Characterizing digital media exchanges in a university campus network

Alexandre M. Mateus[1] and Jon M. Peha[2,3]

Carnegie Mellon University

Abstract

This article presents findings from a large-scale quantitative assessment of online exchanges of copyrighted material on a college campus based on network data collected using deep packet inspection (DPI). We find that use of Peer-to-Peer (P2P) for the transfer of copyrighted content is widespread on campus, although observed P2P is declining. In a month-long monitoring period in Spring 2008, at least 40% of students living on campus were observed engaging in P2P, 70% of those were detected attempting to transfer copyrighted content, and each of the latter was observed transferring copyrighted titles at an average rate of 4 titles per day. Nevertheless, from Spring 2007 to Spring 2008, the daily percentage of detected P2P users fell 10%, and the daily percentage of users observed attempting to transfer copyrighted content out of those detected doing P2P fell 20%. These changes could be the result of decreasing use of P2P, or increasing use of encrypted P2P to evade detection.

We also find that, given a couple weeks or more, current DPI technology identifies most users attempting to transfer copyrighted material, out of users whose P2P traffic it can detect. This shows that even if DPI does not detect every transfer of copyrighted material, it can effectively identify individuals who make these transfers, provided they do not use encryption. However, detection of copyrighted content is less accurate for video than for audio, so it may take far longer to identify individuals who use P2P to transfer copyrighted video but not copyrighted audio.

Finally, to shed light on the impact of P2P on sales of content, we find that 22% of P2P users also purchase content from the iTunes Store (iTS), each buying on average about as much content as non-P2P users who purchase from the iTunes Store. This refutes the hypothesis that all P2P users view the ability to obtain free content from P2P as a complete substitute for paying for content. On the other hand, we also find that among iTs users, those who use P2P are somewhat more likely than those do not use P2P to access iTs only for the free samples.

---

[1] Alexandre M. Mateus, PhD candidate in Engineering and Public Policy at Carnegie Mellon University, U.S.A., and at Instituto Superior Técnico, Portugal, amateus@cmu.edu, http://andrew.cmu.edu/~amateus.

[2] Jon M. Peha, Carnegie Mellon University, Professor in the Dept. of Engineering & Public Policy and the Dept. of Electrical & Computer Engineering, www.ece.cmu.edu/~peha

[3] Jon M. Peha contributed to this work in his capacity as a professor at Carnegie Mellon University, and dissertation advisor to Alexandre Mateus. Any opinion expressed herein is that of one or both of the authors, and does not represent the views of the Federal Communications Commission.

1 Introduction

Peer-to-Peer (P2P) networks are used to illegally transfer copyrighted content, although opinions differ as to how large this phenomenon is, how this is evolving, how it affects copyright holders, and how to deal with it. Copyright holders state that P2P piracy heavily impacts their revenues [1]. They have taken legal action against P2P developers and file sharers in the past, and continue to lobby for legislation against file sharing. College students are among the biggest users of file sharing [2, 3], thus drawing attention to P2P in university campuses. Some university networks in the US have recently assumed a larger role in combating illegal transfers of copyrighted media over P2P by deploying deep packet inspection (DPI) technology in an attempt to detect illegal transfers. Such approaches have been mandated for Internet service providers (ISPs) in a few countries [4, 5], and similar mandates have been proposed in the US and elsewhere for ISPs and universities [6, 7].

This paper focuses on online exchanges of media in university campuses, namely using P2P, the iTunes Store and YouTube. We seek to fulfill three main objectives. First, to quantify the extent of P2P usage and of transfers of copyrighted content using P2P on campus, and how these are changing over time, to help assess the need for intervention, namely the need for deployment of technology for copyright protection or the need for new legislation regarding this matter. Our second objective is to assess the effectiveness of DPI in detecting P2P transfers of copyrighted content, as effectiveness is a major factor to be considered when deciding whether to deploy DPI technology, or whether to mandate or adopt a policy that requires its deployment. Finally, we seek 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. We do so by quantifying the extent to which media is obtained from different online sources such as P2P, the iTunes Store or YouTube, and by correlating usage of these sources.

There have been previous assessments of the extent of P2P and illegal transfers of copyrighted content using P2P on university campuses. By means of surveys, such studies found that over half of college students engaged in P2P file sharing [2] and that college students amounted to 21% all P2P users [8, 9]. Moreover, that a significant share of students' media libraries was composed of music obtained from P2P [10, 11] and that college students got more of their music from P2P than the rest of the population [3]. Results from such survey-based studies depend on the memories and openness of survey respondents, or in how survey instruments are designed and subjects 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 this paper, we present 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, we can also access information that Internet users may not know, such as the volume P2P transfers or the time of such transfers.

Music industry representatives assert that P2P is responsible for billions of dollars in lost sales and thousands of lost jobs [10]. 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 [12], different authors present contradictory results. While most papers focusing on this subject find that P2P file sharing contributes to the decrease in music and video sales [13-17], averaging at a sales displacement rate of 20%, i.e., each title downloaded through P2P displaces sales of 20% of a title, others find that P2P has positive effects on sales [13, 14]. Accounts of P2P file sharing in these articles are based on self reported data collected by means of surveys or on rough approximations using proxies such as internet penetration. In the middle stands a set of papers that uses actual measures of file sharing to find that P2P transfers are unrelated to changes in content sales [15-18]. In this article, by observing the online behavior of students towards usage of P2P as well as usage of the iTunes Store, we seek to contribute with some empirical evidence about the extent to which P2P users still purchase content online.

The remainder of this article is organized as follows. In section 2 we present relevant background to frame our analysis. Section 3 presents an overview of our data collection methodology, followed by a summary of the collected data in section 4. Section 5 is our main results section, in which we first draw the general picture of P2P usage on campus, followed by results on technological limitations in detection of copyrighted content transferred P2P, with particular focus on differences in ability to detect different types of content and differences in ability to detect content transferred using different P2P networks, and finally we compare usage of P2P and usage of the iTunes Store and of YouTube to obtain media online. Section 6 concludes this article with a summary of our findings and the implications that can be derived from them in terms of policy.

2 Background

2.1 P2P File Sharing, U.S. Law and implications for Universities

By U.S. law [19], except where “fair use” provisions apply, transfers of copyright-protected works without permission from the copyright holder are infringements of the holder’s rights. 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) [20] 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 violation.

In the past, the music industry, through the Record Industry Association of America (RIAA), used these legal provisions in several lawsuits against P2P companies [21, 22] and users [23]. 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 using “pre-litigation settlement letters” requesting

that infringing students be identified and that the letter be forwarded to them [24]. Since these letters were not legally binding, some universities ignored them, while others forwarded them to students [23]. Upon reception of the letters, students could avoid court action and settle the case using the phone or a website[4]. More recently, the industry reported the end of their lawsuit campaign against P2P users [25], and announced plans to start collaborating with ISPs for copyright protection.

Illegal file sharing has been debated at the lawmaker's level, both in the U.S. and abroad. In the U.S., Congress held at least six hearings on online copyright infringement in universities since 2003 [10, 26, 27] and discussed possible interventions to deal with it [6, 7]. Abroad, the focus was mostly on ISPs, particularly in the E.U., where France is in the process of approving legislation requiring ISPs to disconnect users detected transferring copyrighted material [28-30], and in the U.K., where the possibility of similar legislation was a subject of dispute between ISPs, the copyright industry and government [31]. All this activity, both in the U.S. Congress and in some of the larger E.U. countries indicates potential for policy change.

Universities across the U.S. and around the world have adopted many different practices in response to P2P use among students [23]. These approaches, both technical and non-technical, include identifying the users of infringing IP addresses reported by copyright holders and forwarding the notices to those users, disconnecting those users from the network, investing in education, facilitating access to legal services, and attempting to prevent users from illegally sharing copyrighted content [23]. Among those measures is the use of deep packet inspection (DPI) technology in an attempt to detect illegal transfers over P2P. This technology has recently been deployed by some university networks in the US and recommended in legislation recently enacted [32]. Its deployment represents a considerable investment for universities [33, 34] and investigation is needed to determine whether it will be able to effectively detect transfers of copyrighted content in the long term.

2.2 The Arms Race Between Network Monitoring and Evasion Technology

P2P technology is attractive for distributing information online because all users participate in the distribution and share the burden of transmitting material to other users, thus leaving content originators responsible for only a fraction of the cost of distributing their material. Many legal applications take advantage of this. CNN, BBC and Joost, for instance, distribute video using P2P, while Skype uses P2P for internet telephony, and BitTorrent[5], Gnutella, Ares and eDonkey, today's top general-purpose file sharing networks, are legally used to transfer open source software and popular game updates. However, the latter general-purpose P2P networks also make it easy to illegally transfer copyrighted content, a feature used by thousands of file-sharers worldwide.

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

[5] 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.

There is contention regarding the dimension of general-purpose P2P networks, in terms of number of users, number of shared files, and traffic that is generated. In a recent online report, Zhang [35] reviews multiple sources in the literature in an attempt to establish boundaries for the percentage of P2P traffic in networks. Such review found a wide range (9% to 93%) of measures of P2P traffic depending on when (year, time of day) and where (geographical location, backbone vs. edge, campus vs. commercial ISPs) the measurements happened, but with the large majority of measurements reporting over one third of network traffic as being P2P.

In an effort to deal with the burden that P2P imposes on their networks, many ISPs have turned to traffic shaping technology[6] to identify and throttle P2P traffic [36-38]. Early traffic shaping relied on information about the ports used by each pair of communicating parties to infer the type of traffic that composed the communication[7]. P2P protocols evolved in response to that shaping and started to use random ports or to masquerade as other protocols by using ports traditionally employed by those protocols. This caused traffic shaping technology to adopt deep packet inspection (DPI) to unravel the protocol in each packet independently of the ports used for communication. Once again P2P protocols were enhanced, this time with the ability to encrypt P2P control traffic (or all traffic) and thus make protocols more opaque to traffic shapers. Yet again, traffic shaping technology moved forward and started taking advantage of behavioral[8] patterns in P2P traffic to detect and restrain it [39]. In response to this latest advancement users can turn to virtual private network (VPN) providers, which allow them to establish private encrypted tunnels to servers that then relay all their traffic. VPN providers have existed for a long time, catering to specific needs for enhanced Internet traffic security and user anonymity. However, as more ISPs engage in this latter form of traffic shaping, the number of VPN providers that focus specifically on tunneling of P2P traffic is growing, enabling users to evade their ISP's detection and throttling practices for a small monthly fee [40-42].

Besides traffic shaping, and among many other uses, DPI is also used as a component in detection of copyrighted content being transferred in P2P networks. This is achieved by, after detecting which transfers are P2P, observing the content being transferred and identifying whether it is copyrighted. Such content detectors have been used mostly in smaller networks (such as university campus networks), partly due to their processing power requirements, which prevent them from being used in higher speed links. However, they are ineffective against users that employ encryption of all their P2P traffic because the content of transfers by those users is completely opaque to the detectors.

---

[6] Traffic shaping restricts the amount of data from a traffic class that can be transmitted in a given amount of time, typically in order to improve the performance of other traffic classes.

[7] For instance, HTTP traffic typically uses port 80, whereas FTP communicates through ports 21 or 22, and early BitTorrent used port 6881.

[8] Taking advantage of patterns only observable in P2P communications, such as connecting to many different parties in short time intervals or communicating with identified sets of addresses, among others.

Another technique used to detect copyrighted content in P2P is swarm infiltration (SI). It consists of using a modified P2P client that connects to a P2P network as a regular user and collects information about other network users sharing particular titles. SI is used mostly by copyright owners to collect information about P2P users illegally transferring their content, particularly the IP address of such users and the time at which the detection occurred, which they later use to take action against those users. However, once again, P2P users are presented with means of evading this type of detection. In this case, the same type of VPN service used to evade P2P throttling by ISPs is also marketed as an effective means of protection from SI detection. These VPN providers aggregate traffic from multiple users in a small pool of IP addresses. Since they have no requirement to maintain data that can identify their users based on such IP addresses, and often do not, they provide protection of their users against action from copyright holders.

Clearly, both in the case of traffic shaping and of detection of copyrighted content, an arms race is occurring. On one side, ISPs, network managers or copyright holders implement measures to detect P2P users; on the other side, P2P developers, P2P users or other stakeholders deploy counter-measures to avoid such detection. It is uncertain whether any of the sides will ultimately win. However, enacting ever more intricate traffic obfuscation measures has the potential to produce side effects, such as undermining the effectiveness of tools that can enhance quality of service or improve security.

3 Methodology

This research was performed on data collected in the scope of the Digital Citizen Project (DCP), a project undertaken by 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” [43]. 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. This section describes the methodology utilized for collection and anonymization of the data used in this article, some of which is also described in our previous work [44].

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. The wired connection in student’s dorm rooms has an assigned fixed IP address for the entire semester. Wi-Fi is not allowed in ResNet.

Data was collected through network monitoring performed by two commercially available monitoring appliances that use deep packet inspection (DPI): Packeteer PacketShaper[9] (from now on referred to as Packeteer) and Audible Magic CopySense[10] (from now on referred to as AM). Both devices log relevant attributes of transmissions between users inside the campus network and outside parties, for traffic routed using commodity ISPs. Packeteer had already been deployed before this project to perform traffic shaping as described above. The device classifies communication sessions in over 500 classes[11] according to the type of traffic that composes them. This device neither examines nor retains the actual contents of the communications sessions.

The AM device was purchased to enforce ISU policy before CMU got involved. AM uses header information to identify P2P streams. Within those P2P streams, AM identifies copyrighted media in real time by trying to match the transferred material against a database of audio fingerprints of copyrighted media titles[12] or hash codes[13] used to identify files in P2P networks. The device does not retain any portion of the transmission, but it does record which copyrighted material was matched in the database. When the material being transferred cannot be matched against anything in the database, AM records a piece of the metadata incorporated in the transfer (typically the name of the file being transferred).

AM logs information on communications in the form of *events*. An event corresponds to one or more consecutive TCP or UDP[14] 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, 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 AM's database) over a set of consecutive TCP sessions or consecutive UDP sessions.

[9] 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/

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

[11] 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 [45].

[12] 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.

[13] 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.

[14] 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.

3.2 Connection of Monitored Activity to Users and Devices

The identification of the network user and device responsible for each online activity detected by AM 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[15]) is obtained by performing a lookup in the DHCP[16] lease logs using the IP address of the monitored activity and the time when the activity occurred. Activity detected by Packeteer did not go through this process.

Information about the user that performed each activity consists of the user's University Login Identification (ULID)[17], 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 DCP policy guidelines, which include, but are not limited to, the following measures to protect the privacy of monitored users. 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 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[18]. 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. The monitoring and anonymization processes were controlled by 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 all the research described in this paper.

---

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

[16] 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.

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

[18] 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 Summary of Collected Data and Definitions used in Analysis

AM and Packeteer collected data during three periods of about one month each in the Spring 2007, Fall 2007, and Spring 2008 academic terms. In each of the periods, AM collected a log of events as described in the previous section. In Spring of 2007 Packeteer collected hourly summaries detailing the total amount of bytes and communication sessions entering and exiting the ISU network, broken down by protocol/application. In Fall 2007 and Spring 2008, Packeteer collected one individual record per detected communication session. Each such record is a Netflow v.5 record[19] augmented with identifiers of the protocol/application used in the communication. Table 1 presents the number of people living on campus and provides a brief summary of the data collected by each appliance in each period.

Table 1. 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 / weeks with AM data	648 / 25 / 3	654 / 26 / 2	1,747 / 60 / 6
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 / weeks with Packeteer data	720 / 30 / 4	735 / 29 / 3	858 / 31 / 3
Number of Packeteer events collected	hourly summaries	3.3 billion	4.3 billion
Full hours / days / weeks with both AM and Packeteer data	642 / 25 / 3	541 / 20 / 1	770 / 24 / 1

Data collected through network monitoring is always dependent on the visibility that monitoring devices have of the network that is being monitored. In this case, both monitoring appliances were deployed at the point where the campus network connects to commercial ISPs, which means that only communication sessions in which one party is inside the campus network and another party is in the external Internet can be detected. Hence, none of the data collected by AM or Packeteer contain records of intra-campus communication sessions nor of communications routed through Internet2.

4.1 Definitions used in the Analysis

The analysis in this article revolves around two main types of activity: the usage of P2P and the usage of P2P to transfer copyright-protected media. We define a P2P activity to be any 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 the 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 the monitored period. The remainder of this section presents several important remarks about the concept of a DATCoM.

[19] 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

Not every DATCoM is a copyright violation [19]. For instance, in some DATCoMs, users may be downloading material that will be used in particular ways that fall under the "fair use" doctrine [46]. Using the collected data, we cannot know whether or not the copyrighted material in each DATCoM will be used in any way that can be considered "fair use", therefore such considerations are outside of the scope of this paper. 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[20], and the other party had to search for that content and instruct her P2P client to download it. Our results do not change significantly if we disregard such "empty" DATCoMs because nearly all DATCoM users and copyrighted titles were detected in multiple DATCoMs, at least one of which containing enough bytes to actually correspond to a transfer, instead of only a failed request. Hence, while we do not claim that all DATCoMs detected on campus are copyright violations, most of them probably are, and we believe that DATCoMs 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 economic impact of P2P. However, the available data does not allow drawing significant conclusions regarding downloads vs. uploads.

5 Results

This section presents our main results. We start in section 5.1, with a characterization of the extent and evolution of detected P2P usage by students living on campus and of detected usage of P2P to transfer copyrighted content. This provides a characterization of the activity we were able to observe on campus. To interpret these figures it is necessary to account for limitations in the monitoring technology used to collect data, and the possibility that students are actively attempting to conceal their P2P activity from detection. In the following sections we assess the impact that these factors have in the observed trends, as well as more generally in usefulness of DPI technology for detection of copyrighted content in P2P networks. Namely, section 5.2 demonstrates why the results presented can only be interpreted as lower bounds on the extent of P2P usage on campus. Section 5.3 focuses specifically on limitations in detection of copyrighted content of different types, particularly in detection of copyrighted audio versus copyrighted

[20] Whether or not making copyrighted content available constitutes a copyright violation is currently the subject of legal dispute beyond the scope of this paper [47].

video. And section 5.4 deals with limitations in detection of copyrighted content transferred using different P2P networks. Finally, in section 5.5 we assess the extent to which illegal transfers of copyrighted content using P2P displace legal transfers or purchases of content from other online sources, in particular the iTunes Store and YouTube.

5.1 Extent and Evolution of Detected P2P Activity

Each of the three monitoring periods we observe in our dataset represents part of an academic semester in the 1-year period between Spring 2007 and Spring 2008. In Spring 2008, the latter of those monitoring periods, we found P2P usage, particularly to transfer copyrighted content, to remain widespread on campus. As table 2 shows, in Spring 2008 about 40% of students living on campus were observed performing P2P, 70% of those were detected transferring copyrighted content over P2P at some point, averaging at over 4 copyrighted movies, songs or TV shows per day.

Table 2. Summary of percentages of students detected performing P2P and engaging in DATCoM, and of number of copyrighted media titles detected per student detected overall in the Spring 2008 monitoring period.

	Out of students living on campus	Out of detected P2P users	Out of detected DATCoM users[21]
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)

In spite of remaining widespread, we observed a generalized decrease in P2P activity over the one-year period leading to Spring 2008. This is clear from Figure 1.a, which shows the declining daily percentage of users detected engaging in P2P and transferring copyrighted content, and from figure 1.b, which plots the decrease in the daily average number of copyrighted titles detected being transferred per student living on campus or per detected DATCoM user. To compare each monitoring period we use daily averages[22] because figures for the whole duration of each monitoring period are not meaningful for inter-period comparisons due to different number of monitored hours in each period[23].

[21] 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.

[22] 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. Furthermore, we tried to compare "regular" days in terms of P2P activity, and for that we disregarded outlier days, such as Spring break, Easter or Labor day weekend, in which the percentage of students present on campus was much lower.

[23] Perhaps averages per week would provide more meaningful terms of comparison between periods, but Fall 2007 included only one consecutive week of usable data, which would not provide a fair comparison to the other periods.

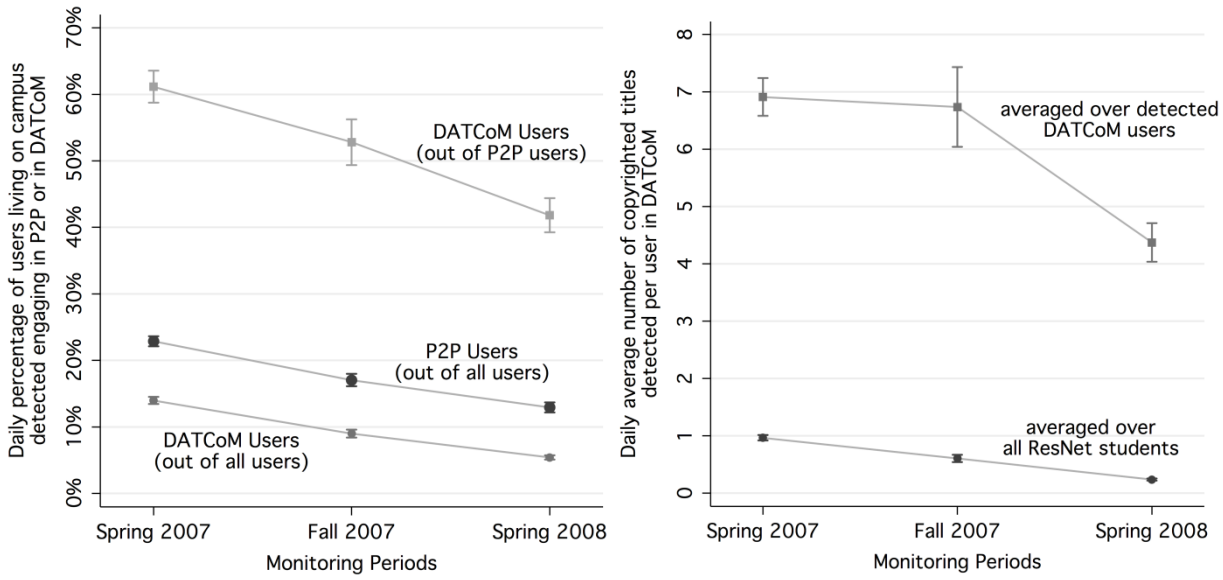


Figure 1. (a) Average for each period of the daily percentage of students detected engaging in P2P out of all students living on campus, and of the daily percentage of students detected engaging in DATCoM, out of all students living on campus and out of detected P2P users in each day. (b) Daily number of copyrighted media titles detected in DATCoM in each of the three monitoring periods, averaged over all students living on campus and over students detected engaging in DATCoM in each day. Caps represent 95% confidence intervals.

Variations of the figures presented above over demographics, i.e., by gender, class, birth year or major, had been reported for the entire Spring 2007 monitoring period in [44]. For that period there were very small differences between demographic categories, some of which were statistically significant yet not particularly relevant in policy terms. The extension of that analysis to the three monitoring periods shows that, while the decrease in percentage of P2P users, DATCoM users, and titles per DATCoM user, for each demographic category, is in line with the decrease observed for the entire population, the differences among categories remain very small. Such result adds more evidence towards the fact that P2P and transfers of copyrighted content were and remain widespread on campus. Besides that, it is clear that demographics are not useful for targeting interventions, and that, whatever the incentives were for the observed decrease, they 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 in others they stopped P2P activity, resulting in the uniform decrease we observe.

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 [2], is still in line with figures previously obtained by means of surveys that reported widespread use of P2P in other university campuses.

5.2 Limitations in Monitoring Technology and Reliability in Detection of Copyright Violations

The trends presented above portray the evolution of detectable P2P activity and the evolution of detectable transfers of copyrighted content over P2P on campus. To interpret those figures we need to consider the extent to which monitoring technology fails to detect P2P traffic or transfers of copyrighted content over P2P, as well as the extent to which it classifies content that is not copyrighted as being copyrighted. In this section we focus on those two points, both to help interpret the trends presented in the previous section, but also to inform stakeholders considering DPI, since accuracy and effectiveness of DPI technology are factors that must be considered when deciding whether to deploy it, or adopt a policy that mandates its deployment.

The percentage of P2P users detected on campus should be interpreted as a lower bound because the monitoring appliances that we used are unable to detect traffic as being P2P if that traffic is encrypted and it is possible that some campus users encrypted their P2P traffic. Since users that encrypt all their P2P traffic will not show up as P2P users in the data we collected, the impact of P2P encryption in our results depends on whether or not users are knowledgeable of encryption and willing to activate it[24]. Use of network monitoring (either DPI or other methods described in section 2.2) 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 such measures, but it remains a question whether such incentive is enough for users to act. As described in section 2.2, further technical advances may yield cost-effective network monitoring tools that detect encrypted P2P as P2P, but no form of network monitoring can determine whether the content transferred is copyrighted.

The percentage of DATCoM users and number of copyrighted titles detected per user should also be interpreted as a lower bound because the monitoring appliances that we used fail to detect some copyrighted content transferred over P2P as being copyrighted. To detect whether content is copyrighted, today's DPI technology, namely AM, extracts pieces of content being transferred over P2P and compares features of that content to features of known copyrighted titles. In the case of AM, such features are either hash codes[25] or audio fingerprints[26]. This approach fails to detect copyrighted titles in several circumstances.

First, it can only identify content whose features are present in a pool of identifiable files or titles. Copyrighted titles not present in such pool will never be identified as copyrighted. In the particular case of AM, the content of the hash code and audio fingerprint databases is not public information, which means that we cannot know what the pool of identifiable media is. However, we do know that such pool contains

[24] 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.

[25] See footnote 13.

[26] See footnote 12.

only features of songs, movies and TV shows and that it is updated regularly with newly released titles[27] using input from the music, movie and television industries. Hence, 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 [48].

Second, it needs to extract enough content from the P2P transfer to allow a meaningful comparison to features in the pool of copyrighted titles. The amount of content needed varies depending on the features being compared and on the type of content. In the case of AM, for hash code comparisons, only the hash code needs to be extracted, but 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. Extracting a large enough piece of media being transferred over a single P2P communication session can be problematic because exchanged files are most of the times broken down in small pieces, each transferred in a different communication session.

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, we cannot rule out the possibility that such comparison fails to produce a positive identification.

The above limitations imply that the percentage of detected DATCoM users out of those detected engaging in P2P obtained from AM monitoring can only be interpreted as lower bounds. However, despite such limitations, and given enough time, AM can still detect most users that attempted to transfer copyrighted content out of those users detected performing P2P. This is shown in figure 2, which depicts the ratio of cumulative detected DATCoM users to cumulative detected P2P users as a function of the number of monitored hours in each period. It is clear that, in the first hours of monitoring the percentage of detected P2P users observed in DATCoM was low, but after some weeks of monitoring that percentage tends to stabilize around a fixed value. Hence, AM is more effective in detection of users that transfer copyrighted content out of those detected doing P2P if it is given a longer monitoring period.

[27] 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).

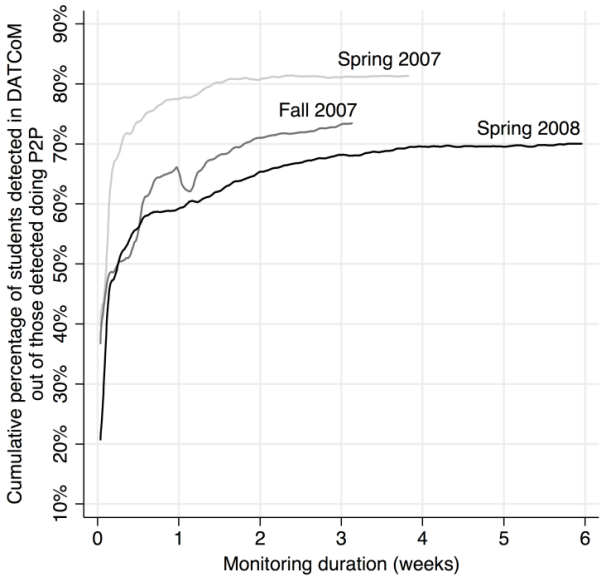


Figure 2. Cumulative evolution over the duration of the monitoring period of the percentage of users detected engaging in DATCoM out of P2P users detected since de beginning of the monitoring period.

Finally, to assess the effect that detecting fewer DATCoMs would have in detection of users transferring copyrighted content, we re-sampled the AM data by removing a percentage of detected DATCoMs at random, and then observed the number of detected users, titles and user×title pairs in the remaining DATCoMs. We performed such re-sampling 1000 times for each percentage between 0% and 100%. The results from this simulation are portrayed in figure 3, which shows that not detecting some DATCoMs at random has little impact on the number of detected DATCoM users, but a higher impact in the number of detected copyrighted titles or user×title pairs.

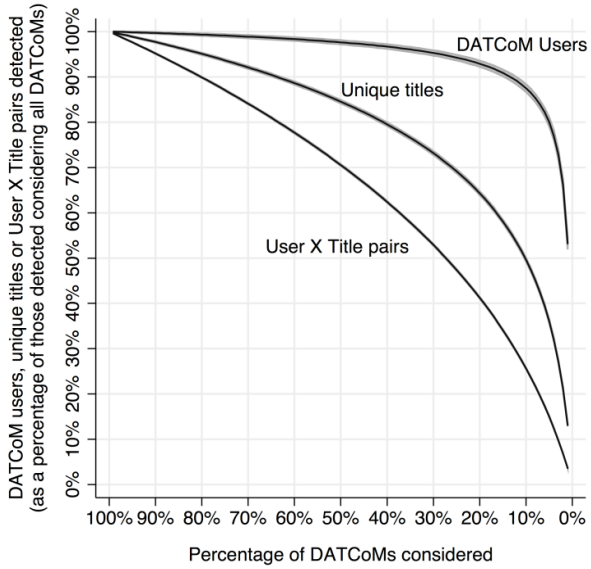


Figure 3. Percentage of DATCoM users, unique copyrighted titles and user×title pairs that would be detected under smaller percentages of detected DATCoMs. Values calculated as percentages of the number of DATCoM users, unique copyrighted titles and user×title pairs detected using all DATCoMs. Lines represent the mean percentages and shaded areas represent two standard deviations from the mean.

The figure shows that it is possible to miss most DATCoMs and still meet the goal of detecting most users who transfer copyrighted content at some point: with only 15% of the DATCoMs detected in a month-long monitoring period, it is still possible to detect 90% of DATCoM users. This is because most users detected with at least one DATCoM were detected with multiple DATCoMs (note that monitoring duration is an important factor.) This interpretation assumes that the number of missed DATCoMs would be uniform across users, an assumption that is violated, 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.

All of AM's limitations discussed above result in false negatives, i.e., communication sessions in which copyrighted content was transferred but that were not classified as DATCoMs. Another type of error that can possibly occur in detection of copyrighted content is to have communication sessions classified as a DATCoMs when the content transferred therein was not copyrighted, i.e., false positives. This type of error is particularly problematic if the results of detection are used to act upon the user supposedly performing the activity in question. In the case of our results, a high false positive rate could prevent the interpretation of the figures reported in the previous section as lower bounds because the number of detected DATCoMs could be inflated and lead to the identification of users transferring copyrighted content when they actually did not. To assess the extent to which this is happening we would need to obtain information on AM's false positive rate, which is not available. However, we have reasons to believe that it is low, given that this particular technology is implemented in various high-visibility outlets[28] where any false positives would lead to great backlash, which has not been observed[29].

In conclusion, AM is more reliable in detecting users that transfer copyrighted content than in detecting individual copyrighted titles transferred by each user, and given enough time it will eventually detect 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. Due to the fact that encryption prevents DPI from looking at the content of communications to detect whether they contain copyrighted content, if the use of encryption in P2P ever becomes generalized, then DPI is rendered useless for detection of transfers of copyrighted content.

5.3 Detection of Different Types of Media Transferred over P2P

In this section we characterize the types of content that students living on campus transfer over P2P and use that information to assess how well DPI can detect transfers of copyrighted content from each of those types. To assess how well DPI detects different types of content we use data collected from AM,

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

[29] 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 [49]. 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.

which, although 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. 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. AM in particular can only detect songs, movies and TV shows as copyrighted 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[30].

To assess which types of content were transferred by students living on campus we use media titles contained in DATCoMs, as well as 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 DATCoM and over 100 thousand distinct filenames in metadata. Overall in the three monitoring periods, DATCoMs were detected for an average of 74% of detected P2P users, and communication sessions 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 we break down all media titles detected in DATCoM and filenames detected in Metadata according to the type of content they advertise[31], and compare rates of detection of particular types using DATCoM to rates using filenames. Filenames are used as a control group against which we test detection of copyrighted titles in DATCoMs. They provide a good 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 separate these two cases we break down detected filenames in different categories of content and analyze those where the probability of a file being copyrighted is greater (filenames indicating known songs, music albums, movies and TV shows). Also, we have no guarantee that the filename actually represents what is contained within the file[32]. For instance, we cannot know whether a file whose filename contains the title of a well known copyrighted work actually contains that work.

[30] 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.

[31] 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.

[32] 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 [50].

However, since today's most popular P2P networks sport some type of content rating system which allows for "fake" files to be tagged 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 3 presents this breakdown of detected titles and filenames by type of content. It also shows that most filenames indicated songs, movies and TV shows, but about 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 3. 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		8.9%
Song	99.2%	62.0%
Album		2.6%
Movie	0.5%	4.7%
TV Show	0.3%	3.7%
Adult / Software / Books / Pictures		18.0%

Focusing only on types detected by both methods, 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 4: 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. One possible explanation for this fact 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 we cannot test, 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 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, and it is also prone to being tested with the data we have available.

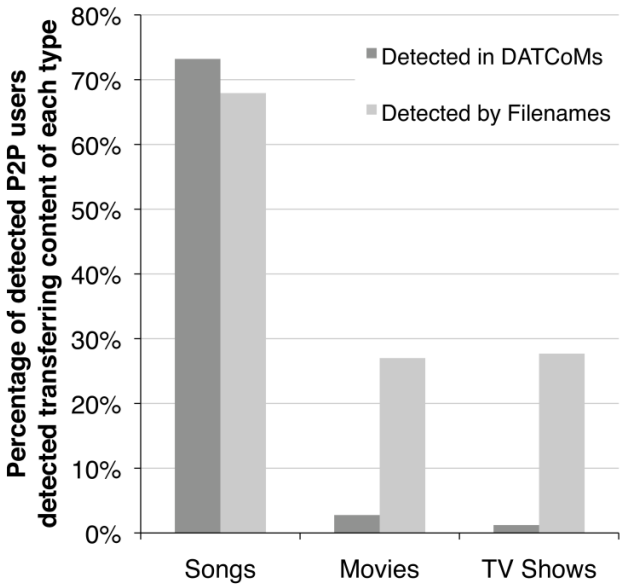


Figure 4. Average for the three monitoring periods of the percentage of detected P2P users detected transferring songs, movies or TV shows by means of DATCoMs and by means of filenames.

By comparing DATCoM and filename detection rates for a set of songs and movies known to be present in AM's database we find that AM fails to classify copyrighted video as copyrighted more often than it fails to classify copyrighted audio. For each of the top 100 copyrighted song and movie titles detected in DATCoMs, we gathered all filenames that indicate the same content and collected the number of users detected by means of DATCoMs, of filenames, and of both DATCoMs and filenames. We assume 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. We also assume that the percentage of filenames that correspond to the actual content in the file is equal for movies and for songs. If both assumptions hold, then, if AM could detect a copyrighted video as being copyrighted as well as it can detect a song we would expect the percentage of people detected by DATCoMs (out of those detected by either DATCoMs or filenames) to be the same on average for movies and for songs.

However, figure 5 shows that this is not the case. The figure presents, in each monitoring period, the average percentage of P2P users detected transferring the song and movie titles in DATCoMs out of all users detected transferring the titles (in DATCoMs or by filenames). It is clear that for songs, more people are detected through DATCoMs than for 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% (15% to 36%) in Fall of 2007 to a high of 48% (36% to 59%) in Spring of 2007.

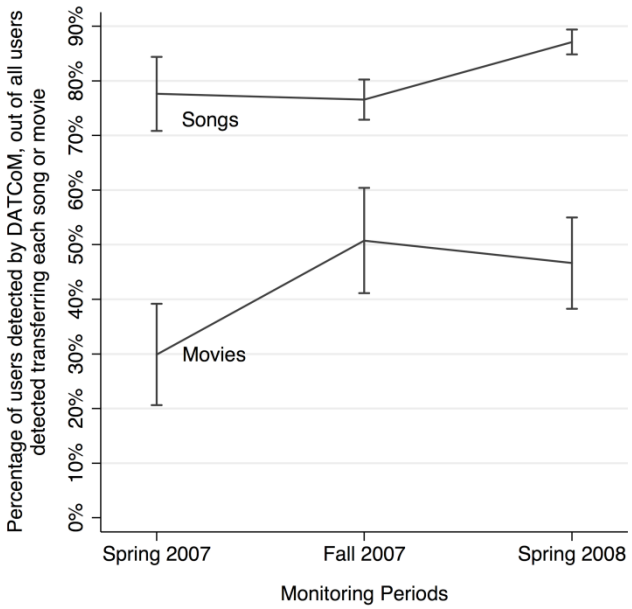


Figure 5. Average percentage of users detected by DATCoM transferring each song and each movie out of all users detected transferring each song or each movie (by DATCoM or by filename).

Hence, we conclude 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 (AV, comprising songs, movies, TV shows and music albums). This is clear in figure 6, 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, clear in 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 difference between the percentage of P2P users detected by DATCoMs and that of P2P users detected by filenames.

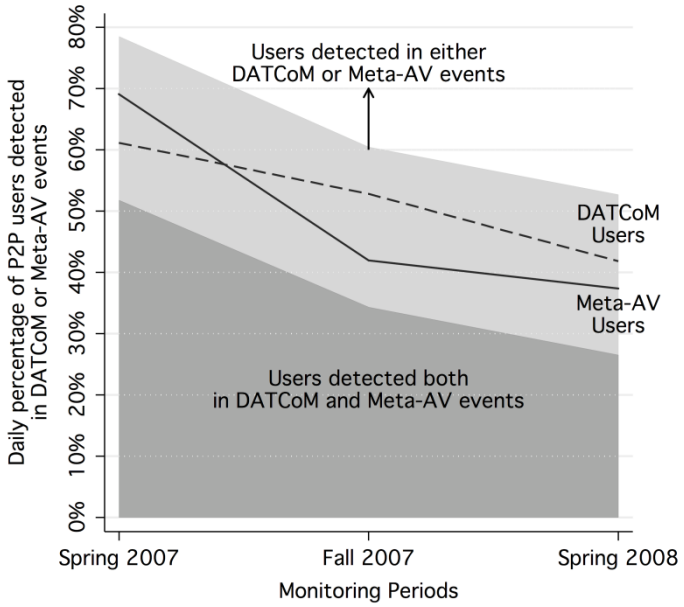


Figure 6. Average daily percentage, for each monitoring period, of users detected engaging in DATCoMs (DATCoM users), in activities containing Metadata whose filenames indicate songs, albums, movies or TV shows (Meta-AV users), in either of those two, or in both of them, out of P2P users detected in each day.

5.4 Detection of Copyrighted Content in Different P2P Networks

Multiple P2P networks are nowadays used on the Internet. P2P networks differ in multiple aspects, the most relevant in this case being differences in availability of various types of content in each network. In this section we characterize the P2P networks detected on campus, differences in usage between them and differences in detection of copyrighted content transferred in each of them to assess whether copyright infringement is easier to detect in certain networks than in others.

In all monitoring periods we detected traffic from several P2P networks on campus[33]. However, most of those networks represent a residual percentage of traffic. The main P2P networks detected on campus were BitTorrent, Gnutella, eDonkey and DirectConnect, altogether accounting for over 95% of all detected P2P traffic. Out of those, as figure 7.a shows, BitTorrent and Gnutella were clearly the dominant networks, accounting for over 90% of traffic. This dominance is consistent with reports on dominant P2P protocols overall in the Internet [51].

Focusing on BitTorrent and Gnutella, we see that BitTorrent dominates in terms of traffic, and its share of traffic is increasing over time, taking up the room made available by a decrease in Gnutella traffic. However, Gnutella dominates in terms of users, as figure 7.b shows. While the decrease in Gnutella traffic

[33] 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.

is easily attributable to the decrease in number of detected users, in BitTorrent we observe a growth in traffic volume despite an even more pronounced decrease in number of detected users. This apparent contradiction can be reconciled when taking into account the differences between content transferred in each of the networks, as discussed next.

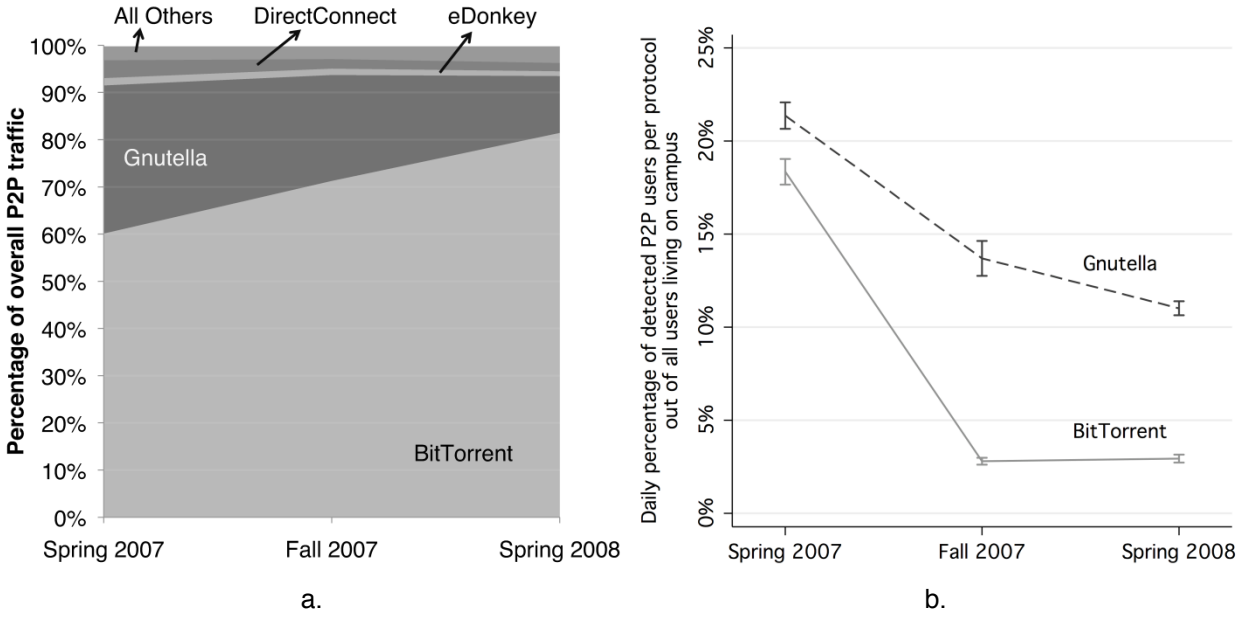
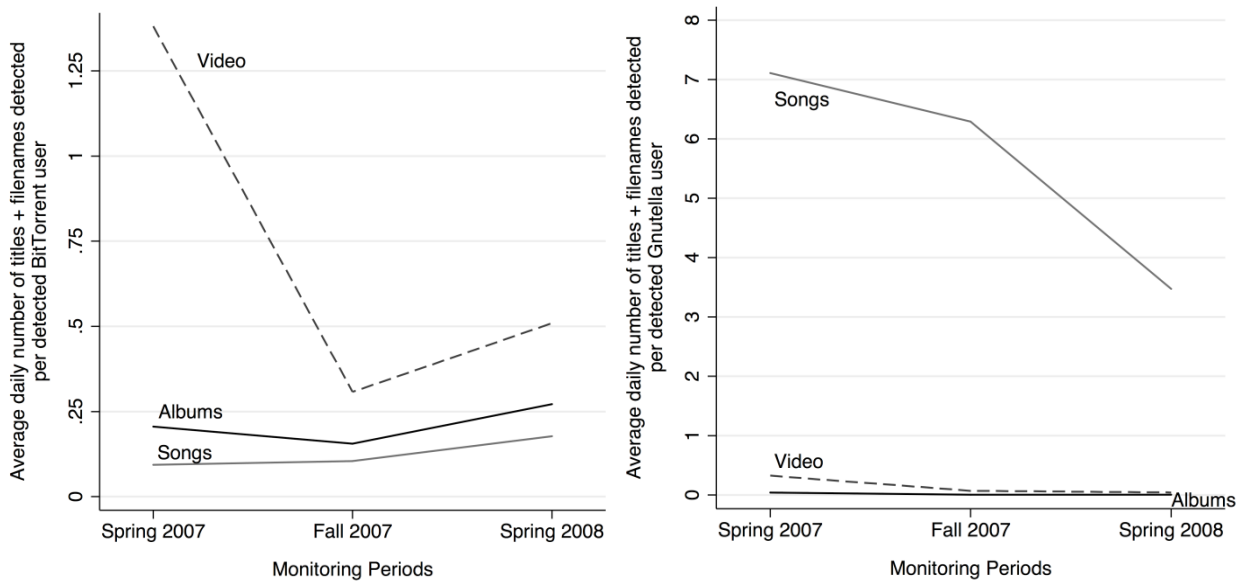


Figure 7. (a) Break down of percentage of P2P traffic by P2P network over the three monitoring periods. (b) Evolution of the percentage of detected BitTorrent users and Gnutella users on campus over the three monitoring periods.

Detected Gnutella users are observed transferring mostly songs, while detected BitTorrent users transfer more video (movies and TV shows) and music albums inside archives. This is clear from figure 8, which plots the average daily number of titles + filenames detected per detected BitTorrent or Gnutella user broken down by type of content. 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 figure 8 explains why there is less traffic detected for Gnutella than for BitTorrent despite the greater number of detected users.

The percentage of users detected transferring copyrighted content is much lower among users detected using BitTorrent than among users detected using Gnutella; figure 9 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 be explained by the content transferred using each of the networks. As we have established before, AM has a harder time detecting copyrighted video than copyrighted audio and it cannot detect full albums transferred inside archives as copyrighted, which are the two principal types of content transferred by detected BitTorrent users. 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.



a. BitTorrent b. Gnutella

Figure 8. (a) Average daily number of titles + filenames detected being transferred per BitTorrent user, broken down by type of content, over the three monitoring periods. (b) Average daily number of titles + filenames detected being transferred per Gnutella user, broken down by type of content, over the three monitoring periods.

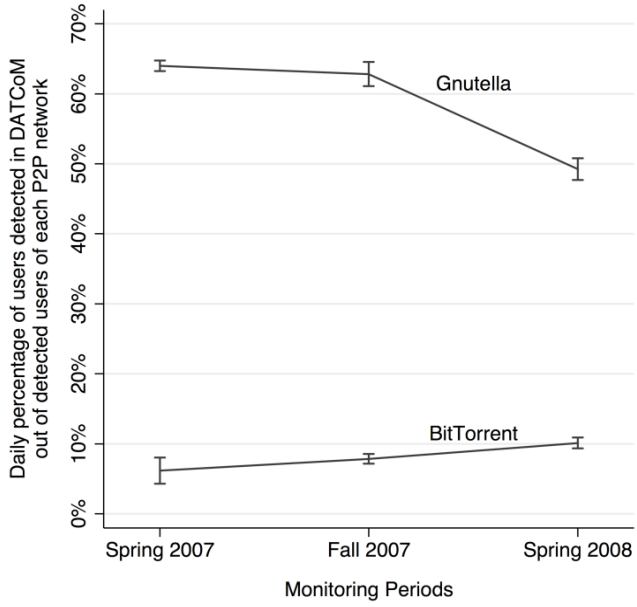


Figure 9. Average daily percentage of detected DATCoM users out of detected BitTorrent users and out of detected Gnutella users, for each monitoring period.

5.5 Relationship between usage of P2P and usage of iTunes or YouTube

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 we cannot assess how many sales of copyrighted content fail to happen due to P2P, we can compare 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. In this section we compare 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 from YouTube, and in the case of the iTS, whether there are differences between download of free content (song samples) and content that is paid for (songs or videos).

We compare P2P, iTunes and YouTube activity for each IP address in the ISU network using communication sessions classified by Packeteer as containing P2P, iTunes or YouTube traffic. While IP addresses do not map necessarily to network users, we believe they provide a good approximation in this case since IP addresses in ISU's residence halls were static for the monitoring period, which allows accounting separately for each user connected to the network in dorm rooms. However, our data set is composed by activity detected for all IP addresses on campus, some of which do not map to a single user during the entire monitoring period. For this reason, we removed two groups of detected IP addresses from this analysis, as discussed in the following paragraphs.

Packeteer was deployed on the Internet side of the Network Address Translator (NAT) at the border of the campus network, which caused all IP addresses from within campus to be detected in the form of ISU-external translated addresses. By means of reverse translation post-processing, implemented using the NAT logs, it was possible to attribute network activity to the respective ISU-internal addresses. However, this reverse translation was not always successful, which resulted in some detected network events impossible to ascribe to ISU-internal IP addresses, and therefore impossible to attribute to a single particular device. The first group of IP addresses that we did not include in the analysis was composed by these addresses, for which the reverse NAT translation failed. Communication by these addresses amounted to 28% of all traffic detected on campus, which we assume to be missing uniformly across protocols. Translation failures are not related to the type of activity contained in the events[34], which makes us expect these 28% bytes to be missing uniformly across protocols, thus not biasing results towards any type of activity in particular. However, we cannot rule out the hypothesis that translation failures occur with higher incidence for certain IP addresses, who happen to engage more in determinate activities. This could possibly introduce bias against those activities.

From the remaining IP addresses, we did not consider for analysis those likely to correspond to short DHCP leases, because they will not capture the behavior of a single network user, but most likely of several users that were assigned that particular address over time. These are not very common in the ISU campus, being mostly assigned to users of the wireless network, which is only available in few places on campus. DHCP-leased IP addresses appear in our data either only once for a short period of time

[34] NAT translation 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.

(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). Hence, we did not use in the analysis IP addresses only seen online for consecutive periods of less than 5 hours[35], which amounted to about 9% of IP addresses that transferred on average 8MB of traffic each in the entire monitoring period. The percentage of such addresses detected doing either P2P, iTunes or YouTube traffic was not statistically different from zero.

To separate transfers of different types of media from the iTS and from YouTube, for each IP address, we categorized each detected inbound communication session based on the amount of bytes transferred in the session. iTS activity was separated into control traffic, sampling of music, downloading of songs and downloading of videos, while YouTube activity was separated into control traffic and viewing of videos.

To separate iTS communication sessions into the different media categories we used the following criterion[36]: 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[37]. Given that an iTunes user can generate traffic without entering the iTS[38], and we are interested song- and video-related activities in the iTS, we consider that an IP address used the iTS if at least one sampling, song or video activity was detected for that address. Such addresses correspond to 41% of addresses with detected iTunes traffic.

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[39]). 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 we can only tell which IP addresses performed P2P, not which ones transferred copyrighted content via P2P. However, we still believe we can draw meaningful comparisons between obtaining music and videos from P2P and from other outlets because, in Spring

[35] 5 hours is the mode of the distribution of number of consecutive hours spent online in any period of consecutive activity detected for any IP address.

[36] This criterion was defied 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.

[37] 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.

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

[39] 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.

2008, at least 70% of the users detected doing P2P were also detected attempting to transfer copyrighted songs, movies or TV shows, and an even higher percentage was detected transferring files whose filenames indicated songs, movies or TV shows. Since we are interested in the activity of P2P users that likely transferred some copyrighted media from a P2P network, we consider in this analysis only those IP addresses detected transferring enough P2P bytes to constitute a copyrighted title, a threshold we set at 3 MB (about the amount of traffic necessary to transfer one song).

Using the above criteria to classify IP addresses as P2P, iTS or YouTube users, we find use of P2P to be correlated with use of the iTS and to be correlated with use of YouTube. 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. This is clear in the cross tabulations in table 4, and shows that, to some extent, P2P and the iTS (or YouTube) complement each other. The fact that P2P and the iTS complement each other 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 sales is lost[40]. 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, also in table 4, we break down iTS users between those who only sampled content and those who actually purchased content. We find that about one third of P2P 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. Despite the fact that P2P users who use the iTS are slightly more likely to use it only for sampling than non-P2P users (2.5% vs. 2.1% of all users), most P2P users who use the iTS clearly purchase content there at some point, thus not all iTS revenue is lost to P2P.

Table 4. 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.

		Not P2P	**P2P**
		76.4%	*23.6%*
Did not use the iTS	*84%*	68.0%	16.0%
Used the iTS only to sample content	*4.6%*	2.1%	2.5%
Used the iTS to purchase songs and videos	*11.4%*	6.2%	5.2%
Did not use YouTube	*61.6%*	51.1%	10.5%
Used YouTube	*38.4%*	25.2%	13.1%

[40] 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.

Focusing on users who purchase content from the iTS at some point in the monitoring period, we find no statistically significant difference in percentage of users purchasing songs or videos, or in number of songs or videos purchased per user, between those who did P2P and those who didn't. As figure 10.a 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 songs (or videos) on average as each non-P2P user, as depicted in figure 10.b.

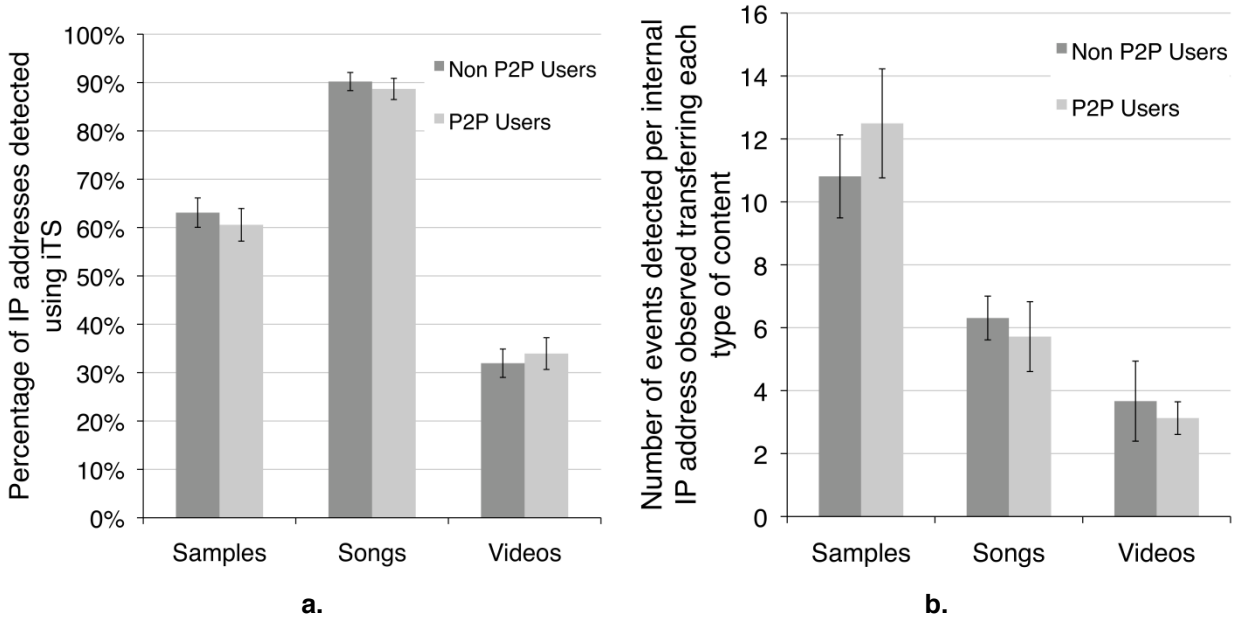


Figure 10. (a) Percentage of IP addresses detected sampling music, downloading songs and downloading videos out of those detected using the iTS to purchase content, broken down by P2P usage. (b) Average number of samples, songs and videos downloaded per IP address detected downloading each of such media from iTS, broken down by P2P usage. Caps represent 95% confidence intervals.

We cannot tell precisely what impact P2P has on paid services from the activity we detected alone, but there are certainly things we can learn. We find 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, we find 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.

6 Conclusions and Policy Implications

In this article, we analyze data collected from a university campus network using DPI network monitoring to assess three main aspects of online activity related to copyrighted content of students living on campus: the extent and evolution of transfers of copyrighted media using P2P networks, the limitations of monitoring technology in detecting such transfers and their implications for the obtained results in particular, and in general for copyright protection of material transferred online, and the relationship between usage of P2P to obtain copyrighted media for free and usage of other online sources of content, such as the iTunes Store and YouTube.

Over three 1-month periods of network monitoring, altogether spanning the duration of one year between the Spring 2007 and Spring 2008 academic semesters, we find P2P activity and the use of P2P to transfer copyrighted content to be widespread on campus. In Spring 2008, the latest monitoring period, at least 40% of the students living on campus were observed using a P2P protocol and at least 70% of those were found attempting to transfer over P2P an average of at least 4 copyrighted songs, movies or TV shows each per day. P2P users and users observed transferring copyrighted content were found across all demographics, with fairly similar incidence between males and females and among users of different ages, classes or majors. The above percentages should be interpreted as lower bounds because the technology used for monitoring failed to detect some of the P2P activity that occurred during the monitoring periods, as well as some activity involving copyrighted content within detected P2P. In particular, it failed to detect as P2P users those users that encrypted P2P traffic during the monitoring periods, and it failed to detect as copyrighted all content not present in a database (which was limited to songs, movies and TV shows). For copyrighted content present in that database, it failed to detect video more often than audio.

Our measurements also show that detected P2P activity on campus decreased over that 1-year period. Specifically, the daily percentage of observed P2P users decreased about 10%, the daily percentage of users observed transferring copyrighted content out of those detected performing P2P fell close to 20%, and the average number of titles detected being transferred by each of the latter decreased from 7 copyrighted titles to about 4 copyrighted titles on average per day. Similar declines were observed in the number of filenames detected, regardless of whether the DPI system deployed could identify the content as copyrighted. Furthermore, decreases were observed in similar proportions throughout all demographic categories. Such decline can actually mean less P2P activity, but it can also mean greater use of methods to conceal P2P, namely the switch to encryption. Despite the decrease in incidence of detected P2P usage and detected transfers of copyrighted content using P2P networks, observed figures are still consistent with numbers previously reported by the entertainment industry.

The limitations observed in one DPI-based device are indicative of the challenges for DPI-based systems in general. First, DPI, cannot determine whether encrypted P2P carries copyrighted content. Indeed, it is challenging to devise a DPI system that can detect encrypted P2P traffic as being P2P at all. Thus, if

encryption becomes common in P2P, then the usefulness of DPI for copyright protection will be severely hampered. Second, the fact that DPI technology can only identify copyrighted content from titles featured in a central database can act as a limiting factor in the amount and type of content that DPI can detect because continuous updates to the pool of detectable titles are needed in order to keep up with new content releases. And finally, there are reasons to expect detection of copyrighted video content to be more difficult than detection of copyrighted audio, and this was certainly observed with AM, which is one of today's leading DPI appliances. Thus, if use of P2P for video becomes more popular, it is likely that a higher percentage of unlawful transfers will go undetected. More importantly, any individuals who use P2P to transfer copyrighted video but not to transfer copyrighted audio are likely to escape detection for much longer. One side effect of the greater difficulty in identifying video is that fewer users of BitTorrent were observed with DATCOMs than users of Gnutella, since video content is more common in the former P2P network.

Despite their limitations, given enough time, the appliances we used could detect most users that attempted to transfer copyrighted content out of those detected doing P2P, a percentage that tends to stabilize around a fixed value after some weeks of monitoring with the current technology and mix of transferred content. Furthermore, since each user was detected transferring copyrighted content in multiple P2P communication sessions during a multi-week period, the percentage of such users is not sensitive to random decreases in the number of communication sessions for which copyrighted content is detected. This means that most users transferring copyrighted content would still be detected even if monitoring technology is unable to identify a large percentage of communication sessions as containing copyrighted content.

Finally, our results contribute to better understanding the impact of P2P on sales of content. While we cannot tell precisely what the impact of P2P is on paid services from observed data alone, there are some lessons to learn from our results. We find some evidence consistent with the hypothesis that some P2P users go to the iTunes Store (iTS) only to obtain free samples; among iTS users, those who also do P2P were 28% more likely to use the iTS only for free samples than those who do not do P2P. On the other hand, our results contradict the hypothesis that users of P2P consistently view the ability to obtain free content from P2P as a superior substitute for paying for content, since 22% of detected P2P users also purchased content from the iTS. Moreover, P2P users who do purchase from the iTS tend to buy about as much as non-P2P users who purchase from the iTS. Thus, even students who engage in free P2P are still making a significant number of media purchases online.

7 References

[1] Recording Industry Association of America [RIAA], "Piracy Online." vol. 2007, 2007, Available online: http://www.riaa.com/physicalpiracy.php?content_selector=piracy_details_online.

[2] J. Lamy, C. Duckworth, and L. Kennedy, "RIAA Launches New Initiatives Targeting Campus Music Theft," RIAA -- The Record Industry Association of America, 2007, Available online:

http://www.riaa.com/newsitem.php?news_year_filter=&resultpage=5&id=0BB7A35D-544B-2DD2-F374-4F680D6BAE9B.

[3] J. Lamy, C. Duckworth, and L. Kennedy, "RIAA Continues College Deterrence Campaign Into 2008," The Record Industry Association of America, 2008, Available online: http://www.riaa.com/newsitem.php?id=36720A8F-FF55-2886-C2A2-EAB629C662BD.

[4] E. Bangeman, "France's plan to turn ISPs into copyright cops on track," Ars Technica, 2008, Available online: http://arstechnica.com/news.ars/post/20080128-frances-plan-to-turn-isps-into-copyright-cops-on-track.html.

[5] C. Bremner, "France to ban illegal downloaders from using the Internet under three-strikes rule." vol. 2008: The Times Online, 2008, Available online: http://technology.timesonline.co.uk/tol/news/tech_and_web/article4165519.ece.

[6] K. Fischer, "Bill would force "top 25 piracy schools" to adopt anti-P2P technology." vol. 2007: Ars Technica, 2007, Available online: http://arstechnica.com/news.ars/post/20070723-bill-would-force-top-25-piracy-schools-to-adopt-anti-p2p-technology.html?rel.

[7] E. Bangeman, "New bill would punish colleges, students who don't become copyright cops." vol. 2007: Ars Technica, 2007, Available online: http://arstechnica.com/news.ars/post/20071111-new-bill-would-turn-colleges-into-copyright-cops.html.

[8] A. Guess, "Downloading by Students Overstated." vol. 2008: Inside Higher Ed, 2008, Available online: http://www.insidehighered.com/news/2008/01/23/mpaa.

[9] S. Oster, "MPAA Statement on Motion Picture Industry Losses due to Piracy among College Students," MPAA, The Motion Picture Association of America, Press Release, January 22, 2008.

[10] C. Sherman, "An Update: Piracy on University Networks," in *110th Congress House Hearings*, SN. 110-29 ed, 2007.

[11] L. Smith, G. Miller, H. McKeon, H. Berman, and H. Coble, "Letter sent to universities on May 1, 2007," 2007.

[12] F. Oberholzer-Gee and K. Strumpf, "File-Sharing and Copyright ": Harvard Business School Working Paper, 2009.

[13] B. Andersen and M. Frenz, "The Impact of Music Downloads and P2P File-Sharing on the Purchase of Music: A Study for Industry Canada," University of London Working Paper, 2008.

[14] R. D. Gopal and S. Bhattacharjee, "Do Artists Benefit from Online Music Sharing?," *Journal of Business* vol. 79, pp. 1503-1533, 2006.

[15] F. Oberholzer-Gee and K. Strumpf, "The Effect of File Sharing on Record Sales: An Empirical Analysis," *Journal of Political Economy,* vol. 115, pp. 1-42, 2007.

[16] S. Bhattacharjee, R. D. Gopal, K. Lertwachara, J. R. Marsden, and R. Telang, "The Effect of Digital Sharing Technologies on Music Markets: A Survival Analysis of Albums on Ranking Charts," *Management Science,* vol. 53, pp. 1359-1374, 2007.

[17] T. Tanaka, "Does file sharing reduce CD sales?: A case of Japan," in *Conference in IT Innovation* Hitotsubashi University, Tokyo, 2004.

[18] M. D. Smith and R. Telang, "Competing with free: The impact of movie broadcasts in DVD sales and Internet piracy," Carnegie Mellon University working paper, 2008.

[19] Library of Congress, "Copyright Law of the United States and Related Laws Contained in Title 17 of the United States Code," Copyright Office, Circular, 2007.

[20] United States Copyright Office, "The Digital Millennium Copyright Act of 1998," Summary, December, 1998.

[21] C. Macavinta, "Recording industry sues music start-up, cites black market." vol. 2007: CNET News.com, 1999, Available online: http://news.com.com/Recording+industry+sues+music+start-up,+cites+black+market/2100-1023_3-234092.html?tag=st.rn.

[22] E. Oswald, "RIAA Sues LimeWire Over Piracy." vol. 2007: BetaNews, 2006, Available online: http://www.betanews.com/article/RIAA_Sues_LimeWire_Over_Piracy/1154722015.

[23] EFF - Electronic Frontier Foundation, "RIAA v. The People: Five years later," Whitepaper, August, 2008.

[24] E. V. Buskirk, "A Poison Pen From the RIAA." vol. 2007: Wired, 2007, Available online: http://www.wired.com/politics/onlinerights/news/2007/02/72834.

[25] S. McBride and E. Smith, "Music Industry to Abandon Mass Suits," in *The Wall Street Journal*, December 19, 2008.

[26] "The Internet and the College Campus: How the Entertainment Industry and Higher Education are Working to Combat Illegal Piracy," 109-58 ed, 2006, Available online: http://frwebgate.access.gpo.gov/cgi-bin/getdoc.cgi?dbname=109_house_hearings&docid=f:30049.pdf.

[27] "The Role of Technology in Reducing Illegal Filesharing: A University Perspective," in *Committee on Science and Technology*, 2007, Available online: http://science.house.gov/press/PRArticle.aspx?NewsID=1858.

[28] E. Pfanner, "France Approves Crackdown on Internet Piracy " in *The New York Times*, May 12, 2009.

[29] N. Anderson, "French court savages "three-strikes" law, tosses it out," *Ars Technica,* June 10, 2009.

[30] N. Anderson, "France govt goes into spin mode to salvage three-strikes law," *Ars Technica,* June 11, 2009.

[31] N. Anderson, "UK ISPs don't want to play umpire to "three strikes" rule," Ars Technica, 2008, Available online: http://arstechnica.com/news.ars/post/20080215-uk-isps-dont-want-to-play-umpire-to-three-strikes-rule.html.

[32] "Higher Education Opportunity Act of 2008," in *P.L. 110-135* United States of America, 2008.

[33] A. Guess, "The Costs of Policing Campus Networks " in *Inside Higher Ed*, October 20, 2008 ed, 2008, Available online: http://www.insidehighered.com/news/2008/10/20/p2p.

[34] J. Goodchild, "Universities Cope with New Anti-Piracy Requirement," in *Network World*, 2009, Available online: http://www.networkworld.com/news/2009/061509-universities-cope-with-new-anti-piracy.html.

[35] M. Zhang, "Internet Traffic Classification " CAIDA: The Cooperative Association for Internet Data Analysis 2009.

[36] M. Geist, "ISP must come clean on `traffic shaping'," Toronto Star online 2007, Available online: http://www.thestar.com/comment/columnists/article/203408.

[37] A. Hussain, "The 20-minute broadband limit," Times Online, 2007, Available online: http://business.timesonline.co.uk/tol/business/money/broadband/article2982965.ece.

[38] R. Paul, "EFF study confirms Comcast's BitTorrent interference," Ars Technica, 2007, Available online: http://arstechnica.com/news.ars/post/20071128-eff-study-reveals-evidence-of-comcasts-bittorrent-interference.html.

[39] A. Fisher and M. Feyen, "Allot Communications NetEnforcer is First to Detect and Manage Encrypted BitTorrent Traffic," Allot Communications, Press Release, August, 2006.

[40] TorrentFreak.com, "More BitTorrent Users Go Anonymous," 2009, Available online: http://torrentfreak.com/more-bittorrent-users-go-anonymous-090622/.

[41] TorrentFreak.com, "Download Torrents Anonymously with TorrentPrivacy," 2008, Available online: http://torrentfreak.com/download-torrents-anonymously-with-torrentprivacy-080812/.

[42] J. Cheng, "The Pirate Bay to roll out secure €5 per month VPN service." vol. 2009: Ars Technica, 2009, Available online: http://arstechnica.com/telecom/news/2009/03/the-pirate-bay-to-roll-out-secure-vpn-service.ars.

[43] Digital Citizen Project at Illinois State University [DCP], "Summary of Project." vol. 2008, 2008, Available online: http://www.digitalcitizen.ilstu.edu/summary/.

[44] A. M. Mateus and J. M. Peha, "Dimensions of P2P and Digital Piracy in a University Campus," in *36th Telecommunications Policy Research Conference (TPRC)* Alexandria, VA, 2008, Available online: http://www.ece.cmu.edu/~peha/dimensions_of_piracy.pdf.

[45] Packeteer, "Applications, Protocols, and Services Classified by PacketWise 7.3," 2007, Available online: http://support.packeteer.com/documentation/packetguide/7.3/reference/services.htm.

[46] 105th Congress of the U.S.A., "H.R.2281 -- Digital Millennium Copyright Act of 1998," 1998, Available online: http://thomas.loc.gov/cgi-bin/query/z?c105:H.R.2281.ENR:.

[47] E. Bangeman, "Judge kills RIAA subpoena: making available not infringement," Ars Technica, 2008, Available online: http://arstechnica.com/news.ars/post/20080403-judge-kills-riaa-subpoena-making-available-not-infringement.html.

[48] TorrentFreak.com, "Top 10 Most Pirated Movies on BitTorrent," Weekly series, 2009, Available online: http://torrentfreak.com/top-10-most-pirated-movies-on-bittorrent-090727/.

[49] S. Smitelli, "Fun with YouTube's Audio Content ID System," 2009, Available online: http://www.csh.rit.edu/~parallax/.

[50] N. Christin, A. S. Weigend, and J. Chuang, "Content availability, pollution and poisoning in file sharing peer-to-peer networks," in *Proceedings of the 6th ACM conference on Electronic commerce* Vancouver, BC, Canada: ACM, 2005.

[51] R. Menta, "Top P2P Applications: 1.6 Million PCs Rank Them," MP3newswire.net, 2008, Available online: http://www.mp3newswire.net/stories/8002/p2p.html.