than what is typically observed in sales of other online media outlets. Distribution of sales from online outlets

providing a large catalog of titles are typically long tailed (C Anderson 2006), with a small set of titles accounting for most sales (the head of the distribution, accounting for the top couple thousand titles, the number of titles typically available in a brick-and-mortar store), but also with a great number of relatively unpopular titles contributing to a significant share of overall sales (the long tail of the distribution). While we cannot examine the tail of the distribution of BitTorrent transfers, our estimates presented in figure 30 clearly show that the head of the distribution is much heavier than what is found for typical online outlets (C Anderson 2006). This may be in part due to the fact that less popular content tends to be short-lived in BitTorrent. A BitTorrent swarm only exists while there are enough connected peers to guarantee that transfers can take place, and this may cause swarms sharing unpopular titles with low frequency of transfers to disappear because they can't maintain sufficient connected peers.

Table 23. Percentage of transferred copies and percentage of swarms taken by the top 100 Song, Music Album, Movie, TV Show Episode and TV Show Season titles, out of all transferred copies and all swarms for each type of media. Average number of swarms sharing each title for the top 100 titles in each type of media.

	Percentage of copies	Percentage of swarms	Correlation between rank and number of swarms per title
Song	59%	3.4%	- 0.28
Music Album	20%	1.2%	- 0.43
Movie	57%	21%	- 0.60
TV Show Episode	22%	1.7%	- 0.28
TV Show Season	34%	9.1%	- 0.46

Furthermore, the very popular titles that account for the bulk of BitTorrent transfers are not only popular in BitTorrent, they are also among the most popular titles in terms of legal sales. Looking at users' preferences for specific titles we find that titles that rank high in terms of worldwide sales also rank high in terms of BitTorrent transfers. This is visible in the comparison between sales rank and BitTorrent transfers rank for the 10 top selling singles, albums and box-office movies worldwide and for the 10 top selling DVDs in the U.S. presented in tables 19 through 22 (section 6.2.4). The tables show that titles in the sales top 10 also rank high in terms of Bittorrent transfers, most of them being part of the BitTorrent top 50. Hence, BitTorrent serves as a source of popular content that is widely available for sale in legal outlets, not catering only to those seeking titles that can't be easily found for sale. Such BitTorrent transfers of widely available popular content are likely to displace more potential sales than those of

content that is hard to find or even unavailable in legal outlets, and thus they are expected to have a large impact on revenues of copyright holders.

A direct consequence of the distribution we find for BitTorrent transfers is that a large inventory is not that important a factor when trying to compete with P2P transfers. In fact, the diversity of offered titles seems to be a minor factor when compared to the importance of selecting which really popular titles to offer, especially in the case of music and movies for which a catalog of 100 titles can capture close to 60% of the “market”. These results differ from results we obtained from monitoring the ISU campus (presented in chapter 4), which, despite finding a significant share of transfers taken up by popular titles, also found heavy tails in the distribution of popularity of music and video titles. Various differences between our ISU analysis and the current analysis may hypothetically explain the differences. In particular, at ISU we monitored the activities of students in a university campus, whose particular tastes may be different from those of BitTorrent users in general, and our monitoring was limited to the titles and filenames that the appliances could extract from transfers that could be detected on the network, whereas we are now looking at data collected from the largest BitTorrent tracker and covering the majority of titles shared publicly using BitTorrent. Furthermore, over three years separate both data collection moments, and P2P usage may have changed, particularly in face of changes in influencing factors (such as available Internet connection speeds, for instance).

The shape of the popularity distribution also bears clear implications for enforcement. It means that preventing illegal transfers from 100 titles may cut the number of illegally transferred copies of copyrighted content by more than half in the case of movies and songs (and by a smaller, yet very significant percentage, in the case of other types of media). In addition, because the bulk of transfers of each title often comes from one or two of the swarms sharing that title (despite there being multiple swarms sharing each title) it is not only possible to cover a large percentage of shared content by acting upon a small number of titles, it is possible to do so by acting only upon the most popular swarms for each of those titles.

6.2.6 Technical Characteristics of Transferred Content

This section looks at technical characteristics of content transferred in BitTorrent, focusing on the file types under which each type of media is shared, on the digitalization methods used to capture the video content shared, and on the preferred video resolutions and audio bit rates. Understanding which technical characteristics of content users prefer can be useful for those seeking to provide legal alternatives to P2P. Furthermore, such characteristics have implications for enforcement to the extent that they can affect the performance of technological methods of detection of transfers of copyrighted content, in particular Deep Packet Inspection detection, whose detection success can be affected by the type of content being transferred.

By observing the file types shared in swarms for each type of media shared in BitTorrent we find that there is a preferred file type for each type of media, which in most cases accounts for more than three quarters of all transferred copies of content of that type of media. Table 24 shows this by presenting, for each media type, the main file type transferred, the percentage of swarms that contain that file type and the percentage of copies transferred from those swarms. For each type of media, the preferred file format in BitTorrent coincides with the file type that is generally most well known, widespread, and widely supported in terms of hardware and software readers/decoders (*mp3* for music, *avi* for video, Windows executable files for software and *pdf* for documents and books). The second most transferred type of file, for most media types, corresponds to archives. This has implications for copyright enforcement using deep packet inspection (DPI). On one side, it implies that, nowadays, content recognition technology needs only to be able to decode a small set of formats to be able to access the media transferred inside most files shared in BitTorrent. On the other side, it shows the already significant share of content transferred in BitTorrent that DPI cannot detect because it is transferred inside archives (it is practically impossible for DPI to perform content recognition if content transferred using P2P is stored inside

archives[102]). Using DPI for enforcement may lead to P2P users' behavior changes, in particular, enforcing copyrights for only a small number of file types may lead to users switching to more obscure file types that are not being enforced or to archived content, which in turn will further increase the amount of content that cannot be identified by DPI.

Table 24. Preferred file types supplied and transferred for each media type.

	Predominant file type	Percentage Swarms	Percentage Copies	Other file types (by decreasing percentage of transfers)
Song	mp3	74.3%	92.8%	rar, zip, ogg, flac, m4a, wma, ape, wav, 3gp, aac
Music Album	mp3	66.1%	91.9%	rar, flac, zip, vob, ogg, ape, iso, wma, m4a
Movie	avi	61.1%	69.1%	rar, mkv, wmv, mp4, vob, rmvb, iso, zip, mpg
TV Show Episode	avi	58.3%	82.3%	rar, mkv, mp4, rmvb, wmv, mpg, zip, m4v
TV Show Season	avi	51.3%	77.5%	mkv (18.5% swarms, 13.2% copies), vob, rar, iso, mp4, ts
Adult Content	avi	42.5%	39.3%	wmv (22.7% swarms, 23.2% copies), rar, zip, mpg, jpg
Software	exe	45.1%	88.2%	rar, zip, iso, ipa, cab, dmg, msi
Game	exe	8.4%	46.1%	rar (51.5% swarms, 30.5% copies), iso, zip, mdf, nds
Book	pdf	49.9%	60.2%	rar (20.6% swarms, 20.2% copies), zip, cbr, chm, txt, html
Document	pdf	91.2%	93.1%	cbr, rar, chm, zip, cbz, djvu, m4b, doc

We looked at method of digitalization of movies and TV shows shared using BitTorrent, and at resolution of movies, TV shows, songs and music albums. To do so, we examined tags indicating the method of digitalization[103], video resolution and audio bit-rate of content found in detected video and music swarms and broke down both the number of swarms and the number of transferred copies by the different categories for each of those variables. Such breakdowns show that high quality copies of movies, TV shows and music are supplied in BitTorrent, and that users transfer preferentially the high quality copies. It is only natural that users prefer the highest quality when there is no difference of price between different qualities of the same content. Considering that the cost of obtaining content from BitTorrent is a function of the number of bytes transferred and of the time spent transferring those bytes, and that many fixed broadband Internet connections use flat rate plans where number of transferred bytes does not influence price, then users' preference for higher quality also shows that they are not sensitive to the time spent in the transfer. For business, the direct consequence of the availability and preference for high quality

[102] DPI needs to gather a fraction of the content being transferred in order to perform content recognition. Archives need to be expanded in order to access the content contained therein, which is only possible if the archive is complete, or at least if specific parts of the archive are present. In P2P, given the fragmented nature of transfers, it is often difficult, if not impossible, to obtain all the parts of an archive via network monitoring. Furthermore, maintaining the archive parts while waiting for the possibility to expand them would require a large storage. All these become increasingly difficult as the speed of the monitored link increases.

[103] Please refer to Appendix B for a list of tags referring to methods of digitalization of video content shared in P2P and their respective meanings.

content in BitTorrent is that those providing legal alternatives can no longer use quality as a differentiating factor to attract customers away from the free but illegal BitTorrent transfers.

In the case of movies, we find that most swarms contain high quality DVD and Blue-ray Rips and that those are the types of most movie copies transferred. Figure 31.a shows a breakdown of the 56% of swarms sharing movies that contained information about the digitalization method. It shows that over 70% of movie swarms contain high quality formats, and those account for close to 70% of transferred copies. The percentage of swarms offering content digitalized prior to DVD release (Cam Rip, Telesync, and Telecine) is smaller, but it is about 15% of swarms and transferred copies. Concerning resolution, the breakdown of the 47% of movie swarms that had such information is portrayed in figure 31.b. DVD quality content accounts for close to 90% of movie swarms and transferred copies, with the remaining swarms containing higher definition content.

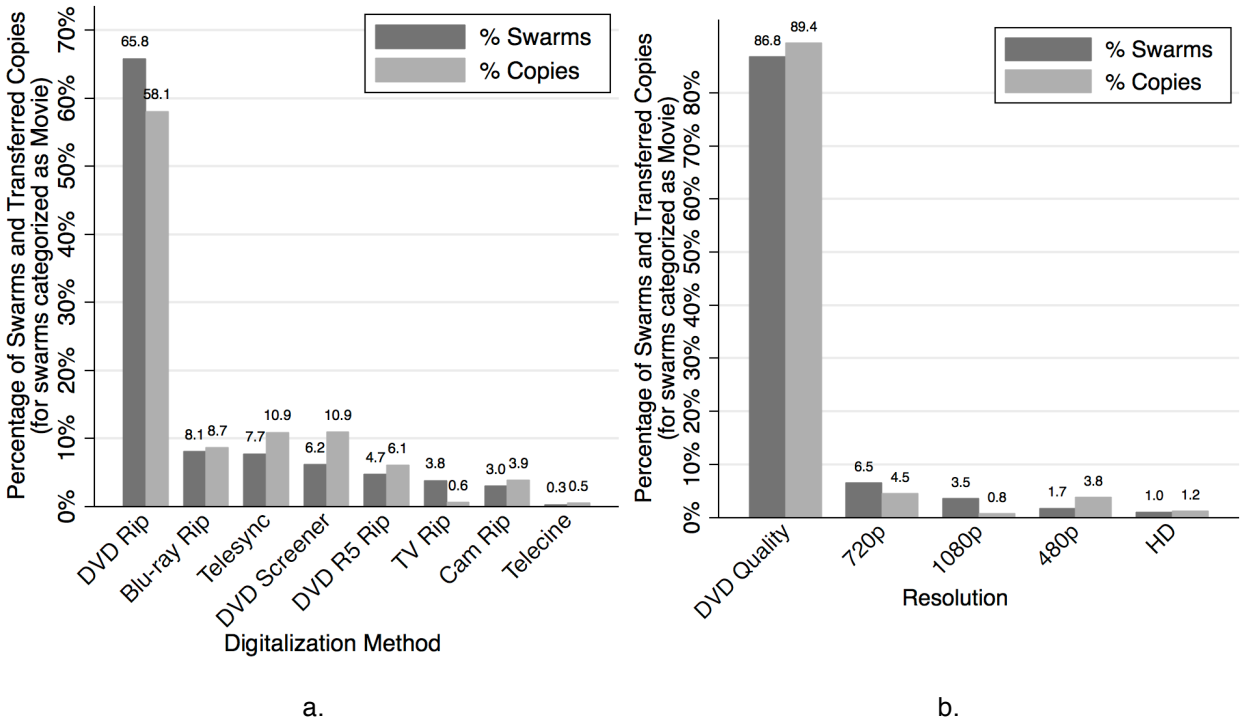


Figure 31. Breakdown of movie swarms and of transferred movie copies by methods of digitalization and resolution. a) Breakdown by method of digitalization. b) Breakdown by resolution.

High quality content is also prevalent in supply and consumption of TV content. However, consumption patterns are different for single episodes or whole TV show seasons. We found information on digitalization method in 21% of the TV show episode swarms and in 24% of the TV show season swarms. Resolution information could be found in 48% of TV show episode swarms and in 37% of TV show season swarms. Both supplied and transferred TV show episodes are high quality content. Most swarms and most transferred copies are TV Rips, obtained from digitally recoding the episode as it is airing, as shown in figure 32.a. Single episodes extracted from DVD rips are the second digitalization method with most swarms and most transferred copies. In terms of resolution, as shown in figure 32.b, most swarms contain copies in HDTV resolution, and that is the preferred resolution in transferred copies as well.

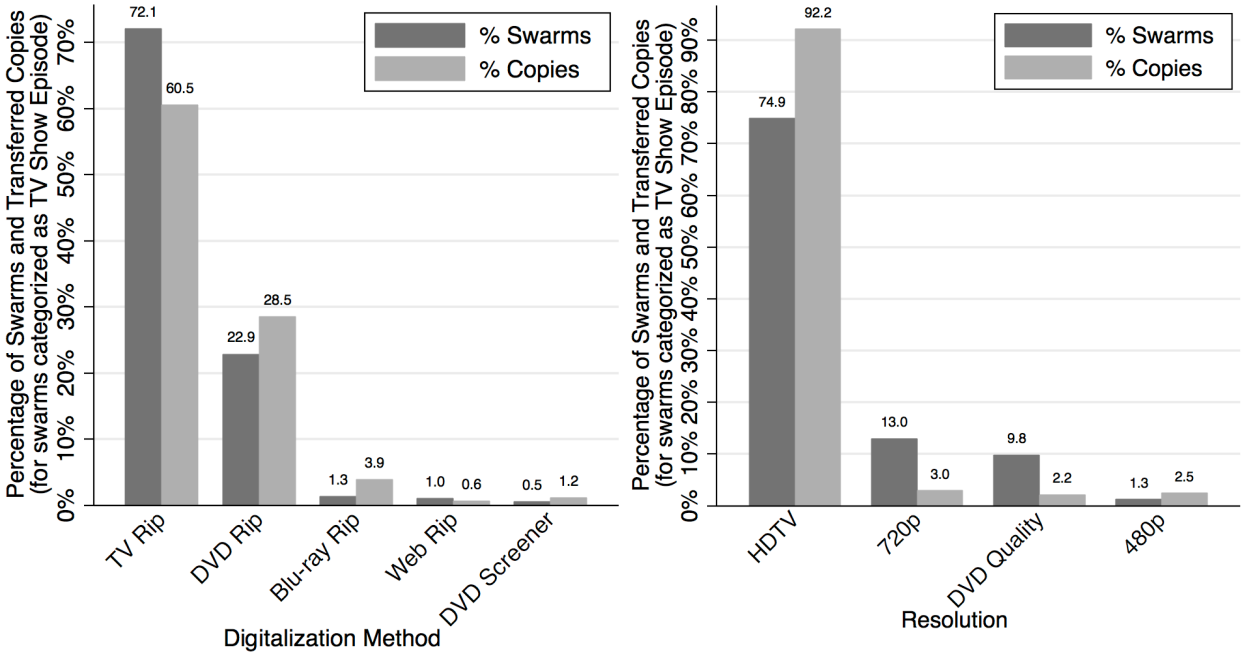


Figure 32. Breakdown of TV show episode swarms and of transferred TV show episode copies by methods of digitalization and resolution. a) Breakdown by method of digitalization. b) Breakdown by resolution.

When it comes to full seasons of TV shows, most content supplied is from DVD Rips, and those DVD Rips account for close to three-quarters of transferred copies. This is portrayed in figure 33.a, which also shows a higher percentage of Blu-ray Rips in the case of full seasons than in the case of single episodes, both in supply and consumption of content. As for resolution of transferred content, as shown in figure

33.b, the higher share of swarms and transfers are DVD quality, but the share of high resolution content, in particular 720p and 1080p content, accounts for more than a quarter of swarms and transferred copies.

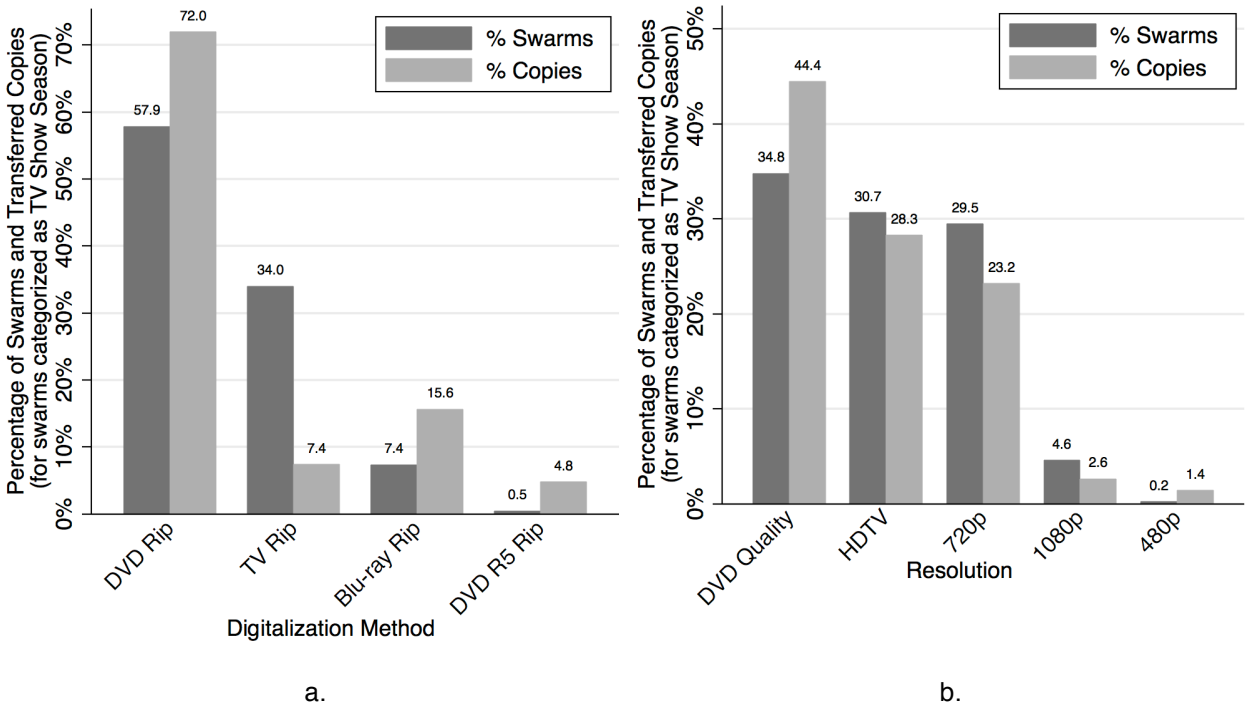


Figure 33. Breakdown of TV Show season bundle swarms and of transferred TV show season bundle copies by methods of digitalization and resolution. a) Breakdown by method of digitalization. b) Breakdown by resolution.

The differences between single episodes and complete seasons in terms of preferred types of capture and resolution shows that users care much more for high quality when transferring entire seasons than when transferring single episodes. One possible explanation for this fact is that single show downloads are for immediate consumption, and therefore the user wants to get the content as fast as possible and start enjoying it, whereas users transferring a full season of a TV show might wish to keep that content archived for repeated consumption in the future, and therefore be willing to allow the extra time to obtain higher quality copies.

Finally, concerning music, we find that higher-quality content is also preferred to lower quality content. However, since only 6% of song swarms and only 8% of music album swarms contained information on bit-rate, conclusions should be carefully drawn from these data. As shown in figures 34.a and 34.b, most single songs with bit rate information were supplied at a bit rate of 192kbps (higher quality than a regular

songs sold in iTunes, which is 128kbps) while most album bundles are supplied at a bit rate of 320 kbps. However, when it comes to consumption, the majority of copies transferred are of the high-end 320kbps media, both for single songs and for album bundles. One possible explanation for this fact concerns available download bandwidth. If the size of transferred songs was a concern in the past due to bandwidth limitations, it is no longer a concern nowadays for most users, who prefer to obtain the higher-quality versions of the content. However, the fact that we also observe a very low supply and number of transfers of lossless music (wav, flac, ape), indicates that users prefer most popular formats in high quality, perhaps due to the convenience allowed by the widespread support for those formats from music playing software and portable music players.

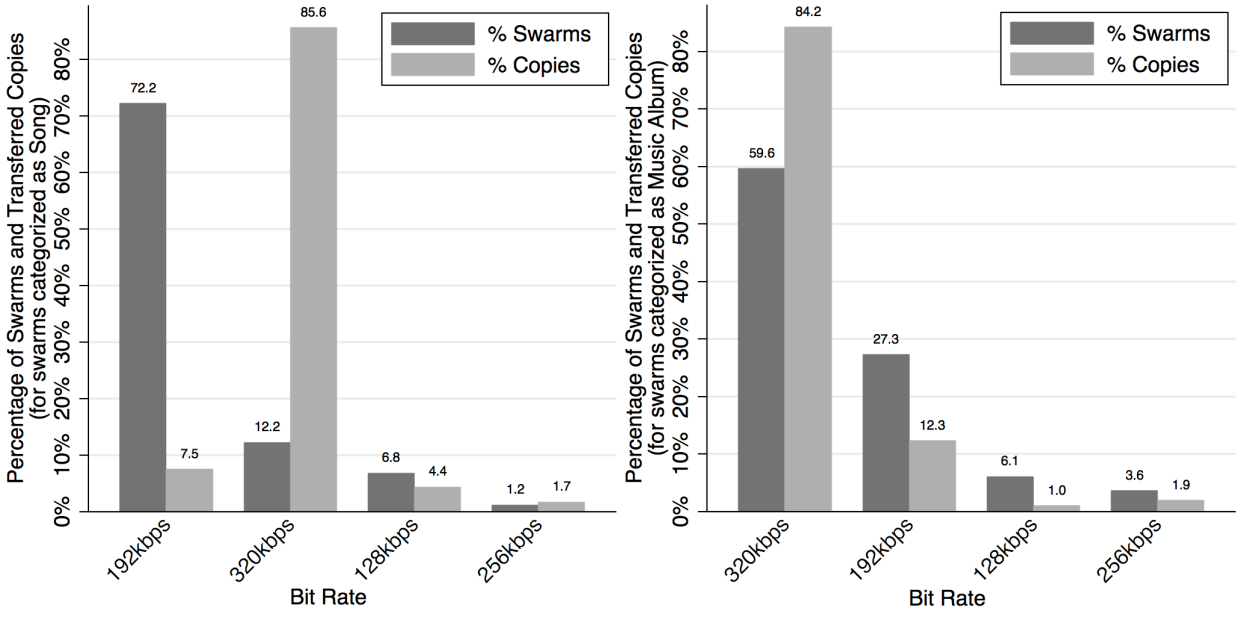


Figure 34. Breakdown of song and album bundle swarms and transferred songs and album bundles by bit rate. a) Songs. b) Album bundles.

6.3 Conclusions and Policy Implications

Using data collected from the most popular public BitTorrent tracker for 115 days between August 2010 and February 2011, we estimate lower bounds for the supply of content in BitTorrent and for the number of copies of content transferred using that P2P network.

We assess how much and what type of content is supplied in BitTorrent and investigate the extent to which BitTorrent is used for copyright infringement by providing a reasonable empirically derived lower bound for the number of copies of content transferred, a figure that has not been accurately quantified before. We find an average of 2.6 million BitTorrent swarms active at any given time. This means that 2.6 million bundles of content were available for download using BitTorrent at any moment. Taking into consideration the number of peers connected to each BitTorrent swarm and assuming that the breakdown of Internet connection technologies used by those peers is the same as the breakdown of fixed broadband connection technologies observed in OECD countries, we estimate that at least 800 million copies of content are transferred using BitTorrent on average per day.

We break down number of swarms and number of transfers by type of content shared to understand what content is more affected by illegal transfers using BitTorrent. We find that movies have the highest supply in BitTorrent (39% of swarms), followed by music albums (17.2%), TV show episodes (15.1%), and then by software (6.9%). These results confirm results from previous studies that indicated video as the most supplied type of content in BitTorrent. However, when considering transferred copies, which provide a more accurate approximation of the number of copyright violations performed using BitTorrent, software becomes the top type of media (37.9% of transfers), followed by individual songs (18.8%), movies (16.4%) and music albums (16.4%). Hence, despite being the principal type of content supplied, video (movies and TV shows) is not the type of content for which most copies are transferred. This finding exposes the main shortcoming in previous accounts of content shared in BitTorrent (Envisional 2011 , for instance), which estimated BitTorrent activity by looking only at the breakdown of swarms or number of

peers connected to each swarm, and didn't take into account how the number of bytes shared in each swarm influences the rate at which copies of content can be transferred.

The majority of content made available and transferred using BitTorrent is likely copyrighted content whose transfers result in copyright violations. We reach this conclusion based on the metadata found for the swarms we monitored, most of which indicating that the shared copies were obtained by digitalizing or re-encoding copyright protected content, and based on the fact that only a very small share of the content made available and transferred using BitTorrent could be found in index websites that specialize in content that can be legally transferred using BitTorrent. Swarms indexed by websites specialized in legal content correspond to under 0.2% of all detected swarms and the number of copies of content transferred represents about 0.02% of overall transferred copies of content. Hence, despite the effort from these index websites to promote legal transfers in BitTorrent, the number of copies of legal titles transferred is close to insignificant in practical terms when compared to the number of transferred copies of titles indexed by the better-known general-purpose indexes. Furthermore, while these websites attempt to promote legal video and audio, which account for the majority of indexed swarms, when it comes to legal content, BitTorrent users transfer mostly documents and software. This is unlike what we observe for all monitored swarms, where audio and video show both high supply and high number of transfers.

To put the number of transfers of movies and music performed using BitTorrent in perspective we compare it to legal sales of those types of content and find that the number of transfers is many times greater than the number of legal sales. The overall number of songs transferred using BitTorrent (individually or part of albums) was 13.1 times greater than the estimated worldwide sales of songs (digital or part of physical media), and the number of movies transferred using BitTorrent was 6.8 times greater than worldwide box-office sales and 16.4 times greater than U.S. DVD and Blu-ray disc sales. Furthermore, when looking at the most sold music and movie titles worldwide, BitTorrent transfers greatly exceed legal sales. To compare how individual titles performed in sales and in BitTorrent we estimated the number of copies transferred using BitTorrent for the 10 most sold music singles and albums worldwide, as well as for the top 10 grossing movies in worldwide box offices and top 10 most sold DVDs

in the U.S. market during our monitoring period. We find that BitTorrent transfers exceed sales by over an order of magnitude for most music titles, and by a smaller, yet very significant margin for the vast majority of movie titles. Furthermore, among the top 10 titles in terms of sales, we find that sales ranks are higher than BitTorrent transfers ranks, which means that BitTorrent transfers exceed sales by an even greater margin if the comparison is performed using the top titles in BitTorrent.

For music albums as well as for DVDs, we find large variation in the ratio of BitTorrent transfers to sales between titles, and hypothesize that this is possibly due to the nature of the media transferred and the demographics it typically appeals to. Titles appealing to the teenager and young adult demographics have disproportionately higher ratios of BitTorrent transfers to sales than titles that appeal to an older segment of the population. Using this information, it may be possible to predict which titles in copyright holders' catalogs are more likely targets of illegal sharing and estimate how that sharing will affect sales of those titles. Teenagers and young adults are an important segment of the population whose members are typically avid consumers of media, but who, at the same time, may have less willingness to pay or disposable income to purchase such media. This segment of the population is typically tech-savvy and has in P2P a free, yet illegal, alternative, which they seem to be taking advantage of. Copyright holders can use this information to try to drive those consumers away from P2P, either by deploying selective enforcement focusing on the titles that typically appeal to those demographics, or by further investigating which factors drive such consumers away from purchasing content in order to devise more compelling legal alternatives.

One of the factors that those providing legal alternatives can no longer use as a differentiating factor to attract customers away from the free but illegal BitTorrent is the quality of content. We find that high quality copies of movies, TV shows and music are supplied in BitTorrent, and that users transfer preferentially the high quality copies. In the absence of a price differential between different qualities of the same content, it is natural that users prefer the highest quality. Users' preference for higher quality also shows that they are not sensitive to the time spent in the transfer, given that the cost of obtaining content from BitTorrent is a function of the number of bytes transferred and of the time spent transferring

those bytes, and that in today's flat rate Internet connection plans the number of transferred bytes does not influence price.

Hence, we find that BitTorrent users have access to the popular titles they seek, in high quality and in formats that they can use in multiple platforms, and that they transfer large numbers of those titles for free from BitTorrent. We cannot tell the extent to which those consumers would purchase the same media if they didn't have the option of getting it from BitTorrent, but the fact that the number of BitTorrent transfers is one order of magnitude greater than legal sales means that getting even a small percentage of BitTorrent users back to purchasing could represent a significant increase in the legal market. In order to start monetizing this potential market, it is necessary to understand why BitTorrent users choose to transfer media illegally instead of purchasing it. One of the reasons may be that BitTorrent users' willingness to pay for content is below the price of that content. If that is the case, then perhaps by adjusting pricing and bundling models, or by segmenting the market, copyright holders could capture some of the current BitTorrent users with highest willingness to pay and, despite possibly realizing less revenue per sale, increase sales in a way that would lead to an overall increase in revenue.

Concerning the popularity of media titles transferred using BitTorrent, we find that transfers concentrate in a small number of very popular titles, especially in the case of movies and songs for which the top 100 most transferred titles account respectively for 57% and 59% of all transferred copies. This means that a lot of the copies transferred (and consequently copyright violations, and possibly impact on revenues of copyright holders) come from a very small number of titles, more so than what is typically observed in legal sales of other online media outlets. Furthermore, we find that the most popular titles in terms of legal sales are also popular in BitTorrent and are among the top titles that account for the bulk of BitTorrent transfers. This shows that BitTorrent is not catering only to those seeking titles that can't be easily found for sale, it serves as a source of popular content that is widely available for sale in legal outlets and whose transfers are likely to displace more potential sales than those of content that is hard to find or even unavailable in legal outlets. Besides that, the high concentration of transfers in few very popular titles shows that, unlike what we had observed when monitoring media transfers in the ISU university

campus in 2007-2008, a large inventory may not be that important a factor anymore when trying to compete with P2P transfers. In fact, the diversity of offered titles seems to be a minor factor when compared to the importance of selecting which really popular titles to offer, especially in the case of music and movies for which a catalog of 100 titles can capture close to 60% of the “market”. When it comes to enforcement, the main implication is that, if the most popular titles can be identified, preventing illegal transfers from as few as 100 titles can potentially cut the number of illegally transferred copies of copyrighted content by more than half (in the case of movies and songs).

Looking at technical characteristics of content transferred using BitTorrent, we find a preferred file type for each type of media (*mp3* for music, *avi* for video, *exe* for software and *pdf* for documents and books), which in most cases accounts for more than three quarters of all transferred copies of content of that type of media, and that coincides with the file type that is generally most well known, widespread, and widely supported in terms of hardware and software readers/decoders. For most media types, the second most transferred type of file corresponds to archives (*rar*, *zip*, etc.). This has implications for copyright enforcement using deep packet inspection (DPI). On one side, content recognition technology needs only to focus on a small set of formats to be able to access the media in most files shared in BitTorrent. But on the other side, there is already significant share of content transferred in BitTorrent that DPI cannot identify as copyrighted because it is transferred inside archives. Moving forward, if DPI is used for enforcement, and in particular if enforcement focuses on a small number of file types, P2P users may start switching to more obscure file types that are not being enforced or to archived content altogether, which in turn will increase the amount of content that cannot be identified by DPI.

7 Conclusions and Policy Implications

This research analyzed data collected from a university campus via network monitoring and from the largest public BitTorrent tracker to fulfill two main goals. The first goal is to characterize the extent to which P2P is used to perform unauthorized transfers of copyrighted content, which provides a perspective of how much copyright infringement is happening using P2P and how that is evolving over time. Our findings contribute to understanding the extent and the evolution of P2P usage on a college campus, as well as to characterize what content is transferred using P2P, both on campus and overall on the Internet. The second goal is to evaluate how well existing network monitoring technologies, in particular Deep Packet Inspection (DPI), can detect P2P transfers and whether they carry copyrighted content, to inform those seeking to employ DPI about the effectiveness of the technology and its limitations. Towards this goal, we assess how well DPI performs in detection of P2P activity, in particular how well it detects users of P2P and whether or not they transfer copyrighted content. We also focus on the extent to which counter-measures such as traffic encryption impede DPI from detecting P2P activity by comparing the performance of DPI to the performance of other monitoring technologies that rely only on summaries of exchanged traffic.

We find that use of P2P and transfers of copyrighted content using P2P were widespread in a university campus in Spring 2008, and that in late 2010 transfers of copyrighted content using BitTorrent overall on the Internet amounted to many times the number of legal sales of content. In Spring 2008 we could detect 40% of the students living on campus using a P2P protocol, 70% of which were observed attempting to transfer copyrighted material, and most such attempts were likely to be copyright violations. On average, users detected transferring copyrighted material were observed attempting to transfer four distinct copyrighted media titles during a day of monitoring. In late 2010, using data collected for about 6 months from the main public BitTorrent trackers, we find an average of 2.6 million BitTorrent swarms active at any given time, which we estimate to originate transfers of at least 800 million copies of content on average per day, most of which were likely copyrighted content transferred without the authorization of copyright

holders. Both figures, which correspond to lower bounds, show that P2P is heavily used to transfer copyrighted content.

P2P can be used to transfer content without violating copyright, but we found no evidence that this was common both on campus and overall on the Internet using BitTorrent. On campus, we found no evidence to support the hypothesis that a significant fraction of students use P2P to obtain content that would not be flagged as copyrighted and not to obtain copyrighted music and video as well. Overall on BitTorrent, we found that swarms indexed by websites specialized in legal content correspond to under 0.2% of all detected swarms and the number of copies of that content transferred represents about 0.02% of all transferred copies. Thus, despite the effort from these websites specialized in legal content to promote legal transfers in BitTorrent, the number of copies of legal titles transferred is close to insignificant in practical terms when compared to the number of transferred copies of titles indexed by the better-known general-purpose indexes, most of which are likely copies of copyrighted media.

Hence, evidence collected both from monitoring a university campus and from BitTorrent trackers corroborates the fact that copyright law is violated frequently using P2P. This calls for considering significant changes, which could be changes in policy, business practices, enforcement methods, technology, consumer education, or a combination of these. Our results alone cannot tell us exactly what approach should be followed, but they can help answer some important questions and inform policymaking.

One such question is whether particular interventions, such as a campaign to educate Internet users about copyright law, should be targeted at a specific group, or applied generally to all. Among college students, we found no reason to target any specific group, because P2P users and users transferring copyrighted content were detected across all demographics, with fairly similar incidence among different genders, ages, classes and majors. However, if education campaigns prove to be useful, there is reason to reach students before they get to college. We find high incidence of P2P activity for freshmen in their first month on campus, followed by a gradual decrease in subsequent semesters, which indicates that

students became P2P users prior to entering college and is consistent with claims by higher education officials that students “learn” to use P2P at a younger age and establish their P2P habits prior to entering college (U.S. Congress 2007b).

Another question concerns how the industries that produce and distribute copyrighted content are affected by P2P and how they can respond to the challenge of competing with a free but illegal alternative. Focusing on music and movies, we find that the overall number of songs transferred using BitTorrent (individually or part of albums) was 13.1 times greater than the estimated worldwide sales of songs (digital or part of physical media), and the number of movies transferred using BitTorrent was 6.8 times greater than worldwide box-office sales and 16.4 times greater than U.S. DVD and Blu-ray disc sales. Although it is impossible to quantify from our data how the number of copies of content transferred both on campus or overall in BitTorrent translates to lost sales, as this depends on many factors, it is reasonable to assume some sales are lost to P2P. In particular, looking at the popularity of media titles transferred using BitTorrent, we find that most transferred copies are from a small number of very popular titles that are also available in legal media outlets. BitTorrent transfers of such titles are likely to displace more potential sales than transfers of content that is hard to find or even unavailable in legal outlets.

While we cannot tell the extent to which consumers would purchase the same media if they didn’t have the option of getting it from BitTorrent, our findings from campus show that consumers who transferred content from P2P still purchased content from the iTunes Store, the main online seller of music at the time. We found that 22% of P2P users detected on campus still made purchases from the iTunes Store, that they were more interested in content from the iTunes Store than non-P2P users, and that the average amount of purchases done by each user who purchased from the iTunes Store was about the same for P2P and non-P2P users. Hence, while illegal transfers using P2P are widespread, there are still P2P users who decide to pay for content at times, and to get content for free from P2P (risking legal action by copyright holders) at other times. This means that users’ decision to purchase content or get it from P2P depends on other factors besides price or risk of legal action. Perhaps by understanding which factors

influence that decision copyright holders can take actions that would drive at least a fraction of P2P users back into the legal market.

By looking at the content transferred in BitTorrent and how it compares to legal sales, we can start to understand some of those factors. We compare sales figures to number of copies transferred in BitTorrent for the worldwide top 10 music singles, music albums and box office movies and for the U.S. top 10 DVDs. BitTorrent transfers exceed sales by over an order of magnitude for most music titles, and by a smaller, yet very significant margin for the vast majority of movie titles. However, music albums and DVDs showed a large variation in the ratio of BitTorrent transfers to sales between titles, which is possibly due to the nature of the media transferred and the demographics it typically appeals to. Titles appealing to the teenager and young adult demographics have disproportionately higher ratios of BitTorrent transfers to sales than titles that appeal to an older segment of the population. This information can be useful to predict which titles in copyright holders' catalogs are more likely targets of illegal sharing and estimate how that will affect sales. Teenagers and young adults are typically avid consumers of media, but at the same time, may have less willingness to pay or disposable income to purchase such media. They are typically tech-savvy and may see in P2P a free, yet illegal, alternative to obtain the media they are not willing to pay for. This information can be useful to try to drive those consumers away from P2P, either by deploying selective enforcement focusing on the titles that typically appeal to those demographics, or by devising more compelling legal alternatives.

Quality of the content may have been one differentiating factor for legal alternatives in the past, but offering high quality content may no longer be useful to attract customers away from illegal BitTorrent. We find that high quality copies of movies, TV shows and music are supplied in BitTorrent, and that users transfer preferentially the high quality copies, which is natural in the absence of a price differential between different qualities of the same content. This also shows that users are not sensitive to the time spent in transfers.

The selection of content available in legal outlets may be another relevant factor. We looked at the most popular titles transferred using P2P on campus, as well as worldwide using BitTorrent, and found that small shares of high demand titles accounted for most transfers in both cases. In global BitTorrent, for movies and songs, close to 60% of the copies transferred (and consequently copyright violations) come from the top 100 titles. This is a greater concentration around popular titles than what is typically observed in legal sales of other online media outlets, and shows that one important factor to compete with BitTorrent is the selection of popular titles that are offered, especially in the case of music and movies, for which a catalog of 100 titles can capture close to 60% of the “market”. However, worldwide BitTorrent is more concentrated around really popular titles than campus P2P, where it takes a couple thousand top titles to get 50% of transfers. College students that use P2P seem to behave differently from worldwide BitTorrent users in what concerns their media tastes and which media they obtain from P2P, which indicates that there may be different factors that legal offers of content need to take into account depending on which type of market they try to reach. While the global market seems to value blockbusters, college students seem to value a larger and more diverse catalog. This is corroborated by our findings from looking at transfers of content on campus, where we find a heavy tail of low demand titles adding up to a significant share of transfers. Hence, to appeal to college students, an important demographic for media sales and in particular for music, legal services should offer a large selection of titles besides offering the top hits of the moment. While providing such a large selection of content may be possible for online services, it is typically not possible for bricks and mortar providers of paid content (CDs, DVDs), who are left at an inherent competitive disadvantage.

Besides the above, it is probably worth exploring other factors that might influence a user’s decision to choose P2P or a legal paid service for copyrighted content. Such factors may include the convenience or ease of use of legal services, ease of search, transfer speed, or usage restrictions in obtained content. In this way, those legally distributing content online may perhaps make their offerings more attractive in order to compete with the free but illegal alternative. This may reduce the impact of P2P use on the revenues of copyright holders, but it probably would not eliminate that impact, considering that the

number of transferred copies using BitTorrent alone is over an order of magnitude greater than legal sales and that more than three quarters of P2P users detected on campus were not seen purchasing from the iTS, at the time the top music retailer in the U.S. (Neumayr and Roth 2008).

Another possible course of action, besides attempting to offer compelling legal alternatives to P2P, is to enforce copyright by detecting users attempting to illegally transfer copyrighted content. Both alternatives have been put in place to some extent in the last decade, with copyright holders changing their content offers as a way to fight decreasing sales figures, and at the same time engaging in various attempts to enforce their copyrights and prevent P2P users from performing illegal transfers. One way to determine whether those measures were successful is to assess how P2P activity evolved over time. We looked at the evolution of P2P activity on campus over time using Deep Packet Inspection (DPI) and found a decrease in detected P2P between Spring 2007 and Spring 2008, both in terms of users detected using P2P (10% decrease) and of users transferring copyrighted material (20% decrease, out of detected P2P users), as well as in the average number of unique copyrighted titles detected being transferred per user in a day of monitoring (decreasing from seven to four titles per user detected transferring copyrighted content). Such decrease can be due to multiple factors. It can be a result of students abandoning P2P for legal services, or for other methods of obtaining content online that can also violate copyright law, such as video streaming websites, direct download link providers (e.g., RapidShare), or Usenet. It can also be a result of students shifting away from observable P2P by adopting counter-measures against detection. Using different technologies to detect P2P activity, we find that, in Fall 2007, students that used BitTorrent were already adopting encryption as a countermeasure against detection, and that there was a shift from unencrypted to encrypted BitTorrent from Fall 2007 to Spring 2008. While we cannot conclude that such shift was sufficient to explain the observed decrease in P2P activity detected via DPI, it certainly contributed towards a smaller number of users that DPI could detect.

Deep Packet Inspection (DPI) has the ability to detect P2P traffic and to identify whether that traffic carries content that is part of copyrighted media titles. As with any technology, those considering deploying DPI must determine whether intended benefits outweigh costs and any unintended side effects, and whether

or not user behavior would ultimately reduce a system's ability to detect transfers of copyrighted material is certainly one of the factors to consider. We investigated how DPI performs as a tool for detection of P2P activity and transfers of copyrighted content using P2P, focusing only on its detection effectiveness (not exploring other relevant issues in use of DPI, such as cost and privacy). We find that after a couple of weeks of monitoring, DPI technology was able to detect most of the users that attempted to transfer copyrighted content out of network users that could be detected engaging in P2P on campus (between 70% and 80% of detected P2P users could be detected transferring copyrighted content). This means that, at least in the short term, DPI could be an effective tool to assess which network users transfer copyrighted content using P2P after some weeks of monitoring, which may be sufficient for purposes such as putting in place user education campaigns and/or for enforcement. However, we found some limitations in the technology that reduce its effectiveness for purposes other than detecting users that transfer copyrighted content, and others that prevent it from detecting users if they adopt counter-measures.

DPI could not detect all communication sessions that carried copyrighted content, and while this didn't affect much detection of users because each user engaged in multiple communication sessions over time and was detected sooner or later, it impacts detection of specific media titles transferred over P2P, or which users transferred each media title, because these are more sensitive to failures in detection of individual communication sessions. As a result, DPI will not be as effective for purposes such as estimating the impact of illegal transfers on revenues of copyright holders. Furthermore, DPI appliances used for monitoring, which were leading products in the market, were significantly more successful in detecting copyrighted audio than in detecting copyrighted video. Hence, if DPI is used to collect data on which media titles are transferred using P2P, its accounts will be significantly biased against video content, one of the types of content most transferred in P2P nowadays. As a result, DPI should not be used for purposes such as implementing a revenue sharing model that would split collected money among copyright holders if some copyright holders specialize in music and others in video (unless it is

possible to make a correction to account for this difference). Besides that, if DPI is used for copyright enforcement, it is likely to benefit the music industry more than the movie/TV industries.

Perhaps the main limitation of DPI is that it cannot detect P2P users if they encrypt their P2P traffic, and neither can it detect whether or not those users transfer copyrighted content. Since the two dominant P2P networks (BitTorrent and Gnutella) support encryption, this limitation can significantly diminish the usefulness of DPI for copyright enforcement should users start to employ encryption at a large scale as a countermeasure against detection. We found BitTorrent users already activating encryption on campus in Fall 2007, and shifting from unencrypted to encrypted BitTorrent during the 2007-2008 academic year, despite the fact that no active enforcement was taking place via DPI. While we cannot be sure of the reasons that motivated the shift to encrypted BitTorrent on campus, the fact that some users started encrypting means that, in the long term, use of DPI for enforcement can drive even more P2P activity underground, a switch that could significantly hinder the purpose of using DPI in the first place.

The fact that encryption caused DPI to miss a significant share of BitTorrent activity in the 2007-2008 academic year, and that it has the potential to cause even more activity to go undetected as time goes by, implies that DPI alone may not be the most effective solution for online copyright enforcement. Methods not vulnerable to encryption, such as behavioral detection methods that rely on traffic summaries to detect hosts engaging in P2P, while not suitable for the type of enforcement that can be implemented via DPI, can be useful tools for copyright protection. The behavioral classifiers that we implemented were able to detect up to 85% of the users of unencrypted BitTorrent that DPI detected on average per hour, and could also detect users of encrypted BitTorrent. These were first attempts at implementing this type of classifiers and thus leave great room for improvement, but the obtained results prove the usefulness of this technology as a complement to DPI's detection, as well as on its own, for specific applications such as issuing warnings about copyright issues to users engaging in BitTorrent without imposing any penalties, or estimating the extent of P2P usage.

8 Future Work

We propose three main directions for future work that build on the results presented and methodologies developed in current research and further improve the knowledge about the underlying issues. The first direction is the logical follow-up to our first study and consists on mapping and evaluating strategies regarding the online exchange of copyrighted material in place in universities across the USA. The second direction is the continuation of the development of behavioral classifiers that we initiated in our second study. And the third direction is a follow-up to our third study to further understand the dynamics of file sharing using BitTorrent.

Our first study demonstrated that P2P and transfers of copyrighted content are widespread on campus. This brings about adverse consequences for universities, in particular concerning the burden of P2P traffic on campus communications, the administrative costs of processing DMCA requests from copyright holders, and more recently, the costs of complying with the requirements of the Higher Education Opportunity Act (110th Congress 2008), which started being enforced on July 1, 2010. Design of effective methods to address illegal file sharing on university campuses would benefit both universities and copyright holders. In fact, some of those methods may already exist and be in use on some campuses, since universities across the USA and around the world have adopted many different practices in response to P2P use by students (including methods for identification of students from IP addresses, policies for disconnection of users from the network, education about copyright, promotion of access to legal media services, or deployment of technology to deter illegal transfers of copyrighted content).

By collecting and analyzing data on practices put in place in universities across the country, as well as data on copyright infringement figures for those universities (for instance, from copyright holders, which have been monitoring P2P networks for years and collecting the IP addresses of nodes sharing copyrighted content) it would be possible to understand the factors influencing the extent to which copyright violations occur on campus networks. This would be accomplished by exploring how the pattern

of alleged copyright violations varies from university to university, how university practices and policies vary, and whether these are correlated. Furthermore, by putting together information on the extent to which practices adopted by universities manage to decrease copyright violations and on the costs of implementing them, as well as other potential advantages and drawbacks of each practice, would allow for a ranking of the strategies by cost-effectiveness, which would permit understanding which practices are most effective in decreasing copyright infringement in universities.

A second direction for future work concerns the further development of behavioral detection of P2P traffic. In our second study, we developed a classifier of BitTorrent traffic that uses behavioral information extracted from traffic summaries to infer whether hosts are engaging in BitTorrent. Using that classifier, we were able to detect over 85% of the hosts engaging in unencrypted BitTorrent on average per hour, and a significant amount of hosts engaging in encrypted BitTorrent. However, this was our first approach to developing one such classifier, and we identified many aspects that can be improved. Future work could build on the behavioral classifier we developed and improve the following four aspects. First, it could introduce further tests for detection of BitTorrent hosts, for instance, tests involving the cardinality of the set of ports that a host communicates to, which is likely higher for P2P than for other protocols. Second, the training process (parameter tuning) could be improved in two fundamental aspects: the training algorithm should be revised to try to avoid local minima and search for globally optimal solutions, and new training and test sets should be collected, preferably with real information (not collected by DPI) about whether or not each communication session was originated by BitTorrent (encrypted or not). Third, the organization of tests and relative weight of each of them in the final classifier could be further developed, namely to give more importance, and thus take better advantage, of tests with greater accuracy. Finally, P2P protocols other than BitTorrent could be incorporated in detection, in particular Gnutella, which is the other network supporting encryption. The objective of such improvements would be to enhance the performance and scope of the behavioral classifier, so that it could be used in real applications for detection of P2P hosts, perhaps alongside DPI, as suggested in our conclusions section.

The final direction for future work is essentially a continuation of our third study. We believe that there is a lot more to understand about the dynamics of sharing of copyrighted content using BitTorrent from analyzing data collected from trackers and using the process of estimation of number of copies transferred that we developed. One important further development is relative to the method of estimating transferred number of copies, in particular to the estimation of transfer speed achieved by leechers. We believe there is much to gain in terms of accuracy from engaging in further collection of data on transfer speeds using a wider range of Internet connection technologies and, if possible, using actual swarms sharing copyrighted content, as swarms sharing illegal content typically reach higher counts or seeders and leechers. This would make the estimates more robust and would yield closer approximations of the number of copies transferred.

Other further developments concern analysis of dynamics of BitTorrent over time. The methodology for estimating the number of copies of content transferred per unit of time can be applied periodically, to observe how use of BitTorrent is changing over time. Continuing to collect data at regular intervals and analyzing those data over a longer period of time can lead to several interesting findings we were not able to achieve. For instance, it can show whether BitTorrent is growing or diminishing as a means of transferring copyrighted content, or whether it is diminishing as a means of transferring some categories of copyrighted content, and expanding for other categories. A data set with longer timespan would also allow for analysis focusing on the effect of new illegal versions of a title in the popularity of existing versions, and of the effect of relevant external events related to certain titles on the popularity of illegal versions of that content. This could be accomplished by analyzing the evolution over time of the number of transferred copies for a selection of titles for which there were events of interest. For instance, using movie titles for which a CAM version existed and then a DVD Rip version was introduced, and comparing copy transfer rates before and after the introduction of the DVD Rip version. Or using music albums that were available in BitTorrent prior to the actual release, and comparing copy transfer rates before and after the album release (and associated marketing campaign).

9 References

105th Congress. 1998. *Digital Millennium Copyright Act (DMCA). Pub. L. 105-304*. October 28.

110th Congress. 2008. *Higher Education Opportunity Act (HEOA). Pub. L. 110-315*. August 14.

ACTA Negotiators. 2010. *Anti-Counterfeiting Trade Agreement*. December 3. http://trade.ec.europa.eu/doclib/docs/2010/december/tradoc_147079.pdf.

Adamic, Lada A. 2000. *Zipf, Power-laws, and Pareto - a ranking tutorial*. Online tutorial. HP Labs, Information Dynamics Lab. http://www.hpl.hp.com/research/idl/papers/ranking/.

Adamic, Lada A., and Bernardo A. Huberman. 2002. Zipf's law and the Internet. *Glottometrics* 3: 143-150.

Andersen, Birgitte, and Marion Frenz. 2008. *The Impact of Music Downloads and P2P File-Sharing on the Purchase of Music: A Study for Industry Canada*. University of London Working Paper.

Anderson, C. 2006. *The Long Tail: Why the Future of Business Is Selling Less of More*. New York: Hyperion.

Anderson, N. 2009. 3.9M or 7M? Behind the UK's dodgy file-sharing numbers. *Ars Technica*. October. http://is.gd/drSWJ.

———. 2010a. The RIAA? Amateurs. Here's how you sue 14,000+ P2P users. *Arstechnica.com*. June 1. http://arstechnica.com/tech-policy/news/2010/06/the-riaa-amateurs-heres-how-you-sue-p2p-users.ars.

———. 2010b. P2P settlement factory expects £10 million from... mailing letters. *Arstechnica.com*. September 29. http://arstechnica.com/tech-policy/news/2010/09/p2p-settlement-factory-expects-10-million-from-mailing-letters.ars.

Bangeman, Eric. 2007. New bill would punish colleges, students who don't become copyright cops. *Ars Technica*. November 11. http://tinyurl.com/ref-bman-2007.

———. 2008a. Judge kills RIAA subpoena: making available not infringement. *Ars Technica*. April 3. http://tinyurl.com/ref-bman-2008b.

———. 2008b. Study: BitTorrent sees big growth, LimeWire still #1 P2P app. *Ars Technica*. April 21. http://arstechnica.com/old/content/2008/04/study-bittorren-sees-big-growth-limewire-still-1-p2p-app.ars.

Bartlett, Genevieve, John Heidemann, and Christos Papadopoulos. 2007. Inherent Behaviors for On-line Detection of Peer-to-Peer File Sharing. In *Proceedings of the 10th IEEE Global Internet*. Anchorage, Alaska, USA: IEEE, May. http://www.isi.edu/~johnh/PAPERS/Bartlett07a.html.

Bhattacharjee, Sudip, Ram D. Gopal, Kaveepan Lertwachara, James R. Marsden, and Rahul Telang. 2007. The Effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts. *Management Science* 53, no. 9: 1359-1374.

Buskirk, Eliot Van. 2007. A Poison Pen From the RIAA. *Wired*. February 28. http://tinyurl.com/ref-buskirk-2007.

Cheng, Jacqui. 2009. The Pirate Bay to roll out secure €5 per month VPN service. March. http://tinyurl.com/ref-cheng-2009.

Christin, Nicolas, Andreas S. Weigend, and John Chuang. 2005. Content availability, pollution and poisoning in file sharing peer-to-peer networks. Proceedings of the 6th ACM conference on Electronic commerce. doi:http://doi.acm.org/10.1145/1064009.1064017.

Cisco. 2008. *Cisco Visual Networking Index: Forecast and Methodology, 2007-2012*. Cisco.

———. 2009. *Cisco Visual Networking Index: Usage Study*. Cisco.

———. 2010. *Cisco Visual Networking Index: Forecast and Methodology, 2009-2014*. Cisco.

Claise, B. 2004. *Cisco Systems NetFlow Services Export Version 9*. The Internet Engineering Task Force, Network Working Group.

Cohen, Bram. 2008. The BitTorrent Protocol Specification. *The BitTorrent Protocol Specification*. February 28. http://www.bittorrent.org/beps/bep_0003.html.

Collins, Michael P., and Michael K. Reiter. 2006. Finding Peer-to-Peer File-Sharing Using Coarse Network Behaviors. In *Computer Security – ESORICS 2006*, 1-17. http://dx.doi.org/10.1007/11863908_1.

Constantinou, Fivos, and Panayiotis Mavrommatis. 2006. Identifying Known and Unknown Peer-to-Peer Traffic. *Fifth IEEE International Symposium on Network Computing and Applications (NCA 2006)*: 93-102.

EFF. 2008. *RIAA v. The People: Five years later*. Electronic Frontier Foundation.

Envisional. 2011. *Technical report: An Estimate of Infringing Use of the Internet*. Envisional ltd. http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf.

Fischer, Ken. 2007. Bill would force “top 25 piracy schools” to adopt anti-P2P technology. *Ars Technica*. July 23. http://tinyurl.com/ref-fischer2007.

Froemling, Todd. 2006. ISU develops new file sharing options. *Daily Vidette at Illinois State University*, May 2. http://is.gd/eFciC.

Gannes, Liz. 2009. A Sector Assembles to Turn Video Pirates Into Gold. April. http://gigaom.com/video/a-sector-assembles-to-turn-video-pirates-into-gold/.

Gopal, Ram D., and Sudip Bhattacharjee. 2006. Do Artists Benefit from Online Music Sharing? *Journal of Business* 79, no. 3: 1503-1533.

Guess, Andy. 2008. Downloading by Students Overstated. *Inside Higher Ed*. http://tinyurl.com/ref-guess-2008.

Hong, Seung-Hyun. 2007. Measuring the Effect of Digital Technology on the Sales of Copyrighted Goods: Evidence from Napster. *Available at SSRN*.

IETF. 1983. *RFC 850 - Standard for Interchange of USENET Messages*. RFC. Internet Engineering Task Force, June. http://www.ietf.org/rfc/rfc0850.txt.

Illinois State University. 2006. *ISU FactBook 2006-2007*.

———. 2008. Digital Citizen Project at Illinois State University, Summary of Project. *Digital Citizen Project at Illinois State*. February. http://www.digitalcitizen.ilstu.edu/summary/.

Jupiter Media Metrix. 2001. Global Napster Usage Plummets, But New File-Sharing Alternatives Gaining Ground. http://www.thefreelibrary.com/GLOBAL+NAPSTER+USAGE+PLUMMETS,+BUT+NEW+FILE-SHARING+ALTERNATIVES...-a076784518.

Kantor, Brian, and Phil Lapsley. 1986. *RFC 977 - Network News Transfer Protocol, A Proposed Standard for the Stream-Based Transmission of News*. RFC. Internet Engineering Task Force, February. http://www.ietf.org/rfc/rfc0977.txt.

Karagiannis, T., A. Broido, M. Faloutsos, and K. claffy. 2004. Transport layer identification of P2P traffic. In *Proceedings of the 4th ACM SIGCOMM conference on Internet measurement - IMC \uc0\u8217{}04*, 121. Taormina, Sicily, Italy. doi:10.1145/1028788.1028804. http://portal.acm.org/citation.cfm?doid=1028788.1028804.

Karagiannis, T., K. Papagiannaki, and M. Faloutsos. 2005. BLINC: multilevel traffic classification in the dark. Proceedings of the 2005 conference on Applications, technologies, architectures, and protocols for computer communications (October). doi:10.1145/1080091.1080119. http://doi.acm.org/10.1145/1080091.1080119.

King, David. 2007. Latest content ID tool for YouTube. October. http://tinyurl.com/ref-king-2008.

Labovitz, C., S. Iekel-Johnson, D. McPherson, J. Oberheide, F. Jahanian, and M. Karir. 2009. *ATLAS Internet Observatory 2009 Annual Report*.

Lamy, Jonathan, Cara Duckworth, and Liz Kennedy. 2007. *RIAA Launches New Initiatives Targeting Campus Music Theft*. News Release. The Record Industry Association of America, February 28. http://tinyurl.com/ref-riaa-2007.

———. 2008. *RIAA Continues College Deterrence Campaign Into 2008*. News Release. The Record Industry Association of America. http://tinyurl.com/ref-riaa-2008.

Layton, Robert, and Paul Watters. 2010. *Investiagation into the Extent of Infringing content using BitTorrent networks*. ICSL - Internet Commerce Security Laboratory, April. http://www.afact.org.au/research/bt_report_final.pdf.

Legout, Arnaud, G. Urvoy-Keller, and P. Michiardi. 2006. Rarest first and choke algorithms are enough. In *Proceedings of the 6th ACM SIGCOMM conference on Internet measurement*, 203–216. IMC 06. New York, NY, USA: ACM. doi:10.1145/1177080.1177106.

Liebowitz, Stan J. 2008. Research Note--Testing File Sharing's Impact on Music Album Sales in Cities. *Management Science* 54, no. 4: 852-859. doi:10.1287/mnsc.1070.0833.

Lyn, Jett. 2008. The Deal with User-Generated Content. *Xbiz.com*. April 12. http://www.xbiz.com/articles/92416.

Macavinta, Courtney. 1999. Recording industry sues music start-up, cites black market. *Cnet News*. December 7. http://tinyurl.com/ref-mac-1999.

Madden, Mary. 2007. *Online Video*. Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2007/Online-Video.aspx.

Magrino, Tom. 2008. World of Warcraft hits 11 million. *Gamespot.com*. October 28. http://tinyurl.com/ref-magrino-2008.

McBride, Sarah, and E Smith. 2008. Music Industry to Abandon Mass Suits. *The Wall Street Journal*, December 19. http://online.wsj.com/article/SB122966038836021137.html.

Menta, Richard. 2008. Top P2P Applications: 1.6 Million PCs Rank Them. http://tinyurl.com/ref-menta-2008.

Michel, Norbert J. 2006. The Impact of Digital File Sharing on the Music Industry: An Empirical Analysis. *Topics in Economic Analysis & Policy* 6, no. 1.

Moore, Frances. 2011. *IFPI Digital Music Report 2011*. IFPI. http://www.ifpi.org/content/library/DMR2011.pdf.

MPAA. 2010. *Theatrical Market Statistics 2010*. Report. The Motion Picture Association of America. http://www.mpaa.org/Resources/93bbeb16-0e4d-4b7e-b085-3f41c459f9ac.pdf.

Netflix. 2009. Netflix prize dataset. *Netflixprize.com*. http://www.netflixprize.com/.

Neumayr, Tom, and Jason Roth. 2008. iTunes Store Top Music Retailer in the US. April 3. http://www.apple.com/pr/library/2008/04/03itunes.html.

Oberholzer-Gee, Felix, and Koleman Strumpf. 2007. The Effect of File Sharing on Record Sales: An Empirical Analysis. *Journal of Political Economy* 115, no. 1: 1-42. doi:doi:10.1086/511995.

———. 2009. *File-Sharing and Copyright*. Working Paper 09-132. Harvard Business School.

Oster, Seth. 2008. *MPAA Statement on Motion Picture Industry Losses due to Piracy among College Students*. News Release. MPAA, The Motion Picture Association of America. http://www.wired.com/images_blogs/epicenter/files/mpaa.pdf.

Oswald, Ed. 2006. RIAA Sues LimeWire Over Piracy. *Betanews.com*. August 4. http://tinyurl.com/ref-oswald-2006.

Packeteer. 2007. Applications, Protocols, and Services Classified by PacketWise 7.3. https://bto.bluecoat.com/packetguide/7.3/reference/services.htm.

Parlement Français. 2009. *Loi favorisant la diffusion et la protection de la création sur Internet*. June 12. http://www.senat.fr/dossier-legislatif/pjl07-405.html.

Purcell, Kristen. 2010. *The State of Online Video*. Pew Internet & American Life Project. http://www.pewinternet.org/Reports/2010/State-of-Online-Video.aspx.

RIAA. 2007. *Piracy Online*. News Release. The Recording Industry Association of America. http://tinyurl.com/ref-riaa-2007b.

Rob, Rafael, and Joel Waldfogel. 2006. Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students. *The Journal of Law and Economics* 49, no. 1: 29-62. doi:doi:10.1086/430809.

Sandvine. 2009. *2009 Global Broadband Phenomena*. http://www.sandvine.com/downloads/documents/2009%20Global%20Broadband%20Phenomena%20-%20Executive%20Summary.pdf.

Schaefer, Lyndsay. 2011. *DEG Year-end 2010 Home Entertainment Report*. Press Release. The Digital Entertainment Group. http://www.degonline.org/pressreleases/2011/f_Q410.pdf.

Schulze, Hendrik, and Klaus Mochalski. 2007. *iPoque Internet Study 2007*. iPoque. http://www.ipoque.com/resources/internet-studies/internet-study-2007.

———. 2009. *iPoque Internet Study 2008/2009*. iPoque. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009.

Singel, Ryan. 2009. Peer-to-Peer Passé, Report Finds. *Wired*. http://www.wired.com/epicenter/2009/10/p2p-dying.

Sjoden, Kerstin. 2009. The Pirate Bay's Anonymity Service Signs 100,000 Users Pre-Launch. June. http://www.wired.com/threatlevel/2009/04/the-pirate-bays/.

Smitelli, Scott. 2009. Fun with YouTube's Audio Content ID System. March. http://www.csh.rit.edu/~parallax/.

Smith, J. 2006. ISU staff speaks in Washington on issues of campus piracy. *Daily Vidette at Illinois State University*, October 3. http://is.gd/eFbpI.

Smith, L, George Miller, Howard McKeon, Howard Berman, and Howard Coble. 2007. *Letter sent to universities on May 1, 2007*. http://tinyurl.com/ref-letter-2007.

Smith, MD, and Rahul Telang. 2008. *Competing with free: The impact of movie broadcasts in DVD sales and Internet piracy*. Working Paper. Carnegie Mellon University. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1028306.

Smith, M, and Rahul Telang. 2007. *Research Note: Internet Exchanges for Used Digital Goods: Empirical Analysis and Managerial Implications*. Research Note. Tepper School of Business, Carnegie Mellon University.

Swasko, Mick, Deanne St. John, Eric Strand, and Mary Yurgil. 2007. Being a pirate is against the law. *Daily Vidette at Illinois State University*, February 21. http://is.gd/eFbUA.

Tanaka, Tatsou. 2004. Does file sharing reduce CD sales?: A case of Japan. In *Conference in IT Innovation*.

U.K. Parliament. 2010. *Digital Economy Act 2010. 2010 c 24*. April 8. http://www.legislation.gov.uk/ukpga/2010/24/contents.

U.S. Congress. 2003. *Peer-to-Peer Piracy (P2P) on University Campuses*. Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 108th Cong. 1st sess. SN 2, February 26. http://purl.access.gpo.gov/GPO/LPS59277.

———. 2004. *Peer-to-Peer Piracy (P2P) on University Campuses: An Update*. Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 108th Cong. 2nd sess. SN 112, October 5. http://purl.access.gpo.gov/GPO/LPS58080.

———. 2005. *Reducing Peer-to-Peer (P2P) Piracy on University Campuses: A Progress Update*. Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 109th Cong. 2nd sess. SN 109-56: U.S. House of Representatives,, September 22. http://purl.access.gpo.gov/GPO/LPS66466.

———. 2006. *The Internet and the College Campus: How the Entertainment Industry and Higher Education are Working to Combat Illegal Piracy*. Committee on Education and the Workforce. Subcommittee on 21st Century Competitiveness. 109th Cong., 2nd sess. SN 109-58, September 26. http://purl.access.gpo.gov/GPO/LPS76269.

———. 2007a. *An Update: Piracy on University Networks*. Committee on the Judiciary. Subcommittee on Courts, the Internet, and Intellectual Property. 110th Cong., 1st sess., SN 110-29, March 8. http://purl.access.gpo.gov/GPO/LPS85994.

———. 2007b. *The Role of Technology in Reducing Illegal File-sharing: A University Perspective*. Committee on Science and Technology. 2007. 110 Cong., 1st sess. SN 110-34, June 5. http://purl.access.gpo.gov/GPO/LPS88842.

US. Congress. 2009. *Communications Networks and Consumer Privacy: Recent Developments*. Committee on Energy and Commerce, Subcommittee on Communications, Technology, and the Internet: U.S. House of Representatives,. http://goo.gl/UENY.

Van der Sar, Ernesto. 2009a. More BitTorrent Users Go Anonymous. http://tinyurl.com/ref-tf-2009b.

———. 2009b. Top 10 Most Pirated Movies on BitTorrent. July. http://tinyurl.com/ref-tf-2009.

Williams, Becky. 2009. Thousands of World of Warcraft fans descend on southern California for Blizzard's epic gaming convention. *Telegraph.co.uk*. August 24. http://tinyurl.com/ref-williams-2009.

Zentner, Alejandro. 2006. Measuring the Effect of File Sharing on Music Purchases. *The Journal of Law and Economics* 49, no. 1: 63-90. doi:doi:10.1086/501082.

Zhang, Mia. 2009. *Internet Traffic Classification*. CAIDA: The Cooperative Association for Internet Data Analysis. http://www.caida.org/research/traffic-analysis/classification-overview/.

Appendix A Grouping of Majors by Area of Study

Category	Major
Gene al	General Student
S cial ciences	Commun cation
	Economics
	Anthropo o y
	Pol t cal Sc ence
	Mass Communication
	Psychology
	Criminal Justice Sciences
	Sociology
	Public Relations
	Applied Economics
Sci & Tech	**Information Systems***
	Computer Science*
	Industrial Technology
	Exercise Science
	Chemistry
	Telecommunications Mgmt*
	Geology
	Physics
	Biological Sciences
	Agriculture
	Biochem/Molecular Biology
	Mathematics
	Technology*
	Geography
	Information Systems*
Humanities	French
	German
	English
	History
	Journalism
	Communication Studies
	Spanish
	Philosophy
	Languages Lit & Cultures
	Historical Archaeology

Category	Major
Health	Safety
	Environmental Health
	Health Education
	Health Information Mgmt
	Nursing (bsn)
	Bachelor Of Social Work
	Clinical Laboratory Sci
	Social Work
	Kinesiology & Recreation
	Athletic Training
Education	Interdisciplinary Studies
	Coll Stud Personnel Admin
	University Studies
	Special Education
	Technology Education*
	Middle Level Teacher Edu
	Physical Education
	Speech Path &Audiology
	Elementary Education
	Early Childhood Education
	Educational Admin
	Alt Secondary Certificate
Business	Recreation & Park Admin.
	International Business
	Management
	Business Teacher Edu
	Finance
	Accountancy
	Business Administration
	Family & Consumer Science
	Insurance
	Business Information Syst*
	Marketing
	Accountancy Bs/Mpa
	Master Of Business Admin
Arts & Music	Art
	Bachelor Of Music
	Arts Technology
	Theatre
	Music-Liberal Arts Ba/Bs
	Music
	Bachelor Of Music Educ

* Majors considered as IT Savvy.

Appendix B Meaning of Tags Present in Video Torrent Titles

Adapted from http://www.vcdq.com/faq#cam

9.1.1.1.1 CAM

“A cam version is capture at a movie theater usually with a digital video camera. A mini tripod is sometimes used, but a lot of the time this will not be possible, so the camera may shake. Also seating placement isn't always ideal and it might be filmed from an angle. If cropped properly, this is hard to tell unless there's text on the screen, but a lot of times these are left with triangular borders on the top and bottom of the screen. Sound is taken from the onboard microphone of the camera, and especially in comedies, laughter can often be heard during the film. Due to these factors picture and sound quality are usually quite poor[…]”

9.1.1.1.2 TELESYNC (TS)

“A telesync is the same spec as a CAM except it uses an external audio source (most likely an audio jack in the chair for the hearing impaired). A direct audio source does not ensure a good quality audio source, as a lot of background noise can interfere. A lot of the times a telesync is filmed in an empty cinema or from the projection booth with a professional camera, giving a better picture quality. Quality ranges drastically […]. A high percentage of Telesyncs are CAMs that have been mislabeled.

9.1.1.1.3 TELECINE (TC)

“A telecine machine copies the film digitally from the reels. Sound and picture should be very good, but due to the equipment involved and cost telecines are fairly uncommon. Generally the film will be in correct aspect ratio, although 4:3 telecines have existed. […] Most R5 releases are Telecines.”

9.1.1.1.4 *R5*

“Typically these are high quality Telecines intended for the East European market (released in Russian language only) to combat piracy in that region. Ironically, these movies then have the English audio track from another source (such as a CAM) dubbed over them and get released. Until recently these releases have been tagged as R5 but we're starting to see similar sources from other regions tagged as R3 or R6. The number is derived from the DVD region the source came from.”

9.1.1.1.5 SCREENER (SCR)

“A VHS tape, sent to rental stores, and various other places for promotional use. A screener is supplied on a VHS tape, and is usually in a 4:3 (full screen) a/r, although letterboxed screeners are sometimes found. The main draw

back is a "ticker" (a message that scrolls past at the bottom of the screen, with the copyright and anti-piracy telephone number). Also, if the tape contains any serial numbers, or any other markings that could lead to the source of the tape, these will have to be blocked, usually with a black mark over the section. This is sometimes only for a few seconds, but unfortunately on some copies this will last for the entire film, and some can be quite big. Depending on the equipment used, screener quality can range from excellent if done from a MASTER copy, to very poor if done on an old VHS recorder thru poor capture equipment on a copied tape."

9.1.1.1.6 HDTV

"Commonly used to tag high definition TV rips."

9.1.1.1.7 DVD-SCREENER (DVDscr)

"Same premise as a screener, but transferred from DVD. Usually letterbox format but without the extras that a retail DVD would contain. The ticker is not usually in the black bars, and will disrupt the viewing. If the ripper has any skill, a DVDscr should be very good."

9.1.1.1.8 DVDRip

"A copy of the retail DVD and should be excellent quality with no markers/tickers. DVD screeners are sometimes mislabed as DVD rips"

9.1.1.1.9 WORKPRINT (WP)

"A workprint is a copy of the film that has not been finished. It can be missing scenes, music, and quality can range from excellent to very poor. Some WPs are very different from the final print (Men In Black is missing all the aliens, and has actors in their places) and others can contain extra scenes (Jay and Silent Bob). WPs can be nice additions to the collection once a good quality final has been obtained."

9.1.1.1.10 BLURAY

"Blu-ray Disc (also known as BD, BDRIP or Blu-ray) is an optical disc storage medium designed to supersede the standard DVD format. Blu-ray discs are in high definition format and as such are the best quality source commonly available."