Dimensions of P2P and digital piracy in a university campus

Alexandre M. Mateus[1] and Jon M. Peha[2]

Carnegie Mellon University

Abstract

This article presents findings from the first large-scale quantitative assessment of Peer-to-Peer (P2P) exchanges of copyrighted material on a college campus based on actual observation. Through passive monitoring and deep packet inspection (DPI), we assess the extent to which P2P is used to transfer copyrighted material. We also characterize the demographics of P2P users, the relative popularity of the material, and how the burden on the campus network varies over time. We found that at least 51% of students living on campus engaged in P2P, at least 42% attempted to transfer copyrighted material, and the mean number of copyrighted media titles whose transfer is attempted per week was at least 6 per monitored student. Some students use P2P legally, e.g. to transfer Linux software or non-copyrighted adult material, but we found no evidence that large numbers of students use P2P for these legal purposes and not to transfer copyrighted material. Students of all genders, ages, classes and majors engaged in file sharing, to the extent that demographics were not helpful in identifying likely file-sharers so as to target interventions. This study also provides lessons for those who would use DPI technology to reduce illegal use of P2P. If given enough weeks to observe, current technology is effective at identifying users who attempt to transfer copyrighted material, provided that their traffic is identifiable as P2P. Thus, DPI can be used to estimate the extent of piracy, and to notify individuals who may be violating copyright law. However, encryption is available and can be easily activated in most P2P clients. Once turned on, encryption prevents DPI from detecting whether transferred material is copyrighted, rendering it ineffective. If DPI is used for copyright enforcement that includes imposition of penalties, then P2P users or P2P developers may have the incentive to use encryption as a way of evading detection.

---

[1] Corresponding author: Alexandre M. Mateus, Ph.D. Candidate, Department of Engineering & Public Policy, Carnegie Mellon University, amateus@cmu.edu, www.andrew.cmu.edu/~amateus

[2] Jon M. Peha, Associate Director of the Center for Wireless & Broadband Networking, Professor in the Department of Engineering & Public Policy and the Department of Electrical & Computer Engineering, Carnegie Mellon University, peha@cmu.edu, www.ece.cmu.edu/~peha

1 Introduction

The Internet is now one of the main ways of obtaining music and movies [4]. Consumers can access such media legally through on-line music and video stores, or illegally using Peer-to-Peer (P2P) file sharing networks that allow them to freely share information. P2P networks are not the only means of illegally transferring copyrighted material and not all P2P transfers are illegal, but P2P is considered the main vehicle for online media piracy [39]. Digital piracy affects music and movie industries, who claim that it causes billions of dollars in lost sales [44]. In response, those copyright industries have sued P2P software companies and P2P users, and lobbied for legislation on this matter. According to the industry, university students account for a big share of online copyright violations and campus networks are "privileged"[3] repositories of material that is accessed by outside users [39]. This made university students preferred targets for lawsuits [41], and put campus piracy on the Congressional agenda [11, 24, 46, 48].

This paper focuses on P2P in university campuses. One primary objective of this research is to determine the extent of legal and illegal P2P usage on a university campus, in part to help policymakers and copyright-holders assess the importance of intervention. Another primary objective is to understand the capabilities and limitations of current technology for detecting and reducing copyright violations, including use of this technology to block transfers believed to be illegal or to gather evidence for legal action, as some have proposed. This may help universities as well as ISPs decide whether to use the technology, and may help policymakers decide whether to require use of such technology. For example, France mandated that ISPs monitor their networks and cut off Internet access to those costumers detected sharing copyrighted files via P2P [6, 8] and the US Congress is considering legislation that would require universities to invest in monitoring technology for enforcement against illegal transfers of copyrighted material over P2P under the penalty of loosing federal funding otherwise [3, 16].

Apart from the principal objectives stated above, this paper also provides useful information on other dimensions of P2P usage. In particular, it shows the differences (and strong similarities) of P2P usage across demographic groups, as strong differences might help in the construction of more targeted interventions. It examines P2P usage patterns over time, which may help

[3] Privileged in the sense that, in university campus networks, users typically have greater upload capacity than in ISP broadband connections, which makes it more attractive for ISP customers to use university file-sharers as sources .

campus network managers and ISPs to better manage their networks to avoid congestion. Finally, it characterizes the material transferred using P2P, as this may shed light on how to construct competing legal services.

There have been previous assessments of the extent of P2P file-sharing on university campuses. A survey in the 2003-2004 academic year showed that about 58% of the 600 surveyed students in two large colleges[4] admitted to file-sharing, and that about 40% of their music collection was composed of file-shared music [48]. Other figures advanced by the Record Industry Association of America (RIAA) [26], citing a survey by Student Monitor from Spring 2006, stated that more than half of college students downloaded music and movies illegally. Also reported by the RIAA, according to the market research firm NPD, college students "get more of their music from illegal peer-to-peer than the rest of the population: 25 percent vs.16% percent" and "college students alone accounted for more than 1.3 billion illegal music downloads in 2006" [27]. More recently, members of the House Judiciary Committee and the Education and Labor Committee in the US Congress addressed a letter to the 19 top violating universities (according to a list compiled by the industry representatives) requesting the response to a list of questions on current practices regarding illegal file-sharing on campus and asking for urgent action [46]. This letter restated the above numbers advanced by RIAA and added that 21% of all P2P users were college students and that their P2P file sharing accounted for 15% of movie industry domestic piracy losses in 2006 [22, 33], students also reported illegally obtaining more than 2/3 of all music they acquired.

Whereas the previous studies cited above are based on surveys, this paper presents results from the first large-scale quantitative assessment of online media transfers based on actual observation of P2P exchanges on a college campus. Thus, our results do not depend on the memories and openness of survey respondents. Moreover, it is possible to observe things that users may not know, such as the volume of P2P traffic, and the time of transfers.

The remainder of this paper is organized in 5 sections. Section 2 briefly discusses relevant background and section 3 describes the network monitoring methodology and measures to guarantee the privacy of users. Section 4 presents a general summary of the collected data and defines the main concepts used in the analysis. Section 5 presents the paper's results and is divided in three parts. First it focuses on our principal objectives, covering the extent of P2P

[4] Florida Atlantic University (about 26,000 students) and University of Nevada Las Vegas (about 28,000 students)

usage on campus, and the capabilities and limitations of monitoring technology for detecting and reducing copyright violations, with a discussion of why our results are lower bounds. Second, it presents more detailed results on pervasiveness and intensity of P2P usage, covering differences across demographics and over time. And finally, it focuses on the type of transferred material, both copyrighted and not copyrighted. Section 6 closes the paper with our conclusions and the policy implications that we extract from them.

2 Background

2.1 P2P and Sharing of Copyrighted Material

Using current technology, it is easy to obtain media from the Internet. Consumers can obtain legal media from download[5] or streaming[6] services, and illegal media from P2P networks, central repositories[7], DarkNets[8] or newsgroups, or even from user-fed websites where uploaded material is not checked for copyright. This research focuses on P2P networks.

P2P technology can be used for legal and illegal purposes. The advantage of P2P networks to distribute information is that all the users share the burden of transmitting material to other users, thus releasing content providers from bearing the costs of transmitting all the material they distribute. Although there are many examples of legal applications of P2P, such as BitTorrent legal distribution[9], Skype[10] or Joost[11], many P2P networks are used for online piracy. Napster was the first of them. It was released in 1999 and peaked at 26 million users worldwide

[5] Download services allow users to transfer and keep the media on their computers, typically charging a per-title or per-album price. Examples of these are Apple's iTunes store or Amazon's MP3 download service.
[6] Streaming services allow users to play the media directly from on-line servers, typically charging a periodical subscription fee. Examples of these are RealNetwork's Rhapsody or Napster's on-demand service.
[7] Typically, central repositories are setup in servers that can be accessed by a restricted group of users who are given access credentials. Access is performed using well-known protocols, such as FTP (File Transfer Protocol), and does not require client software developed specifically for the purpose.
[8] DarkNets are private virtual networks in which users connect and share material only with other users they trust. The term DarkNet was first used in this context in [7] and also included general-access P2P networks, but due to their dimension and more open nature, most P2P networks are now treated separately from DarkNets. In the context of university campuses, the term DarkNet also refers to file-sharing virtual networks that operate strictly within campus, therefore dark to the outside world.
[9] BitTorrent is the name of three distinct but related entities. It is the name of a P2P protocol originally developed to allow software developers to easily and cheaply distribute their applications, a purpose for which it is still widely used nowadays. It is the name of a client application that implements the protocol. And finally, it is the name of a company that provides legal information distribution services using the protocol as the underlying technology.
[10] Skype is a voice over IP service that allows the placement of phone calls over the Internet (http://www.skype.com)
[11] Joost uses P2P architecture to deliver near-TV resolution images (http://www.joost.com).

in February 2001 [19], after which it was shut down by federal order in the sequence of a lawsuit by the Record Industry Association of America (RIAA) [42]. Napster's shut down allowed other P2P networks to emerge. Today's top P2P networks are BitTorrent, Gnutella, Ares and eDonkey. Concerning popularity, BitTorrent seems to be the most popular network in terms of traffic, accounting for 30% of all Internet traffic in the end of 2004 [20], while LimeWire, a Gnutella client, leads in terms of downloads, with 62% of all P2P downloads in the end of 2005[12] [21]. As for number of users, according to estimates in [35], the monthly average number of simultaneous P2P users grew from 4 to 9 million between 2003 and 2005.

P2P accounts for a big fraction of Internet traffic, with estimates going from 37% [15] to 50-60% or even 80% for some Internet Service Providers (ISPs) [31]. To maintain P2P traffic within acceptable boundaries and provide appropriate level of service for other applications, ISPs have used traffic shaping[13] (some examples in [18, 23, 37]). Every development in P2P traffic detection by producers of traffic shaping appliances has had a response from P2P developers in the form of changes to the protocols as to avoid detection, originating a true arms race between monitoring technology and P2P developments to dodge detection. The latest development on the detection front came from a traffic shaper company that claimed to detect encrypted P2P [17]. However, it remains a challenge for network monitoring to detect encrypted P2P transmissions, and impossible to unveil what material is transferred therein.

2.2 Copyright industry, Congress and the Universities

Copyright industries are the main victims of illegal file sharing. Industry representatives claim that it costs billions of dollars in lost sales and causes thousands of lost jobs [44]. Recent research found links between file sharing and the decline in record sales [28], supporting the argument that file sharing causes harm to copyright owners [43, 50].

According to U.S. law [12], transferring copyright-protected works without the authority of the copyright owner is an infringement of the owner's exclusive rights, and those who aid and

[12] These numbers can be reconciled by taking into account that LimeWire is mostly used to transfer music files (higher number of smaller downloads) while BitTorrent is more popular for movies (less transfers, but more traffic per transfer).

[13] Traffic shaping is a technique commonly used in large networks to maintain certain types of traffic, such as P2P, within boundaries in an attempt to minimize delay of other types of traffic. To implement shaping, the type of traffic in all transfers needs to be identified and, in case the type is one of the target types and the bandwidth threshold has been exceeded, then the transfer is either dropped or delayed. Traffic type identification techniques range from simple detection by port number to deep packet inspection and, more recently, analysis of communication patterns.

support copyright infringement are as culpable as the infringers. This means that both P2P users and P2P developers may be accused of copyright infringement. Concerning ISPs, the Digital Millennium Copyright Act (DMCA) [14] has provisions limiting ISP liability under certain circumstances, but to obtain such "safe harbor" protection, ISPs must respond to subpoenas and provide identification of subscribers accused of violation.

The music industry, through the RIAA, used these legal provisions in a series of court battles against P2P companies [29, 34] and users [41]. To unveil the identity of users, RIAA traditionally used the subpoena mechanism in DMCA. In the case of universities, since early 2007 the music industry started using "pre-litigation settlement letters" requesting that infringing students be identified and that the letter be forwarded to them [9]. Since these letters were not legally binding, some universities ignored them, while others forwarded them to students [41]. Upon reception of the letters, students could avoid court action and settle the case over the phone or using a website[14].

Congress also focused on online copyright piracy in universities, holding six hearings on the issue since 2003 [11, 24, 46, 48] and discussing possible interventions to deal with it [3, 16]. Abroad, the focus lies mostly on ISPs. In the E.U., France passed legislation requiring ISPs to police networks and disconnect users detected transferring copyrighted material [6, 8], and in the U.K., the possibility of similar legislation is a subject of dispute between ISPs, the copyright industry and government [2]. All this activity, both in the U.S. Congress and in some of the larger E.U. countries indicates potential for policy change.

3 Monitoring Methodology

This research was performed on data collected through the Digital Citizen Project, a broader project undertaken by the Illinois State University (ISU) "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" [13]. The Digital Citizen Project aims to address ethical and legal issues by educating college students, putting in place self monitoring and enforcement and providing multiple legal digital media services [13]. It also seeks to investigate K-16 use of peer-to-peer software, to develop a

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

curriculum component to combat illegal downloads, and to work with industry leaders to create educational fair use media definitions and faster copyright use [13]. The project's ultimate goal is to "create a nationally recognized program that will be cost-effective and is replicable on other college and university campuses" [13]. 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.

3.1 Network Monitoring

The ISU network serves the entire campus population. This network connects to the Internet using two commodity Internet Service Providers (ISPs) and Internet 2. ISU uses traffic shaping in the connection to its commodity ISPs and does not impose limits on the amount of traffic generated by each network user. There are several sub networks in the ISU network. ResNet is the sub network that students connect to in their dormitories. ResNet users purchase network access from ISU, which allows one wired connection per user in the dorm room. Wireless routers are not allowed in ResNet, a policy enforced by the ResNet management.

Network monitoring was performed by two commercially available monitoring appliances that use DPI to analyze packets: Packeteer PacketShaper[15] (from now on referred to as Packeteer) and Audible Magic CopySense[16] (from now on referred to as Audible Magic). Both devices log relevant attributes of transmissions between ResNet users and parties outside the campus network, provided that traffic is routed using commodity ISPs. Packeteer had already been deployed before this project to perform traffic shaping as described above The device classifies communications in one of over 500 classes[17] according to the type of traffic that composes them. This device neither examines nor retains the actual contents of the communications sessions.

The Audible Magic device was purchased to enforce ISU policy before CMU got involved. Audible Magic uses header information to identify P2P streams. Within those P2P streams, Audible Magic identifies copyrighted media in real time as the material arrives by trying to match

[15] For more information on the features of Packeteer PacketShapper, refer to http://www.packeteer.com/products/packetshaper/
[16] For more information on the features of Audible Magic CopySense, refer to http://www.Audible Magic.com/products-services/copysense/
[17] 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 [36].

the transferred material against a database of audio fingerprints of copyrighted media titles[18] or hash codes[19] used to identify files in P2P networks. The device does not retain any portion of the transmission, but it does record which copyrighted material in the database was matched. When the material being transferred cannot be matched against anything in the database, Audible Magic only records a piece of the metadata incorporated in the transfer (typically the name of the file being transferred).

Audible Magic logs information on communications in the form of *events*. An event corresponds to one or more consecutive TCP or UDP[20] 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 identified with any media title in Audible Magic's database. Hence, an Audible Magic event means that two peers in a P2P network have exchanged or attempted to exchange a given amount of information (either from an identified copyrighted media title or information that could not be identified as belonging to any copyrighted media title present in Audible Magic's database) over a set of consecutive TCP sessions or consecutive UDP sessions. Audible Magic aggregates consecutive communication sessions in single events for the purpose of simplicity in the logging process.

3.2 Connection of Monitored Activity to Users and Devices

The identification of the network user and device responsible for each detected online activity was implemented using data from several network management databases also collected from the ISU network. For each collected data record, which contains one IP address internal to the network, device information (the device's MAC address[21]) is obtained by performing a lookup in

[18] One technique used by Audible Magic to identify copyrighted material is audio fingerprinting. Audible Magic collects a sample of the audio track of the material that is being transferred (typically 20 seconds) 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.

[19] 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 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 Audible Magic to identify copyrighted material because it allows for faster comparisons and earlier detection than the technique based on audio fingerprinting.

[20] 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.

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

the DHCP[22] lease logs using the IP address of the monitored activity and the time when the activity occurred.

Information about the user that performed each activity consists of the user's University Login Identification (ULID)[23], birth year, gender, major, role (student, staff, faculty), and university title (freshman, sophomore, junior, etc.). This information is retrieved from the ISU directory using the ULID as key. To obtain the ULID associated with each monitored record, different network management databases need to be queried depending on the type of connection used to perform the activity. This procedure assumes that the user that registered the device used to perform an online activity was the one responsible for that activity.

3.3 Privacy Protection

The collection of monitoring data was performed in accordance with the Digital Citizen Project policy guidelines, which include measures to protect the privacy of monitored users such as the following. Data collection was performed at ISU 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 identity 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[24]. Both the data collection process and the generation of pseudonyms were performed in an automated fashion without human intervention, so no human would ever see the raw data, and the keys used in the hashing function were destroyed. Also, the people that controlled the monitoring and anonymization processes were people that would otherwise have access to the raw data (network management team at ISU) and were precluded from analyzing the anonymized data. Conversely, CMU researchers who analyzed the resulting data were not allowed to observe raw data prior to anonymization. Thus, the CMU researchers who performed the analysis had no possible way to connect any of the data to a specific person, computer, or location on campus.

---

[22] 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.
[23] University Logon ID, a unique identifier assigned to each person in the ISU campus.
[24] 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.

All the research described in this paper was approved by both the ISU Institutional Review Board (IRB) and the CMU IRB.

4 Overview of Collected Data and Main Definitions

Network monitoring in April 2007 produced useful data for 620 of the 720 hours in the month, adding to 25 days with 24 full hours of data and 3 weeks with 7 full days of data. Two data sets were collected, one with hourly summaries of traffic detected by Packeteer[25] and another with 24 million P2P communication events detected by Audible Magic[26].

By 2006, ISU had announced to its campus community and the outside world that it planned to manage its own system to monitor traffic over the campus network, and enforce stated policies against use of the campus network for the illegal transfer of copyrighted material [45]. Indeed, the ISU newspaper carried a number of articles and editorials about the university's monitoring plan and a variety of related legal and ethical issues. It is therefore possible that this public discussion deterred some people from engaging in illegal transfers, thereby affecting our results.

The results presented in this paper were obtained from analysis of the Audible Magic data set and are lower bounds for the P2P activity that took place on campus because, as we can observe in Figure 1, Audible Magic does not detect a fair amount of P2P traffic. The figure compares the data collected by Packeteer and by Audible Magic in terms of detected number of bytes and shows that Packeteer detects almost 13% of all observed traffic as being P2P[27] while Audible Magic detects about 7.5%, less than two thirds of what Packeteer detects[28].

[25] Data contains one summary record for each traffic type in each hour, with total number of bytes and total number of TCP sessions or UDP pseudo-sessions detected during that hour.
[26] See definition of communication event detected by Audible Magic in section 3.1, page 8
[27] This looks small when compared to estimates of at least 37% for commercial ISPs [15]. The smaller percentage may be due to ISU's traffic shaping policy that limits P2P to 5% of the available bandwidth. The 12% average is calculated out of the number of bytes transferred in each hour. In hours that are not busy for other types of traffic, the maximum 5% of bandwidth taken by P2P actually corresponds to a much higher percentage of the transferred traffic.
[28] Packeteer and Audible Magic count packet bytes at different layers of the protocol stack. Packeteer counts total layer 3 bytes including TCP+IP headers while Audible Magic counts only layer 4 payloads, i.e., without TCP+IP headers. To make the comparison fair, we assume 44 bytes per IP+TCP header in each P2P packet detected by Packeteer and scale the percentage of P2P detected by Audible Magic accordingly.

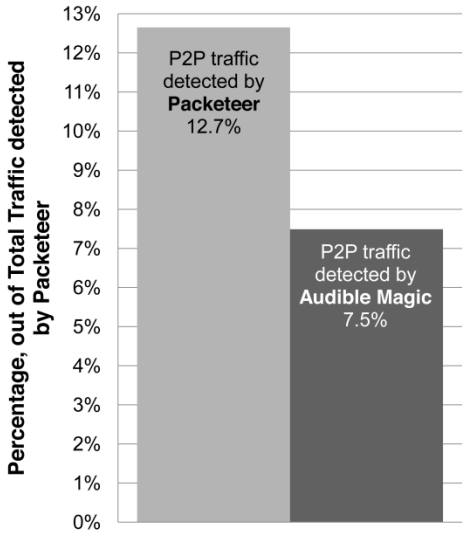


Figure 1. Average percentage of all network traffic detected as being P2P by Packeteer and Audible Magic. The bars in the figure represent hourly averages of P2P bytes detected by Packeteer and by Audible Magic as a percentage the total traffic detected by Packeteer.

We centered the analysis on two types of activity: the usage of P2P networks and the usage of P2P networks to transfer copyright-protected media. Clearly, the latter activities are a subset of the former. To avoid misinterpretation of the results in the remainder of this paper, we clearly define the concepts that were used in the analysis.

We define a P2P activity as a communication event detected by Audible Magic, in which information is transferred using a P2P protocol. A P2P user is a network user detected doing at least one P2P activity in the monitored period. A Detected Attempt to Transfer Copyrighted Media (DATCoM) is a detected transfer or transfer attempt using a P2P protocol, of media identified as being protected by copyright, which corresponds to an event detected by Audible Magic in which the exchanged material was identified as being copyrighted. The set of DATCoMs is a subset of the set of P2P activities. A DATCoM user is a user who is detected doing at least one DATCoM in the monitored period.

Several remarks are due regarding the above definitions. First, it is important to emphasize that analysis is performed only on the activities that could be detected by Audible Magic and, as mentioned above, this leads to lower bound results.

Second, all P2P requests for files whose hash code is part of Audible Magic's database are classified as transfers of copyrighted material, although some of these requests are mere attempts that eventually end up as failed connections with no actual information transferred. However, the fact that a P2P user sends/receives a request for copyrighted material to/from

another P2P user means that the receiver is or was sharing that material on the network and that the sender tried to obtain it. Whether or not merely making material available to share constitutes in itself a copyright violation is currently the subject of a legal dispute [5] that is beyond the scope of this paper.

Third, the definition of DATCoM does not distinguish downloads from uploads. Audible Magic only provides conclusive information on direction of transfers for 14% of detected activities, which makes it impossible to extract any significant or valid conclusion regarding downloads vs. uploads. This remark is particularly important because statistics about ratio of uploads and downloads drawn from the same pool of data were previously published [40]. We believe such statistics should be disregarded.

Fourth, in the present analysis we do not distinguish audio from video in the pool of detected copyrighted material. This is because Audible Magic does not report information on the mime type of identified material, providing only information on the title of the movie/song and in some occasions the author.

Finally, a remark regarding the distinction between a DATCoM and a copyright violation as understood in U.S. legal terms. According to U.S. copyright law, one copyright violation is the transfer of a copyrighted media title between two people without the permission of the copyright holder [12, 14]. A simplistic interpretation of this would see one copyright violation for each pair of communicating parties transferring each distinct media title. However, since we only have information about anonymized IP addresses of parties outside the ISU network (which do not map univocally to people), it is impossible to identify pairs of communicating parties. Also, it is debatable whether or not certain transfers over P2P networks – for instance, when the user that downloads the material already owns a legal copy – are copyright violations or “fair use” [12, paragraph 107]. Our data collected from the network cannot tell us whether or not the transferred material will be used in any way that can be considered “fair use”. Finally, the fact that a DATCoM can be a mere transfer attempt drives the concept further apart from that of a copyright violation. The above examples clearly show that there is not a direct correspondence between DATCoMs and copyright violations: not all DATCoMs are copyright violations, although most of them are; and not all copyright violations occurring on campus were captured as DATCoMs.

5 Characterization of P2P Activity on Campus

5.1 Extent of P2P Activity and Limitations of Monitoring Technology

During the monitored period, Audible Magic detected users in ResNet performing 119,073 DATCoMs, which represented 23,819 distinct media titles. We detected P2P activity by about half the students living on campus (51%) and the great majority of those were also detected transferring or attempting to transfer copyrighted material (42% of ResNet students, or 82% of P2P users). This is conveyed in Table 1, which also shows that during the monitored period, the average number of copyrighted media titles detected being transferred per student was 18. These figures remain unaltered when we restrict the universe of analyzed DATCoMs only to those where there was actual transfer of information[29], which means that users that were detected attempting to transfer material, eventually transferred material.

Table 1: Proportion of detected P2P users and DATCoM users out of all ResNet students and average number of detected copyrighted titles per ResNet student in the monitoring period (95% CI in parenthesis). Results considering all DATCoMs and considering only DATCoMs with transferred media.

	Proportion of P2P Users	Proportion of DATCoM Users	Average copyrighted titles detected per user in the period
All DATCoMs	**51%** (49% - 52%)	**42%** (40% - 43%)	**18** (17 - 20)
Only DATCoMs with transferred media		**42%** (40% - 43%)	**18** (17 – 19)

The above results are obtained from an incomplete record of all the activity that occurred on campus and are therefore lower bounds on the actual quantities that we want to assess. This downward bias is due to limitations of monitoring technology that happen both at the level of detection of P2P traffic and at the level of detection of copyrighted material within P2P traffic. At the level of P2P detection, as we had seen in Figure 1, Audible Magic detects less than two thirds of the P2P traffic that Packeteer detects. Furthermore, Audible Magic does not detect encrypted P2P as being P2P at all. Concerning detection of copyrighted material within P2P traffic, we will focus below on several conditions that cause Audible Magic to fail to identify transferred material as copyrighted. In spite of these, only 9% of all ResNet users (or 18% of P2P users) are detected performing P2P but not attempting to transfer copyrighted material.

[29] Some detected DATCoMs had a number of bytes that allowed only for the P2P protocol control traffic. Those correspond, for instance, to failed connections or requests to which there was no response. We consider a DATCoM to actually transfer information when at least 100 bytes are transferred.

To evaluate how many DATCoM users Audible Magic is possibly missing out of detected P2P users; the percentage of detected P2P users and detected DATCoM users was summarized for different time intervals in Figure 2, and the ratio of detected DATCoM to P2P users in Figure 3. As interval duration increases, the proportion of DATCoM to P2P users also increases, which means that, given enough time, Audible Magic will eventually detect most users with DATCoMs out of the P2P users that it detects, but that it misses most of DATCoMs in each short period.

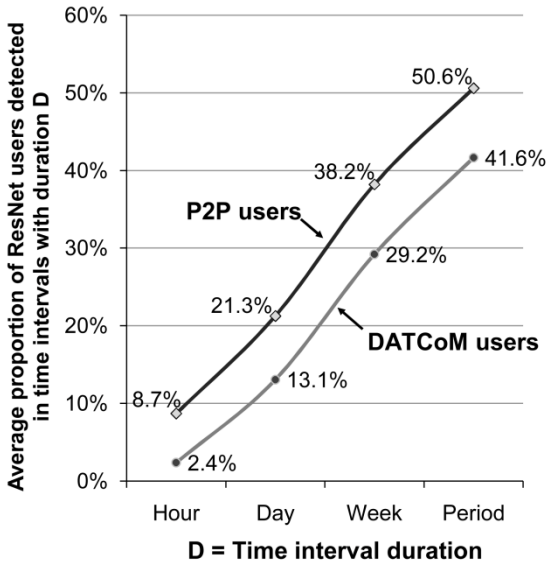


Figure 2. Proportion of ResNet users detected doing P2P or DATCoMs. For each time interval duration, averages of the proportions were calculated over all periods of that duration in the data set.

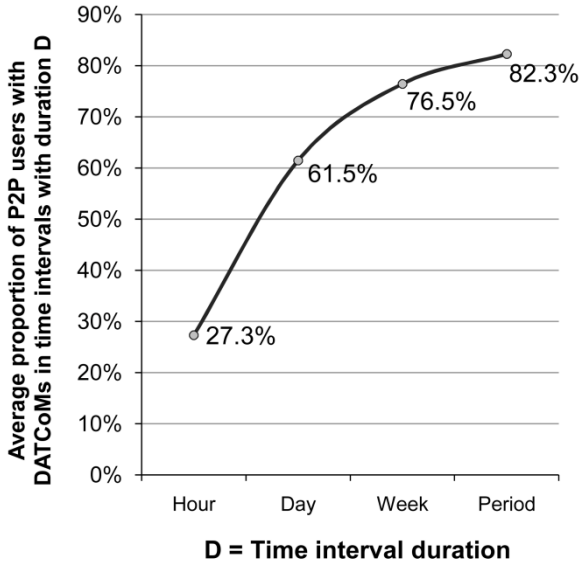


Figure 3. Proportion of P2P users detected performing DATCoMs. For each time interval duration, the average of the proportion was calculated over all periods of that duration in the data set.

Audible Magic only identifies P2P traffic as copyrighted if the transferred material belongs to music or movie titles present in a central database, which automatically excludes other types of copyrighted media such as software or video games. Hence, one reason why P2P users are not detected performing DATCoMs is if such users only transfer media absent from the Audible Magic database. This happens for material legally distributed over P2P – such as Linux distributions – or for material whose distribution over P2P is illegal, but that is not present in the database – such as adult movies. Analysis in section 5.3.2 uses metadata found in transfers not identified as copyrighted to delve deeper into both these cases.

Information regarding which titles are in Audible Magic's database is not public, but the database is updated regularly with newly released music, movies and TV shows. It is fair to expect more recent or higher sales titles to be better represented since those compose the industry's high-revenue fringe and are more likely targets of piracy. Hence, P2P users that only

transfer copyrighted material sufficiently unpopular not to be present in the database will not show DATCoMs. Furthermore, certain types of material are harder to identify than others (movies are harder to identify than music), which makes users more difficult to detect if they transfer more of the harder types.

Apart from the above limitations that result from Audible Magic's database approach, users can take several active measures to prevent detection, which will work for all DPI-based approaches. One such measure is to transfer compressed[30] material, which is typically not done to avoid monitoring, but because that is how certain material is made available. Users that transfer compressed material are detected as P2P users but will show a reduced number of DATCoMs as none of the compressed material is detected as being copyrighted. Another measure is to use encryption, achievable by simply checking this feature in P2P clients. Most P2P clients support encryption only of control traffic or of all traffic. Both cases pose serious challenges at the level of detection of P2P traffic, and the latter makes it impossible to identify transferred material. Audible Magic, as most DPI-based monitoring appliances, cannot detect encrypted P2P traffic as being P2P, thus users who encrypt P2P will not even show up as P2P users in our data.

5.2 Pervasiveness and Intensity of P2P Usage

5.2.1 P2P Usage across Demographics

ISU is a public college. Initially established as a teacher's college, it still offers many education-related majors. In 2006-2007 it had 20,261 students, of which 88% were undergraduate, 34% lived on campus (ResNet) and 74% received financial aid [25]. Data analysis focused on ResNet students, which accounted for over 96% of all detected P2P activity. ResNet students were in 79 different majors listed in the table in Appendix A. Based solely on their major, students were grouped in two categorical variables: area of major[31] captures different scientific

[30] Material transferred inside compressed archives is fairly common in P2P networks. In order for a DPI-based monitoring appliance to be able to identify material within compressed archives, it would need to collect all the packets for the archive in order to be able to decompress it. For this to be feasible, such appliances would have to store all the material transferred in all streams that they detect, which is impractical given the typical bandwidths that the appliances need to cover. However, the reason why users transfer compressed material is not to avoid being monitored. Music albums are often made available as zip or rar archives containing all the individual songs together with album covers as image files. It is not possible to decompress one such archive having only a small fraction of it, which, in practical terms, makes the material within these files impossible to inspect by Audible Magic.

[31] Grouping according to the table in Appendix A

areas, and IT savviness[32] captures the propensity of students to be more IT savvy, which may lead to different online behavior. Figure 4 shows the demographic breakdown of ResNet students.

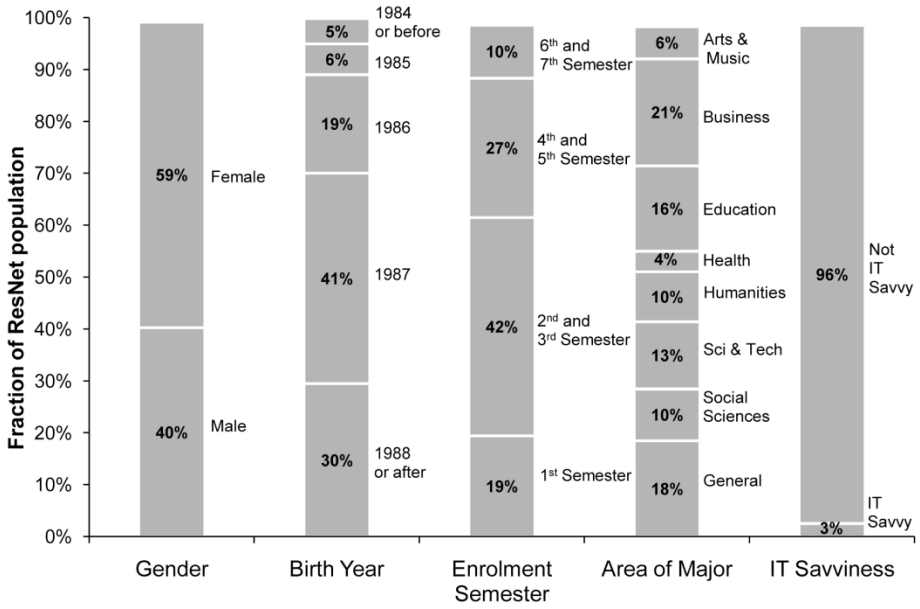


Figure 4. Breakdown of ResNet students by gender, birth year, enrollment semester, area of major and IT savvyness.

P2P and transfers of copyrighted material were prevalent across all analyzed demographics. According to the breakdown of the proportions of P2P and DATCoM users in ResNet and of the average number of titles transferred or attempted per week in Table 2, none of the demographic groups stands out particularly. Nevertheless, statistically significant differences exist between males and females, the former with higher proportions of P2P and DATCoM users but the latter transferring or attempting to transfer more titles; between students with IT savvy majors and students with other majors, the former with a higher proportion of P2P users but the latter transferring or attempting to transfer more titles; between General Student and other majors, the first with higher proportion of P2P and DATCoM users and also transferring or attempting to transfer more media titles; and across birth years and enrollment semesters.

[32] Majors considered as IT savvy are tagged in the table in Appendix A

Table 2. Lower bound on proportions of P2P users and DATCoM users out of all ResNet users and average distinct copyrighted media titles detected per week per ResNet user, broken down by demographics (95% CI in parenthesis).

		Proportion of P2P users	Proportion of DATCoM users	Average copyrighted titles detected per user per week
	All population	51% (49% - 52%)	42% (40% - 43%)	6 (5.4 - 6.0)
Gender	Female	45% (43% - 46%)	40% (39% - 42%)	6 (5.6 - 6.3)
	Male	59% (57% - 61%)	44% (42% - 46%)	5 (4.7 - 5.5)
Enrollment Semester	1st Semester	56% (53% - 58%)	48% (45% - 50%)	8 (7.5 - 8.9)
	2nd and 3rd Semester	51% (49% - 53%)	42% (40% - 44%)	6 (5.3 - 6.1)
	4th and 5th Semester	50% (47% - 52%)	41% (39% - 43%)	5 (4.2 - 5.0)
	6th and 7th Semester	43% (39% - 47%)	31% (27% - 34%)	3 (2.5 - 3.8)
Birth Year	1988 or after	51% (49% - 53%)	42% (39% - 44%)	7 (5.5 - 7.7)
	1987	53% (51% - 55%)	44% (42% - 46%)	6 (5.3 - 6.6)
	1986	51% (48% - 54%)	41% (39% - 44%)	5 (4.2 - 6.1)
	1985	44% (39% - 49%)	31% (26% - 35%)	3 (2.0 - 4.4)
	1984 or before	45% (40% - 50%)	30% (25% - 35%)	3 (1.6 - 3.8)
Area of study	General Student	59% (56% - 62%)	52% (49% - 55%)	9 (8.2 - 9.8)
	Social Sci	50% (46% - 54%)	42% (38% - 46%)	5 (4.3 - 5.6)
	Sci &Tech	55% (51% - 58%)	40% (36% - 43%)	4 (3.7 - 4.9)
	Humanities	47% (43% - 51%)	39% (35% - 42%)	4 (3.4 - 4.8)
	Health	45% (39% - 51%)	40% (34% - 46%)	5 (3.5 - 5.5)
	Education	42% (39% - 45%)	37% (34% - 40%)	5 (4.5 - 5.7)
	Business	52% (50% - 55%)	42% (40% - 45%)	6 (5.0 - 6.1)
	Arts & Music	45% (40% - 50%)	34% (29% - 39%)	4 (2.8 - 5.2)
IT Savvyness	Not IT Savvy	50% (49% - 52%)	42% (41% - 43%)	6 (5.4 - 6.0)
	IT Savvy	62% (54% - 69%)	43% (36% - 51%)	3 (1.8 - 3.2)

There is high correlation between birth year and enrollment semester in the ResNet population[33], therefore similar results are expected. Figure 5 through Figure 8 revisit numbers in Table 2 for these demographics in a graphical fashion. In both cases, the average number of titles per week decreases for older students or students closer to graduation. For birth year, older students (born in 1986 or before) have smaller proportions of P2P and DATCoM users than younger students (born in 1987 or after), but the differences between birth years within each group are not significant. Concerning enrolment semester, the closer students are to

[33] Correlation coefficient of 0.63

graduating the smaller proportions of P2P and DATCoM users. This fact can have two reasonable but conflicting explanations. On one side, as users spend more time in college, they may learn to better conceal their online activity, thus resulting in a smaller proportion of detected activity for similar proportions of actual P2P users. On the other side, students may actually use P2P less the higher the semester they are at, either because they start college with more entrenched P2P habits every year or because they actually decrease their P2P activity as they advance towards graduation (due to satiation, less free time, conscience of illegality, etc). Future research on data from consecutive years will allow testing the above hypotheses.

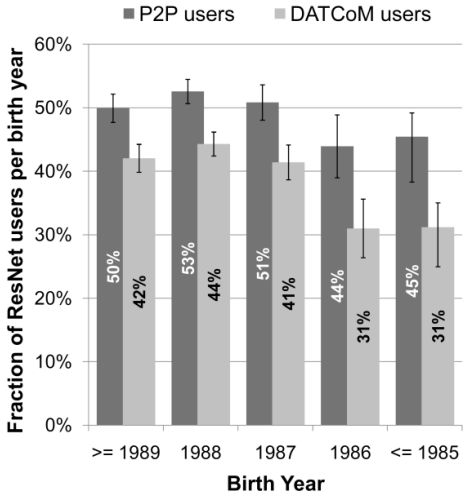


Figure 5. Lower bound on proportion of ResNet users that were detected as P2P users or DATCoM users, broken down by birth year (error bars represent 95% CI).

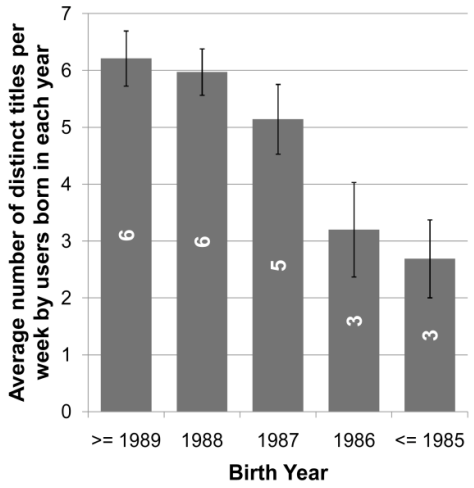


Figure 6. Lower bound average weekly number of distinct titles transferred or attempted to transfer per ResNet student, broken down by birth year (error bars represent 95% CI).

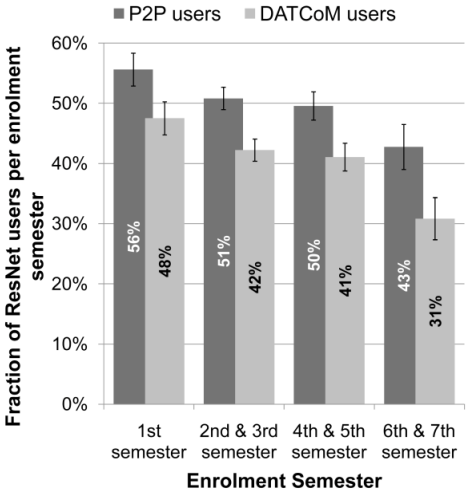


Figure 7. Lower bound on proportion of ResNet users that were detected as P2P users or DATCoM users, broken down by enrolment semester (error bars represent 95% CI).

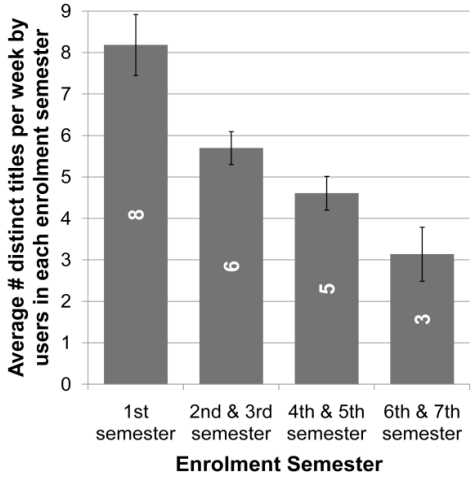


Figure 8. Lower bound average weekly number of distinct titles transferred or attempted to transfer per ResNet student, broken down by enrolment semester (error bars represent 95% CI).

Besides finding that P2P and DATCoMs are prevalent across demographics, we also found that, using demographics only, it is not possible to identify subgroups of the population that are substantially different in terms of P2P usage so as to allow targeting of interventions (such as education campaigns) designed to try to reduce P2P or DATCoMs on campus. This happens because demographics have very low predictive power for either the probability of being a P2P user, the probability of P2P users performing DATCoMs[34], or the number of distinct media titles per DATCoM user. This is in evidence in Table 3, which summarizes and presents goodness of fit results for three models estimated for the above outcomes of interest using as predictors the demographic variables under analysis[35]. Poor goodness of fit for all models shows that none of the outcomes is successfully predicted based solely on the demographics used.

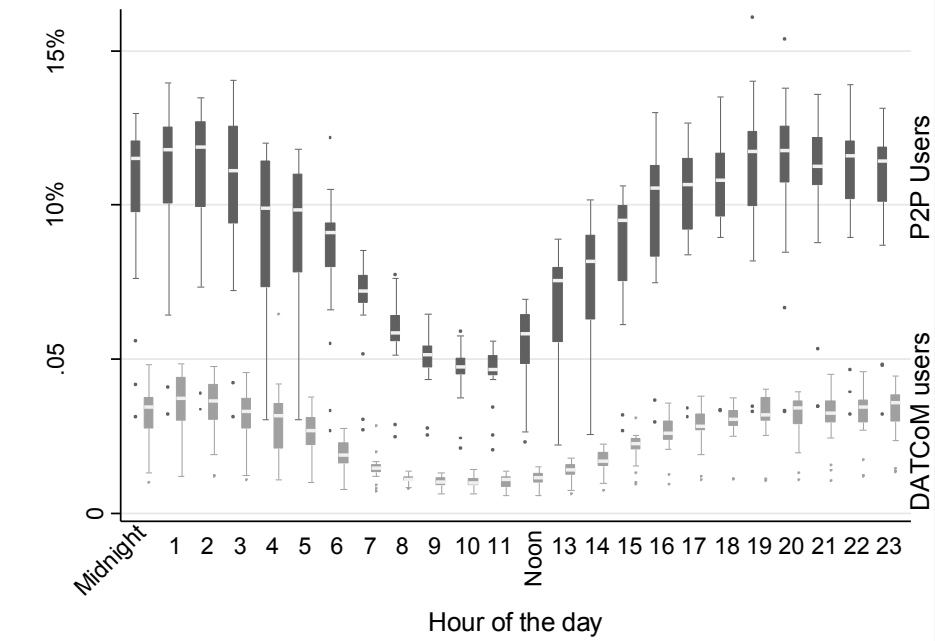
Table 3. Description of regression models (dependent variable, possible values for the dependent variable, type of regression model and goodness of fit metric) used to assess the predictive power of demographics.

Model	Dependent variable	Value	Regression	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. Undefined otherwise.	Logit	0.055
C	Number of distinct media titles per DATCoM user	Log of the number of media titles per DATCoM user. Undef. for non-DATCoM users.	OLS	0.041

5.2.2 P2P Usage over Time

One of the main arguments of ISPs and network managers against P2P is that it is the main contributor for congestion in their networks, making it hard to maintain adequate quality of service for other online applications. We found that both P2P and DATCoMs occur at all hours of the day and night[36] (Figure 9), with an average of 9% of the ResNet population detected engaging in P2P in each hour and an average of 2% detected performing DATCoMs. This is another aspect of how pervasive P2P and DATCoMs are on campus. However, as the table in Figure 9 shows, there is a statistically significant difference between the proportion of ResNet users that engage in P2P and DATCoMs during the day and during the night, with both activities peaking in night hours. This is actually good news for network managers, since night hours are typically less busy hours for other types of traffic.

[34] Models for probability of performing DATCoMs for all users were attempted, but resulted in poorer fits of the data.
[35] 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.
[36] A user is considered to be doing P2P (or DATCoMs) in any given hour if there is at least one P2P activity (or DATCoM) detected for that user during that hour.

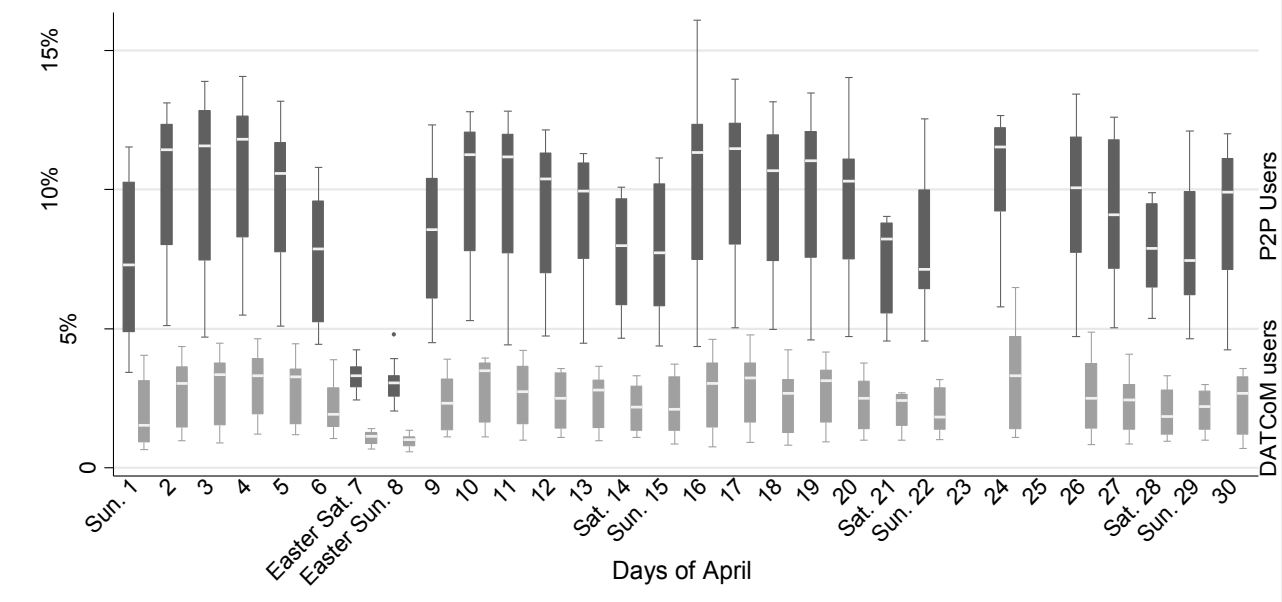


	All monitored hours	Night hours (5pm to 7am)	Day hours (7am to 5pm)
P2P users	9% (8.4% - 8.9%)	10% (10.0% - 10.6%)	7% (6.6% - 7.2%)
DATCoM users	2% (2.3% - 2.4%)	3% (3.0% - 3.2%)	2% (1.5% - 1.7%)

Figure 9. Boxplot (quartiles box divided by median line, 1.5 inter-quartile range upper and lower whiskers, and point outliers) for each hour of the day of the distribution of proportions of detected P2P and DATCoM users in ResNet. Table wih average hourly percentage of ResNet users detected as P2P users or DATCoM users (95% CI in parenthesis). Differences between night and day proportions are statistically significant at 5%.

P2P and DATCoMs were also found to occur in all monitored days (Figure 10[37]), with statistically significantly higher proportions of detected P2P and DATCoM users for weekday hours than for weekend hours, as we can observe in the table in Figure 10. A plausible reason why students don't do as much P2P during the weekends (and especially Easter) is that they might go back to their parents' homes, which makes sense for ISU since 94% of enrolled students (96% of undergraduate students) are from Illinois [25]. Hence, P2P and DATCoMs occur at all times but are mostly passive activities that peak on weeknights when humans are typically not active. This is not good news for the industry, since their copyrighted material is being transferred at all times, even when nobody is actively commanding such transfers.

[37] There was no data from April 25th. April 23rd was not plotted because there was only one hour of data



	All monitored hours	Weekday hours	Weekend hours	Weekend hours (excl. Easter)
P2P Users	9% (8.4% - 8.9%)	10% (9.3% - 9.9%)	7% (6.2% - 7.0%)	8% (7.4% - 8.1%)
DATCoM Users	2% (2.3% - 2.4%)	3% (2.5% - 2.7%)	2% (1.7% - 2.0%)	2% (2.0% - 2.2%)

Figure 10. Boxplot (quartiles box divided by median line, 1.5 inter-quartile range upper and lower whiskers, and point outliers) for each day in the monitored period of the distribution (over the hours of that day) of the proportion of detected P2P and DATCoM users.
Table with average hourly percentage of ResNet users detected doing P2P or DATCoMs in weekday hours and weekend hours (95% CI in parenthesis). Difference between weekday hours and weekend hours (including or excluding the Easter weekend) are statistically significant at 5%.

Finally, we find that P2P is part of the routine of the users who perform it since, as Figure 11 shows, each P2P user on campus is detected engaging in the activity for a median of about 10% of the monitored hours and in a median of about 35% of monitored days, which corresponds to about 6 hours per day in 10 days in the period. The figure also shows that the amount of time spent in P2P is much greater for users with detected DATCoMs than for those without, which gives rise to the hypothesis that some P2P users might not have been detected performing DATCoMs because they didn't engage in P2P for long enough to be detected by Audible Magic.

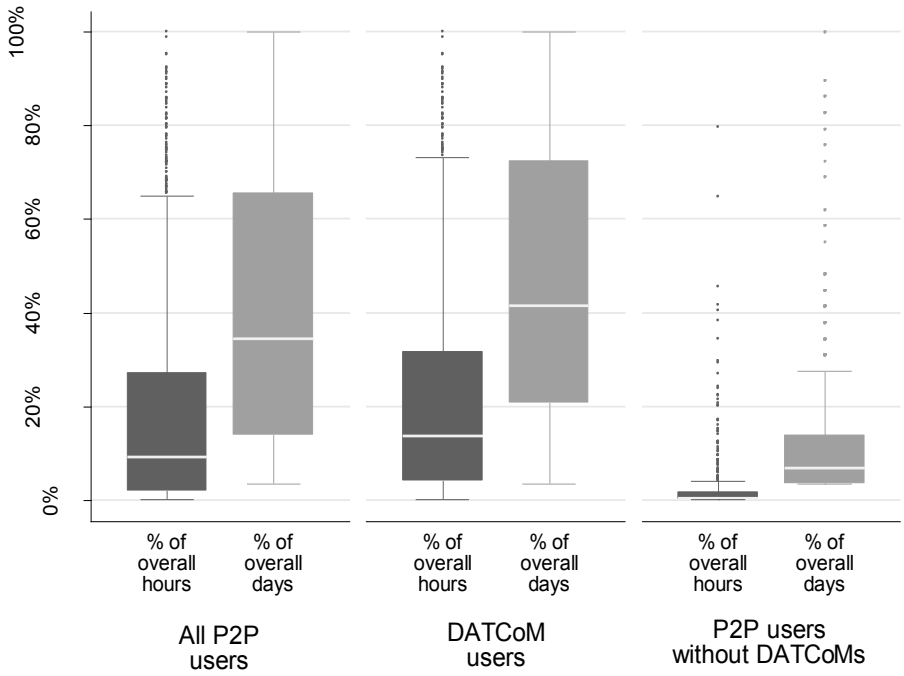


Figure 11. Boxplot (quartiles box divided by median line, 1.5 inter-quartile range upper and lower whiskers, and point outliers) of the distribution of the percentage of overall hours and overall days with detected P2P activity; for all P2P users, DATCoM users and P2P users without DATCoMs.

5.2.3 Intensity of P2P Usage

Another important measure of P2P usage is the mean number of copyrighted media titles detected being transferred per student in ResNet. Figure 12 shows this number when averaged over a day, a week, and the 25 day monitoring period. There is almost one copyrighted title per student per day. In fact, as portrayed in Table 4, the great majority (70%) of users with detected DATCoMs were detected transferring more than one copyrighted title per day. This shows that DATCoMs are not performed heavily by a small group of users, but instead that most users are intensive users of P2P to transfer copyrighted material.

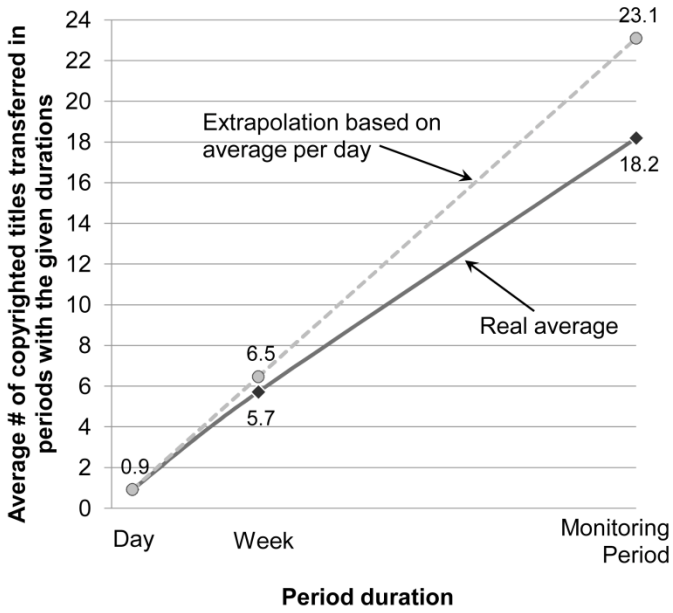


Figure 12. Average number of distinct copyrighted titles detected per ResNet user per day (using the 25 days with 24 hours worth of data), per week (using the 3 weeks with 7 full days worth of data) and during the entire monitoring period (using the whole 25 days). The real average over the periods is compared with the extrapolation of the real daily average (0.9) for a week (multiplied by 7) and for the monitoring period (multiplied by 25).

Moreover, returning to Figure 12, the fact that the mean number of DATCOMs observed in *n* days does not increase linearly with *n* shows that we detect attempts to transfer the same title to or from the same user on multiple days. Thus, a title might be counted separately in two different days, but it is counted only in one week. This is sometimes because the same transfer spanned two days, perhaps because it started just before midnight. It may also occur because additional transfers have occurred on different days within the same week. For example, a title might be downloaded on Monday, and then uploaded to other users on Tuesday and Thursday.

Table 4. Proportion of ResNet users and of users with detected DATCoMs detected transferring more than one copyrighted title in the period, per week, or per day (95% CI in parenthesis).

	% of ResNet users	% of DATCoM users
Detected DATCoM users	42% (40% - 43%)	
With more than one copyrighted media title **in the period**	39% (38% - 40%)	94% (93% - 95%)
With more than one copyrighted media title **per week**	33% (32% - 34%)	79% (77% - 80%)
With more than one copyrighted media title **per day**	29% (28% - 30%)	70% (68% - 72%)

Following from above, we assess how long each user keeps a copyrighted title in its P2P client share list, therefore making it available for upload to other peers. Since campus network users typically enjoy higher upload speeds than regular ISP users, they become preferred sources because they can typically provide faster uploads to other peers. On average, each copyrighted title was detected being shared by each user for at least 59 hours (57.9 – 59.5 hours), which is

about 2.5 days. This is the average of the minimum duration that each user maintains each copyrighted title in her P2P share list, which is calculated as the time difference between the first and last DATCoM detected for each user×title pair[38]. Figure 13 presents the distribution of this minimum duration for the 52% of user×title pairs detected in more than one DATCoM (those detected in a single DATCoM would have a duration of 0) and shows that 22% of all the detected material shared online is shared for at least one day and 14% for least one week. These figures do not account for the fact that users may not have their P2P clients operating during the whole share duration of each title. However, in face of findings in the previous section that users perform P2P on a regular basis, it is fair to assume that, for instance, if a user is found sharing a title for 10 days, then that title should be exposed in the P2P network a couple of hours per night in about 4 of the 10 days.

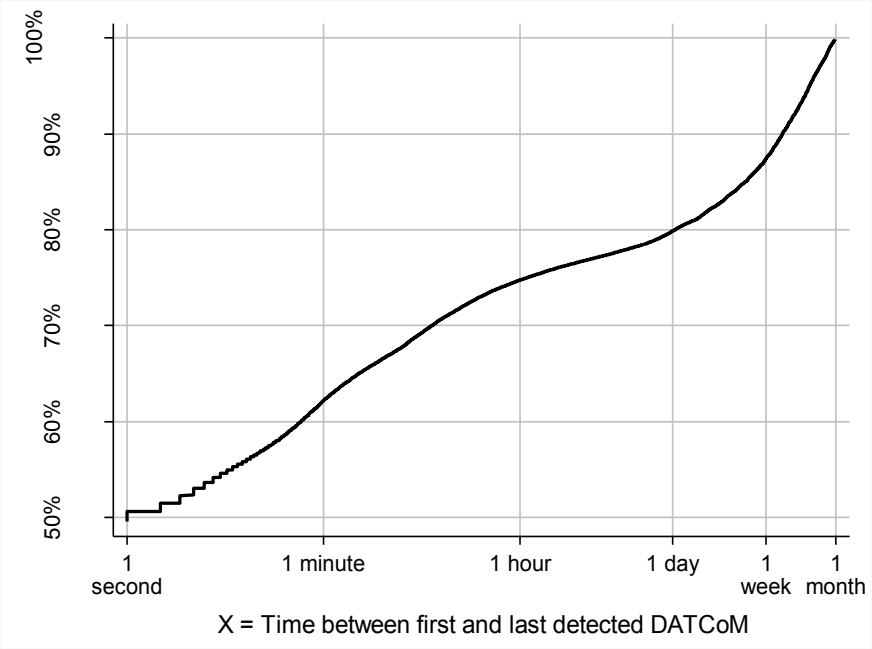


Figure 13. Empirical Cumulative Distribution Function of the minimum duration that each copyrighted title is shared by each user (user×title pairs with a single DATCoM omitted).

[38] This is relevant because in P2P systems a user starts sharing a file (eventually uploading it to other users) as soon he/she has downloaded at least one piece of that file.

5.3 Characterization of Transferred Material

5.3.1 Copyrighted Material

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 likely not available in the store around the corner will allow industry decision-makers to work on alternative ways of reaching P2P users. For instance, by devising new marketing strategies to make media sales competitive with P2P, or by expanding their catalogs to make it easy for P2P users to obtain the material they seek from legal sources. We find that the distribution of popularity of transferred copyrighted titles (Figure 14) appears to be long-tailed, with a small head of very popular titles and a large tail of unpopular titles. The popularity of each title is defined as the "market" share of that title, i.e., the percentage of DATCoMs involving that title out of the total number of DATCoMs detected in the monitored period. This means that P2P is used for both popular and less-popular material, the latter accounting for a fair share of the "market".

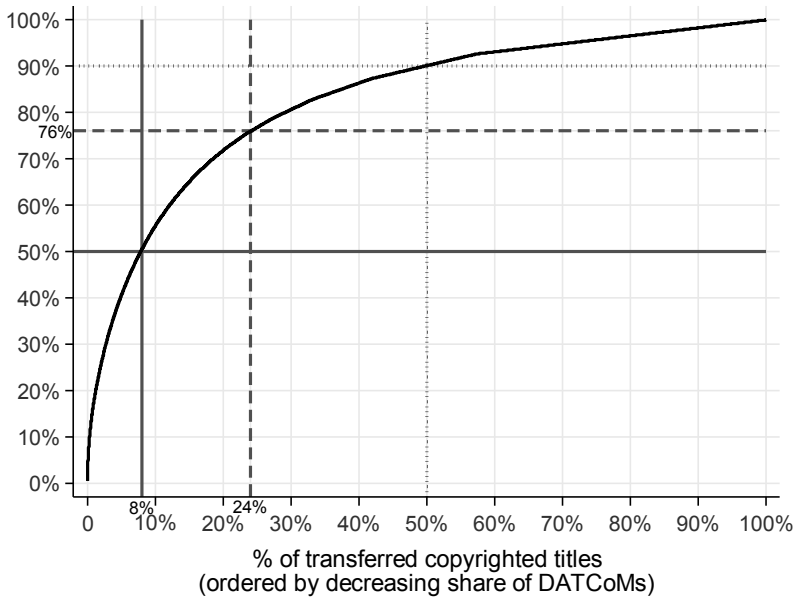


Figure 14. Cumulative Distribution Function of the popularity of copyrighted titles transferred over P2P, i.e., of the percentage of DATCoMs accounted for by the X% top titles. The 8% most popular titles (< 2,000 titles) accounted for 50% of all DATCoMs while the 50% least popular titles (≈12,000 titles) accounted for 10% of the DATCoMs. The balance ratio is 76/24: 76% of DATCoMs are due to the 24% top titles and the 76% bottom titles account for 24% of DATCoMs.

We also compare the above distribution to popularity distributions of Netflix Top 100 movies concerning number of user ratings[39] [32], in Figure 15, and of CD and DVD sales in Amazon.com[40] [47], in Figure 16. Comparison to Netflix focuses on the head of the distribution and shows that DATCoMs are not as dominated by a small number of popular titles. In the case of Amazon.com CDs and DVDs, the focus is on the tail of the distribution, where we see a similar shape but with higher mass in the case of P2P copyrighted titles. Hence, less popular material seems to play a similar role on campus P2P as it does in Amazon.com, reaching market niches that alone do not account for a great market share but that, altogether, add up to a significant percentage of sales[41].

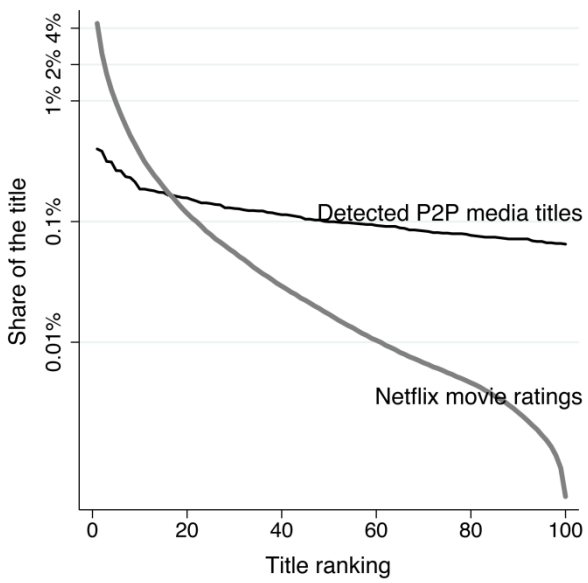


Figure 15: Comparison of the distribution of popularity of the Top 100 most popular copyrighted titles transferred over P2P to the distribution of rating share of the Top 100 Netflix movies with most user ratings.

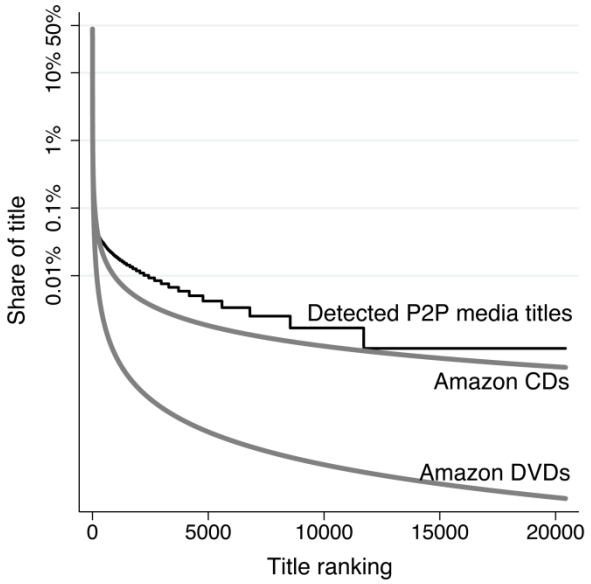


Figure 16: Comparison of the distribution of popularity of copyrighted titles transferred over P2P to the distribution of Amazon CD and DVD sales.

[39] The distribution of Netflix Top 100 movies in terms of user ratings does not reflect actual movie rentals, but it is a fair approximation since movies that are most often rated are also most often rented.
[40] The distribution of CD and DVD market shares was derived using parameters estimated in [47] for a Pareto distribution of sales. To go from the distribution of sales to the distribution of market share, it was assumed that the number of distinct CD and DVD titles sold was the same as the number of distinct titles transferred. Varying this number within reasonable boundaries does not change the shape of the distributions.
[41] Since for material to be identified as copyrighted it must be present in the Audible Magic database, it is possible that extremely unpopular titles are not detected at all and that the distribution is biased towards popular titles. Nevertheless, it is clear that P2P is used to transfer a fair amount of unpopular material.

5.3.2 Material Not Detected as Copyrighted

Many P2P supporters discuss the legal uses of P2P. We observed that 9% of students engaged in P2P but generated no DATCoMs, and males were more likely to fall into this category than females. It is unknown how many of these students transferred copyrighted material but were not detected, and how many were involved only in legal use of P2P. To assess this, this section addresses material that would not generate a DATCoM if transferred using P2P.

One important legal use of P2P is the distribution of open source software, such as Linux. P2P is also used to exchange adult materials, some but not all of which are not copyrighted. Even those adult materials that are copyrighted were not detected as DATCoMs, because the database of copyrighted material did not include adult titles. However, as this section will show, we found no evidence that either of these types can fully account for the 9% of students observed using P2P but not observed with DATCoMs.

We characterized material that was not identified as copyrighted through use of metadata. During the monitored period, 28% of the communication events detected by Audible Magic were found to contain this metadata, which often corresponds to the name of the file that is being transferred, or at least the file that the user had intended to transfer.

Linux is among the more prominent software titles legally distributed using P2P networks. We observed 0.2% of P2P users transferring files with names that include “linux” or specific major Linux distributions on the market: “fedora”, “ubuntu”, “suse”, “red hat”, “mandriva”, “slackware” or “debian”. The percentage was the same among those P2P users with DATCoMs. Among P2P users without DATCoMs the average proportion was 0.02% and it was not statistically different from 0. Hence, there is no evidence to support the hypothesis that the transfer of Linux distributions is a driver for the use of P2P, even among users that do not use P2P for copyrighted material.

Transfers were generally assumed to contain adult material if the metadata included the keywords “porn” or “xxx” and file extensions indicated video , with a few exceptions that were explicitly filtered out[42]. There is certainly a great deal of adult material that does not include these keywords, but it should be possible to make fair comparisons across subsets of P2P

[42] An example of one such title is the 2002 movie “xXx” (http://www.imdb.com/title/tt0295701).

users with these keywords, as the percentage of adult material that contains these keywords is likely to be the same for all student groups. Table 5 and Table 6 show the overall percentage of students, and of various subsets of students, that engaged in transfers that included these keywords.

A significant percentage (53%) of P2P users was observed transferring adult material. However, the dissemination of adult material does not explain the existence of P2P users with no observed DATCoMs. P2P users with no detected DATCoMs are far less likely (12% vs. 62%) to be seen transferring adult material. Moreover, use of adult material is a good predictor of use of copyrighted media and P2P users observed transferring adult material had roughly three times as many DATCoMs as P2P users who were not observed transferring adult material.

Table 5: Percentage of P2P users, of users with DATCoMs and of users without DATCoMs that had communication events whose metadata indicates adult material, broken down by gender (95% CI in parenthesis).

	Proportion of users with detected adult material		
	Both genders	**Out of males**	**Out of females**
Out of P2P users	53% (51.6% - 55.0%)	57% (54.4% - 59.4%)	50% (47.8% - 52.5%)
Out of DATCoM users	62% (60.3% - 64.0%)	71% (68.5% - 73.7%)	55% (52.8% - 57.8%)
Out of P2P users without DATCoMs	12% (9.8% - 15.1%)	15% (11.4% - 18.4%)	7% (3.0% - 10.2%)

Table 6. Proportion of P2P users with detected DATCoMs and average number of copyrighted titles detected in the monitoring period, broken down by users detected transferring adult material and users not detected (95% CI in parenthesis).

	Proportion of P2P users with DATCoMs	**Average distinct copyrighted titles per P2P user**
Out of users detected transferring adult material	96% (94.9% - 96.8%)	56 (51.9 - 60.4)
Out of users not detected transferring adult material	65% (64.4% - 69.1%)	13 (11.6 - 14.1)

Table 5 also shows that males were somewhat more likely (57% vs. 50%) to be seen transferring adult material, but these differences do not appear to be large enough to explain why 15% of male students were observed using P2P with no DATCoMs as compared to only 5% of female students, as shown in Table 2.

6 Conclusions and Policy Implications

This article presents findings from the first large-scale quantitative assessment based on observation of P2P exchanges on a college campus. Because a significant part of the online activity at ISU was not captured, most results are lower bounds.

During April 2007 at least 51% of students living on campus engaged in P2P and at least 42% attempted to transfer copyrighted material. These figures are consistent with the initially controversial claims by the RIAA that over half of college students engaged in illegal file-sharing, which were obtained through survey methods. Our monitoring system detected an average of 6 successful or unsuccessful attempts to transfer (in either direction) copyrighted media titles per student per week from the entire dorm population. Out of the students found transferring or attempting to transfer copyrighted material, 70% were detected attempting more than one copyrighted title per day on average.

The economic impact of this file sharing depends in part on the extent to which students make files available to people off campus, in part because campuses often provide upload speeds far greater than those of residential users. Users were detected with transfers (uploads or downloads) for each copyrighted title spanning a mean of at least 2 days, a time during which those titles are available to other peers, since data is available for upload the instant it has been downloaded.

Students of all genders, ages, classes and majors participated in this file-sharing. Demographics were poor predictors of whether or how much students engage in P2P or in transfers of copyrighted material. This implies that demographics are ineffective for targeting interventions, such as education campaigns, that aim at altering students' behavior.

There is both good and bad news for network managers. File sharing occurred at all times of the day and night, peaking on weeknights when people are typically not using their computers. Hence, peak P2P usage periods occur when congestion is less of a problem. Nevertheless, even in an environment such as the ISU network where detected P2P is policed, P2P activity accounted for about 12% of network traffic, a lower bound that excludes encrypted traffic.

Concerning transferred copyrighted material, its distribution of popularity (i.e., of the share of each title out of total transfers of copyrighted material) resembles a long-tailed distribution comparable that of CD sales from Amazon.com. This indicates that unpopular media reaches market niches, with each title alone not accounting for many transfers but the tail of the distribution adding up to a significant share of all transfers. Hence, in order to "compete" with P2P, content retailers not only need to account for the fact that users can obtain content free of charge, but that they also have access to an extensive catalog of titles to suit their particular taste. While the costly versus "free" factor affects both online and physical retailers, the

extension of catalog factor is more likely to be a problem for physical retailers who have to deal with storage and inventory constraints.

Some might suggest that there are many people who use P2P for the legal transfer of software such as Linux, or for the transfer of adult material (which may or may not be copyrighted), but do not engage in the illegal transfer of copyrighted material. However, we found no evidence of this among college students. Indeed, P2P users who transfer adult material are more likely to attempt to transfer copyrighted material and vice-versa. P2P users that transfer adult material are also found to attempt to transfer, on average, more than three times the number of copyrighted media titles than those who don't transfer adult material. As for the legal transfer of software, the percentage of P2P users found transferring Linux out of those that do not transfer copyrighted media is not statistically different from zero.

In addition to the data gathered, this study provides useful lessons on use of DPI technology to reduce illegal activities. Our results indicate that today's technology is effective in detecting users that transfer copyrighted material using unencrypted P2P. Of students who were detected using P2P, 82% were found attempting to transfer copyrighted material at some point in the one month monitoring period, although this percentage decreases significantly for smaller monitoring periods. Thus, while the technology may not detect every copyrighted file that is transferred, it is likely to catch those who transfer copyrighted files after a few weeks of observation – at least of those whose P2P traffic is detected at all. Thus, DPI can be a useful diagnostic tool for universities or ISPs that, for example, wish to warn their students or customers about possible risk of lawsuits from the copyright industry or, in the case of universities, target education to users that are detected in copyright infringement.

However, this approach may be less effective if it is coupled with a punishment mechanism. The biggest problem with using this approach to assign penalties rather than merely to inform network operators, users, or others about DATCoMs is that the technology probably misses all material sent using encrypted P2P, as well as some material sent using unencrypted P2P. Hence, this form of detection can be rendered ineffective through use of encryption, which is available in all major P2P systems and that users can activate relatively easily if they wish. The assignment of punishments could motivate users to take this step, or even motivate P2P developers to activate encryption by default via a software update. While there are passive monitoring techniques beyond the ones described in this paper that can detect encrypted P2P traffic [10, 17], they cannot tell whether the transferred material is copyrighted.

It is not just network operators who are confronted with decisions concerning DPI technology. Lawmakers could decide to require that operators use DPI to monitor their networks and enforce anti-piracy policies by assigning penalties. Such policies have been considered for college campuses in the US [11, 24, 46, 48], and for commercial ISPs in France [6] and other EU countries [1, 30]. This would mandate an investment in monitoring technologies that might initially lead to the detection of more copyright violations, but because P2P users have the ability to evade this particular technology with tools they already possess, the long-term impact is uncertain. In addition, policymakers need to take into account competition issues if the network operator that is blocking the transfers of copyrighted music and video is also a provider of music or video, as would be the case for many cable TV providers [38].

Since encryption conceals whether material is copyrighted from DPI-based enforcement mechanisms, some policymakers may consider prohibiting use of encryption with P2P. One disadvantage is that some P2P transfers are legal and beneficial, and such a policy would prohibit security practices that protect these legal transfers as well as their illegal counterparts. The long-term outcome of such a measure is difficult to predict, as users engaged in illegal transfers may adopt other obfuscation techniques in addition to encryption, and an arms race could easily ensue.

If large numbers of users do choose to encrypt or otherwise conceal P2P traffic to avoid punishment, this may have unintended side effects. First, it will no longer be possible to use DPI to detect copyright violations for informational purposes. Thus, for example, even if it proves useful for a network operator to warn users that they are at risk of future lawsuits from copyright-holders while the network operator imposes no punishment, that approach could be undermined. Second, it may be more difficult for network operators to identify P2P traffic exclusively for the purpose of ensuring quality of service for non-P2P traffic during periods of congestion.

Note that there are other technical means of detecting network users who share copyrighted material besides passive network monitoring and DPI. One alternative approach is for a node to infiltrate and actively participate in a P2P network, which copyright-holders and their representatives use today to identify the peers that are sharing or requesting copyrighted material. However, currently, network service providers have no particular advantage in infiltrating a P2P network using this technology. Their role is limited to providing the copyright holders contact information for subsequent legal actions against users identified from IP

addresses, or shutting off service for those users who are allegedly violating copyright law, or both. It is conceivable that network providers might be able to play a larger role in the future with technical enforcement that is based on infiltration, perhaps under the new P4P platform [49]; this is a subject for future work.

Acknowledgements

This work was partly funded by the Portuguese Science and Technology Foundation, fellowship reference: SFRH/BD/27350/2006.

References

1. Anderson, N. (2008) *UK ISPs don't want to play umpire to "three strikes" rule*. February 15, online at: http://arstechnica.com/news.ars/post/20080215-uk-isps-dont-want-to-play-umpire-to-three-strikes-rule.html.
2. Anderson, N. (2008) *UK ISPs don't want to play umpire to "three strikes" rule*. February 15 [last checked March 2008] online at: http://arstechnica.com/news.ars/post/20080215-uk-isps-dont-want-to-play-umpire-to-three-strikes-rule.html?rel.
3. Bangeman, E. (2007) *New bill would punish colleges, students who don't become copyright cops*. November 11 [last checked December 2007] online at: http://arstechnica.com/news.ars/post/20071111-new-bill-would-turn-colleges-into-copyright-cops.html.
4. Bangeman, E. (2008) *Apple passes Wal-Mart, now #1 music retailer in US*. April 2 [last checked May 2008] online at: http://arstechnica.com/news.ars/post/20080402-apple-passes-wal-mart-now-1-music-retailer-in-us.html.
5. Bangeman, E. (2008) *Judge kills RIAA subpoena: making available not infringement*. April 3, online at: http://arstechnica.com/news.ars/post/20080403-judge-kills-riaa-subpoena-making-available-not-infringement.html.
6. Bangeman, E. (2008) *France's plan to turn ISPs into copyright cops on track*. January 28, online at: http://arstechnica.com/news.ars/post/20080128-frances-plan-to-turn-isps-into-copyright-cops-on-track.html.
7. Biddle, P., et al. (2002) *The darknet and the future of content distribution*, in *Proceedings of the ACM Workshop on Digital Rights Management,* online at: http://www.dklevine.com/archive/%20darknet.pdf.
8. Bremner, C. (2008) *France to ban illegal downloaders from using the Internet under three-strikes rule*. June 19 [last checked June 2008] online at: http://technology.timesonline.co.uk/tol/news/tech_and_web/article4165519.ece.
9. Buskirk, E.V. (2007) *A Poison Pen From the RIAA*. February 28 [last checked November 2007] online at: http://www.wired.com/politics/onlinerights/news/2007/02/72834.
10. Collins, M.P. and M.K. Reiter. *Finding Peer-to-Peer File-Sharing Using Coarse Network Behaviors*. in *ESORICS*. 2006.
11. *Committee Looks at Technology to Limit Illegal Filesharing*, 2007, U.S. House of Representatives, Commitee on Science and Technology, online at: http://science.house.gov/press/PRArticle.aspx?NewsID=1858.
12. *Copyright Law of the United States and Related Laws Contained in Title 17 of the United States Code*. 2007, Library of Congress, Copyright Office: United States.
13. *Digital Citizen Project at Illinois State - Summary of Project*. (2008) [last checked February 2008] online at: http://www.digitalcitizen.ilstu.edu/summary/.
14. *The Digital Millennium Copyright Act of 1998*, 1998, United States Copyright Office
15. *Ellacoya Data Shows Web Traffic Overtakes Peer-to-Peer (P2P) as Largest Percentage of Bandwidth on the Network*, 2007, Ellacoya, online at: http://www.ellacoya.com/news/pdf/2007/NXTcommEllacoyaMediaAlert.pdf.
16. Fischer, K. (2007) *Bill would force "top 25 piracy schools" to adopt anti-P2P technology*. July 23 [last checked December 2007] online at: http://arstechnica.com/news.ars/post/20070723-bill-would-force-top-25-piracy-schools-to-adopt-anti-p2p-technology.html?rel.
17. Fisher, A. and M. Feyen, *Allot Communications NetEnforcer is First to Detect and Manage Encrypted BitTorrent Traffic*, 2006, Allot Communications, online at: http://www.allot.com/index.php?option=com_content&task=view&id=369&Itemid=18.

18. Geist, M. (2007) *ISP must come clean on `traffic shaping'*. April 16, online at: http://www.thestar.com/comment/columnists/article/203408.

19. *Global Napster Usage Plummets, But New File-Sharing Alternatives Gaining Ground*, 2001, Jupiter Media Metrix

20. Goldfarb, C.B., *Access to Broadband Networks*, 2006, CRS

21. Graham, L., *Legal Music Downloads Were Fastest Growing Digital Music Category in 2006*, 2007, The NPD Group, online at: http://npd.com/press/releases/press_0703141.html.

22. Guess, A. (2008) *Downloading by Students Overstated*. January 23 [last checked April 2008] online at: http://www.insidehighered.com/news/2008/01/23/mpaa.

23. Hussain, A. (2007) *The 20-minute broadband limit*. December 2, online at: http://business.timesonline.co.uk/tol/business/money/broadband/article2982965.ece.

24. *The Internet and the College Campus: How the Entertainment Industry and Higher Education are Working to Combat Illegal Piracy*. (2006) SN. 109-58, 109th Congress House Hearings. September, online at: http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_house_hearings&docid=f:30049.pdf.

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

26. Lamy, J., C. Duckworth, and L. Kennedy, *RIAA Launches New Initiatives Targeting Campus Music Theft}*, 2007, RIAA -- The Record Industry Association of America}

27. Lamy, J., C. Duckworth, and L. Kennedy, *RIAA Continues College Deterrence Campaign Into 2008}*, 2008, RIAA -- The Record Industry Association of America}

28. Liebowitz, S.J., *File Sharing: Creative Destruction or Just Plain Destruction?* Journal of Law \& Economics, 2006. **49**(1): p. 1-28.

29. Macavinta, C. (1999) *Recording industry sues music start-up, cites black market*. December 7 [last checked December 2007] online at: http://news.com.com/Recording+industry+sues+music+start-up,+cites+black+market/2100-1023_3-234092.html?tag=st.rn.

30. Meller, P. (2008) *Europe rejects plan to criminalize file-sharing*. April 10 [last checked May 25 online at: http://www.infoworld.com/archives/emailPrint.jsp?R=printThis&A=/article/08/04/10/Europe-rejects-plan-to-criminalize-file-sharing_1.html.

31. Mennecke, T. (2007) *P2P Remains Dominant Protocol*. June [last checked December 2007] online at: http://www.slyck.com/story1502.html.

32. *Netflix prize dataset*. 2006, Netflix.

33. Oster, S., *MPAA Statement on Motion Picture Industry Losses due to Piracy among College Students*, 2008, MPAA, The Motion Picture Association of America, online at: http://www.mpaa.org/press_releases/Iek%20college%20student%20data_f.pdf.

34. Oswald, E. (2006) *RIAA Sues LimeWire Over Piracy*. August 4 [last checked November 2007] online at: http://www.betanews.com/article/RIAA_Sues_LimeWire_Over_Piracy/1154722015.

35. *P2P Volume Climbs Again in June, User Levels Near 9 Million*. (2005) July 8 online at: http://www.digitalmusicnews.com/yesterday/july2005#070805p2p.

36. Packeteer, *Applications, Protocols, and Services Classified by PacketWise 7.3*. 2007.

37. Paul, R. (2007) *EFF study confirms Comcast's BitTorrent interference*. November 28, online at: http://arstechnica.com/news.ars/post/20071128-eff-study-reveals-evidence-of-comcasts-bittorrent-interference.html.

38. Peha, J.M. (2008) *The Future of Video: Challenges in Promoting Competition and Protecting Intellectual Property*. En Banc Hearing on Broadband and the Digital Future, Federal Communications

Commission. Pittsburgh, July 21, online at: http://www.fcc.gov/broadband_digital_future/072108/peha.pdf.

39. *Piracy Online*. [last checked December 2007] online at: http://www.riaa.com/physicalpiracy.php?content_selector=piracy_details_online.
40. Read, B. (2007) *The First Close Look at Colleges' Digital Pirates*, in *The Chronicle of Higher Education,* September, online at:
41. *RIAA v. The People: Four years later*, 2007, EFF, Electronic Frontier Foundation
42. Richtel, M. (2002) *Napster Says It Is Likely to Be Liquidated*, in *The New York Times,* September 4, online at:
43. Rob, R. and J. Waldfogel, *Piracy on the High C's: Music Downloading, Sales Displacement, and Social Welfare in a Sample of College Students*, 2004, National Bureau of Economic Research, online at: http://www.nber.org/papers/w10874.
44. Sherman, C. (2007) *An Update: Piracy on University Networks*. SN. 110-29, 110th Congress House Hearings. March, online at:
45. Smith, J. (2006) *ISU staff speaks in Washington on issues of campus piracy*, in *Daily Vidette at Illinois State University,* October 4, Normal, Illinois online at: http://media.www.dailyvidette.com/media/storage/paper420/news/2006/10/04/News/Isu-Staff.Speaks.In.Washington.On.Issues.Of.Campus.Piracy-2329590.shtml.
46. Smith, L., et al., *Letter sent to universities on May 1, 2007*, 2007, online at: http://w2.eff.org/IP/P2P/RIAA_v_ThePeople/example_letter_sent_to_universities.pdf.
47. Smith, M.D. and R. Telang, *Research Note: Internet Exchanges for Used Digital Goods: Empirical Analysis and Managerial Implications*. 2007.
48. *An Update: Piracy on University Networks*. (2007) SN. 110-29, 110th Congress House Hearings. March, online at:
49. Xie, H., et al., *P4P: Explicit Communications for Cooperative Control Between P2P and Network Providers*, 2007, P4P Working Group
50. Zentner, A., *File Sharing and International Sales of Copyrighted Music: An Empirical Analysis with a Panel of Countries.* Topics in Economic Analysis & Policy, 2005. **5**(1): p. 1452-1452.

Appendix A Grouping of Majors by Area of Study

Category	Major
General	General Student
Social Sciences	Communication
	Economics
	Anthropology
	Political Science
	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.