Quantifying Global Transfers of Copyrighted Content using BitTorrent

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

Abstract

This paper presents the most accurate empirical study to date to characterize and quantify the amount of content of various types that is transferred worldwide using BitTorrent, the dominant peer-to-peer (P2P) file sharing application. Using data we collected from the largest public BitTorrent tracker over 106 days between August 2010 and February 2011 and a new methodology, we find that for some content types, the number of copies transferred is an order of magnitude greater than the number sold through legal channels. For example, we estimate that 10.7 songs were transferred using BitTorrent for every song sold, 3.6 movies were transferred using BitTorrent for every legal sale or rental of a DVD or Blu-ray, and 227 movies were transferred using BitTorrent for every paid download. We also find that the vast majority of music and video content transferred using BitTorrent is copyrighted, as demonstrated both by the swarm metadata we observed, and the fact that only 0.55% of the transfers were of files indexed by websites that specialize in content that can be transferred legally. Thus, we conclude that BitTorrent transfers result in hundreds of millions of copyright violations worldwide per day, and that copyright holders fail to realize significant revenues as a result. Movies are the type of content most supplied and most transferred in BitTorrent (shared in 38.7% of swarms and accounting for 26.1% of transfers). Songs and software, despite being shared in small percentages of swarms (4.5% and 7.2% of swarms respectively), rank 2nd and 3rd in terms of transfers (with 20.4% and 16.8% of transfers respectively). This shows the limitations of past studies that estimated the economic impact of P2P by looking at which content is available rather than trying to measure the number of actual transfers. Surprisingly, most of the copies transferred using BitTorrent come from a small number of extremely popular titles; 37 song titles account for half of all songs transferred, and 117 movies account half of all movie transfers. Thus, for a global marketplace, the importance of the "long tail" of less popular content is smaller than we and others have observed in more localized studies. In general, the content that is popular in legal channels is also popular with BitTorrent, but we observe some important differences. For example, we find that content that is popular among teenagers is more likely to be disproportionally represented in BitTorrent transfers as compared to content that appeals to an older audience.

[1] Alexandre M. Mateus, Ph.D. in Engineering and Public Policy from Carnegie Mellon University, U.S.A., and Instituto Superior Técnico, Portugal, amfmateus@gmail.com.
[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 U.S. Government.

1 Introduction

The Internet is increasingly being used to obtain content, in particular audiovisual media (Cisco 2010). Peer-to-Peer (P2P) technology enables cost-effective distribution of content online by facilitating transfers of information between hosts (peers) that are part of a self-organizing overlay network supported by the IP network. At the same time, P2P raises significant issues in copyright protection and network management. P2P networks are used to transfer copyrighted content without permission from copyright holders, who claim that such activity has a heavy negative impact on their revenues (RIAA 2007). However, the actual dimension of copyright violations using P2P is far from being a settled matter, and there is still ongoing debate regarding how P2P affects the industries that produce and distribute copyrighted material (Oberholzer-Gee and Strumpf 2009).

This article assesses what and how much content is transferred using BitTorrent, currently the most popular file sharing P2P protocol in use, with the purpose of fulfilling three objectives. The first objective is to provide a reasonable empirically derived lower bound for the number of copies of copyrighted titles transferred using BitTorrent. This important figure helps to show the extent to which BitTorrent is used for copyright infringement.

The second objective is to break that lower bound down into categories depending on characteristics of content transferred. Thus, we can distinguish between transfers that would probably violate copyright law and those that probably would not, as well as estimate the number of copies transferred for distinct content types, such as songs, movies, and software. We further assess the extent of such transfers for each market segment by comparing the number of copies transferred illegally via P2P with the number purchased from a variety of legal outlets, including sale of physical goods (CDs, DVDs), downloads from legal sites, and theater ticket sales. For example, we find that for some content types, the number of copies transferred illegally via P2P exceeds the number of legal sales by an order of magnitude. Finally, we estimate the number of copies transferred by specific title, and differentiate the more popular titles from the less popular. Looking at the distribution of popularity of transferred media titles tells us the extent to which P2P users prefer content that is popular through legal outlets, and the extent to which they seek less popular titles that may not be widely available in legal outlets. These results can help copyright holders provide more compelling legal alternatives to P2P.

The third objective is to understand which content formats and technical characteristics of content (different methods of video digitalization, video resolutions and audio bit rates) users prefer. Such information can be useful to understand consumer demand, and inform those providing legal media outlets. It can also help predict how well enforcement technology is likely to perform, since current

techniques, like Deep Packet Inspection, are more effective at detecting some content types than others. The effectiveness of such technical mechanisms may influence policy decisions.

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

The remainder of this article is organized in three main sections. Section 2 presents the estimation methodology and data collection procedures. Section 3 presents our results, in which we estimate the volume of BitTorrent transfers and characterize what content is transferred using BitTorrent. The article concludes with a summary of findings and policy implications in section 4.

2 Methodology

In order to estimate the rate at which copies of various types of content are transferred using BitTorrent, we estimate the rate at which copies are transferred in each swarm, and then aggregate those estimates for swarms sharing each type of content. We consider the swarms being managed by the two largest public BitTorrent trackers: OpenBitTorrent and PublicBT. These trackers don't manage all existing BitTorrent swarms, but they were by far the largest public trackers at the time of monitoring, so it is likely that they account for a large share of all existing swarms. Thus, an estimation using these two trackers should yield a reasonable lower bound for the amount of content transferred using BitTorrent.

The transfer rate of each swarm is estimated using equation 1, where $Speed$ is the average transfer speed in bytes per unit of time achieved by each leecher in the swarm, L_{active} is the number of leechers actively downloading content in the swarm, and $Bytes$ is the number of bytes of content shared in the

swarm. In this analysis, we assume that all file transfers eventually complete successfully, even if it takes multiple BitTorrent sessions to do so. This assumption generally holds, because BitTorrent clients are designed to automatically resume incomplete transfers upon launch. However there are some cases when leechers abort transfers before obtaining the entire content and do not come back for the rest. We believe such cases are sufficiently unusual that this assumption will not greatly affect estimations.

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

The next sections describe the data collection and estimation processes by which we obtained each of the inputs in equation 1: section 2.1 describes the estimation of $Speed$, section 2.2 describes the estimation of L_{active}, and section 2.3 describes the estimation of $Bytes$.

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

In this section we estimate the average download rate achieved by a leecher in a swarm, which is one of the inputs to equation 1. This rate depends on three main factors: (i) the speed of the leecher's Internet connection, which imposes a ceiling on the transfer speed that the leecher can achieve, (ii) the number of peers connected to the swarm, which limits the number of peers from which the leecher can simultaneously download content, and (iii) the number of bytes shared in the swarm, since for swarms that share few bytes the download may not take long enough for the leecher to reach full speed.

Through experimentation, we determine how transfer rates vary with the number of seeders and leechers in a swarm and with the number of bytes shared in the swarm for different Internet connection technologies. We apply regression analysis to measurements of steady-state transfer speeds obtained using different technologies in swarms with different numbers of seeders and leechers. In the case of swarms sharing smaller amounts of content, we incorporate a scaling factor to account for the fact that leechers will not be able to reach steady-state download speed. The number of copies of content transferred in each swarm can then be determined parametrically using different scenarios for the breakdown of Internet connection technologies among the leechers in the swarm.

Estimation was performed using a data set containing transfer speeds achieved using different Internet connection technologies in swarms with different sizes. We used a set of 20 swarms sharing content that can be legally transferred using BitTorrent, with different numbers of seeders (ranging from 0 to 269) and leechers (ranging from 0 to 67). For each swarm, we collected data from three types of connection

technologies and five locations: from two Fiber/LAN connections in university campuses in the US and in Portugal, from two high-speed cable residential connections in the US and in Portugal and from a slower DSL residential connection in Portugal. We downloaded content for each swarm in each location every two hours over the course of a day and once per second recorded the number of seeders and leechers and download speed achieved in the session.

Our analysis shows that the steady-state download speed achieved in a swarm depends on the type of technology that the leecher is using to connect to the Internet, as figure 1 illustrates. Although there is not much difference between Fiber/LAN and high-speed cable modems in the US or Portugal, probably because they all have sufficiently high capacities that the Internet connection is rarely the bottleneck, the slower DSL connection greatly reduces average download speed.

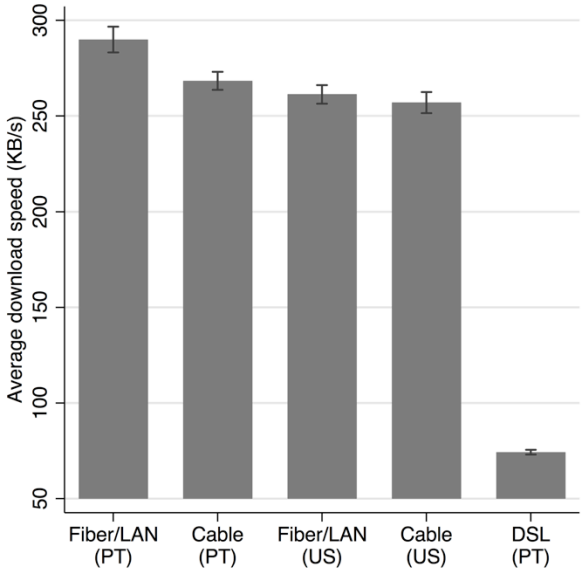


Figure 1. Average download speeds achieved using all monitored swarms for each location/technology monitored.

We also find that download speeds are higher in swarms with higher numbers of seeders and leechers, as shown by the correlations in table 1. Furthermore, we expect decreasing returns to scale in download speed as swarms get larger, yielding a (strictly) concave function, since BitTorrent clients typically define a ceiling for the number of peers from which they can download content at any moment, and that ceiling is independent of the size of the swarm.

Table 1. Correlation coefficients between transfer speeds and number of leechers and seeders in swarms and logarithms of number of seeders and leechers in swarms, for different Internet connection technologies monitored.

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

We use regression analysis to estimate the parameters in equation 2, which shows how steady-state download speed varies with the number of seeders and leechers in the swarm. We perform separate estimations for each connection technology. These parameters explain between 54% and 65% of the observed variance in transfer speeds, which is sufficient for estimating of number of copies transferred in the swarms.

$$Steady\_State\_Speed(s,l) = \beta_s \log(s) + \beta_l \log(l) \quad (2)$$

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

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

However, leechers may not always be able to reach the above steady-state speeds in a transfer. Leechers stack up download connections up to a maximum threshold (which is user-configurable), but there are two situations in which a leecher may not be able to use all its download connections: when the swarm is small and there are not enough peers to download from, and when the content shared in the swarm is broken down in a number of parts smaller than the number of download slots (which is typical for small files). The impact of the number of peers is already reflected in the regression analysis used to derive steady-state speed estimates. To account for file size, we use equation 3, which scales down the steady-state transfer speed linearly with p (the number of parts in which the swarm's content is divided) when p is smaller than 50 (the typical default number of download slots in today's top BitTorrent clients). This method of estimating transfer speed is likely to be more accurate for swarms sharing larger files, which is the case for most files for which copyrights are enforced. (For example, the sizes of movies, TV shows and songs are typically over several megabytes).

$$Speed(s,l,p) = Steady\_State\_Speed(s,l) \times \frac{\min(p,50)}{50} \quad (3)$$

Since Internet users use a variety of technologies, we calculate the number of copies transferred in a swarm (equation 1) using five different scenarios for the breakdown of the swarm's leecher population by Internet connection technologies, as shown in table 3. The "All DSL" and "All Fiber" scenarios assume all the leechers in the swarm have DSL and Fiber/LAN connections, respectively. These scenarios result in the lowest and highest transfer rates respectively. The "Portugal", "USA" and "OECD" scenarios assume that the breakdown of connection technologies is identical to the breakdown of fixed broadband connections that exists in Portugal, in the USA and in all OECD countries respectively[4]. These three scenarios result in intermediate values for the estimates of number of copies transferred in each swarm. The scenario that is likely to yield the most accurate estimates is the "OECD" scenario, given that it represents the breakdown of connection technologies of a wide range of countries with a high penetration of broadband Internet.

Table 3. Scenarios used in estimation of the average transfer speed achieved by a leecher in a swarm. In each scenario, the swarm is assumed to have a breakdown of leechers for each connection technology according to the percentages indicated in each corresponding table cell.

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

2.2 Estimating the number of leechers downloading content in each swarm

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

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

It is possible for a leecher to be waiting for the desired content to become available, e.g. if there are not enough peers sharing the content to satisfy the demand from all the leechers in the swarm, but this is unusual. BitTorrent's *Rarest First* scheduling algorithm prevents this from happening in swarms that have passed the initial ramp-up phase when only the original seeder holds all the pieces of the content (Legout, Urvoy-Keller, and Michiardi 2006). Since this ramp-up state is transient and its duration is typically much smaller than the lifespan of the swarm, for the purpose of calculating how many leechers are actively downloading content, we assume that the number of leechers waiting for content is negligible.

In contrast, the other situation in which a leecher is counted among the swarm but is not downloading, i.e. when the leecher has failed, is too common to ignore. We define failed leechers as those who disconnect from swarms without informing the tracker. This can happen for several reasons, such as users quitting their BitTorrent clients without stopping active downloads, client application crashes, or loss of Internet connectivity. In such cases the tracker takes some time to notice that the leecher has departed[5], and during that period it still accounts for the leecher in the reported counts. It is not possible to determine which leechers are active by communicating with them directly, even if one retrieves all leecher IP addresses, because many attempts to initiate communications with active leechers would fail due to problems with network address translation (NATs). Thus, we instead take failed leechers into account and estimate the number of leechers actively downloading content in each swarm, L_{active}, using equation 4, where L_{all} is mean value of the total number of leechers reported by the tracker and L_{failed} is the mean value of the number of leechers that have failed but are still reported.

$$L_{active} = L_{all} - L_{failed} \qquad (4)$$

We obtain L_{all} for each swarm managed by the PublicBT and OpenBitTorrent[6] trackers using the BitTorrent tracker scraping mechanism[7]. At specific time intervals, we requested the list of swarms managed by the tracker and the counts of seeders and leechers for each of those swarms. Such data was collected from OpenBitTorrent at 1-hour and 2-hour intervals between August 6 and September 23,

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

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

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

2010, and from PublicBT at 10-minute intervals between November 23, 2010, and February 4, 2011. We switched collection from OpenBitTorrent to PublicBT because OpenBitTorrent phased out its support for the tracker protocol over HTTP in favor of UDP. Since some BitTorrent clients do not support tracker protocol over UDP yet, this made the popularity of OpenBitTorrent decline, and PublicBT took its place as the most popular public BitTorrent tracker[8]. We detected an average of 2.6 million swarms being managed by OpenBitTorrent and an average of 2.7 million swarms being managed by PublicBT at any moment. Overall, we detected close to10 million swarms over the entire data collection.

The main challenge in estimating the number of leechers actively downloading content from each swarm is to estimate L_{failed}, the number of leechers reported by the tracker that have failed. We perform such estimation using a novel method that takes advantage of the fact that the actual removal of failed leechers from the tracker lists happens in bursts at regular time intervals. By observing short-interval variations in the number of leechers reported by the tracker it is possible to estimate the percentage of leechers that already failed but that are still accounted in the tracker counts, which we observed to be relatively constant for different swarms and at different monitoring points. While we detail the estimation in terms of leechers, the same process happens for seeders reported by the tracker, and we present results for seeders obtained using a similar estimation process.

BitTorrent trackers use the following timeout mechanism to detect peers that have failed. Peers contact the tracker at least once per pre-defined time interval (t_a, the announce interval). To remove failed peers, the tracker performs a cleanup at regular intervals of t_c wherein a peer is considered to have failed and is removed from the respective list of leechers or seeders if that peer has not communicated with the tracker for a predefined timeout period of t_{to} ($t_{to} > t_a$).

We assume that new peers arrive according to a Poisson process. Let λ be the average leecher arrival rate and f represent the probability that a leecher will fail. Let t_f be the average time that a leecher remains in the tracker lists after failure. Assuming these variables are independent, the average number of leechers that have failed but are still accounted by the tracker at any moment, L_{failed}, is given by $L_{failed} = \lambda \cdot f \cdot t_f$. Under these same assumptions, the average number of leechers removed in each cleanup process, L_r, is given by $L_r = \lambda \cdot f \cdot t_c$. Solving these two equations yields the average number of failed leechers given in equation 5, and then equation 5. Next, we detail the estimation of the three inputs to equation 6: L_r/L_{all}, t_c, and t_f.

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

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

$$\frac{L_{failed}}{L_{all}} = \frac{L_r}{L_{all}} \cdot \frac{t_f}{t_c} \quad (6)$$

2.2.1 Estimating t_c and L_r/L_{all}

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

For a diverse set of 500 swarms[9], we queried the PublicBT tracker for the number of leechers and number of seeders for each swarm at time intervals less than 1 second apart[10] during a period of about 24 hours. This yielded a data set with about 18 million observations, at a median rate of one observation every 0.7 seconds for each swarm. The variation over time in the number of leechers reported by the tracker for each swarm indicates that the tracker removes peers that failed every 60 seconds, i.e. $t_c = 60\ s$. Figures 2.a and 2.b show that $t_c = 60\ s$ by portraying the number of leechers in one swarm during an interval of 10 minutes, and the distribution of number of seconds between decreases in the number of leechers reported by the tracker[11] for all monitored swarms.

We calculate the average number of leechers (L_{all}) by averaging across all samples, and we calculate the average number of leechers removed (L_r) by averaging the decreases in number of leechers every 60 seconds across all observed cleanups. The ratio L_r/L_{all} we seek to estimate is fairly constant across swarms of very different sizes, as shown by the narrow confidence interval in table 4, which presents L_r/L_{all} averaged over the 500 swarms for which we collected data.

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

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

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

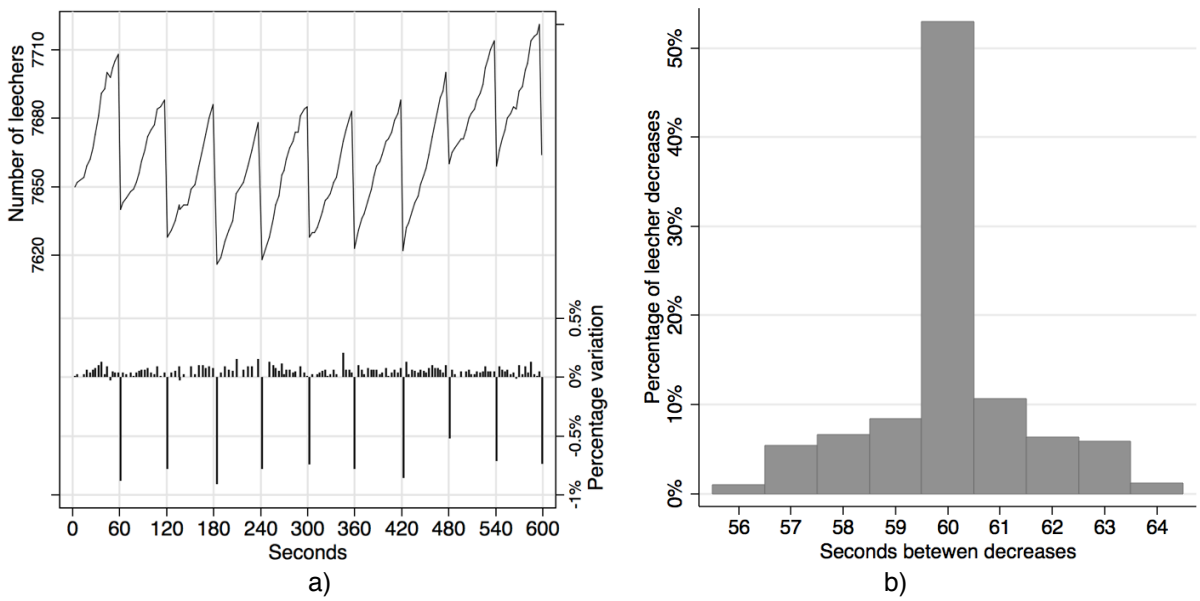


Figure 2. Dynamics of removal of failed peers by the tracker. a) Snapshot of the number leechers reported by the tracker for a swarm. Lines on the top of the graph represent number of peers reported by the tracker. Bars on the bottom represent the percentage variation in reported number of leechers between observations. b) PDF of number of seconds between decreases in number of leechers reported by the tracker, detail of seconds 56 to 64.

Table 4. Ratio of the average number of leechers removed in cleanup processes to the average number of leechers reported by the tracker (L_r/L_{all}), averaged across the 500 monitored swarms (95% CI in parenthesis).

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

2.2.2 Estimating t_f

In this section we estimate t_f, the mean time between the failure of a leecher and its removal from the tracker lists, the final input needed for equation 6. To do so we use a probability model that incorporates information about the tracker's timeout mechanism and that assumes that the time until a leecher fails is distributed exponentially.

The two main parameters that influence the tracker's timeout mechanism are the tracker's timeout period, t_{to}, which is the time between a peer's last contact and it's removal from the tracker lists; and the peer's announce time, t_a, which is the maximum time allowed between successive contacts to the tracker from each peer. Both parameters can be estimated by observing tracker behavior. The announce time, t_a, is set by the tracker, and is communicated to the leecher in the response to every interaction with the tracker. To estimate t_a, we forged announce requests to PublicBT and collected the responses. The resulting t_a were uniformly distributed between 1620 and 1980 seconds (27 to 33 minutes). To estimate the timeout time, t_{to}, we created a swarm sharing a file with random bytes and registered that swarm in

PublicBT. We then consecutively collected seeder and leecher counts at short time intervals while sending forged announces for new peers in that swarm, which we would then let timeout. We collected the time difference between the last announce sent by each peer and the moment that peer stopped being counted by the tracker, which yielded an estimate for t_{to} of 45 minutes[12].

Let F be the distribution of time until a leecher fails. We assume that F is exponential with parameter γ, the average leecher failure rate, and the probability density function, $f(x)$, in equation 7.

$$f(x) = \gamma e^{-\gamma x}, \qquad x \geq 0 \tag{7}$$

Let G be the distribution of time between the failure of a leecher and its removal from the tracker list, with probability density function $g(x)$. Clearly, $g(x) = 0$ in its entire domain, except when $x \in [t_{to} - t_a, t_{to}]$, or since $t_{to} = 45$, except when $x \in [45 - t_a, 45]$. For a particular value of the announce time, t_a, the density function, $g(x|t_a)$, is the one in equation 8.

$$g(x|t_a) = \begin{cases} \dfrac{f(x-(45-t_a))}{\int_0^{t_a} f(x)\,dx}, & 45 - t_a < x < 45 \\ 0, & \text{otherwise} \end{cases} \tag{8}$$

As we determined empirically, t_a is a uniformly distributed random variable that ranges from 27 to 33 minutes. Let $h(x)$ be the density function of t_a, defined according to equation 9 if we consider minutes as the time unit.

$$h(x) = \begin{cases} \dfrac{1}{6}, & 27 \leq x \leq 33 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

Assuming that the time until a failed leecher is removed is independent from the announce time, i.e., that G and t_a are independent, then $g(x)$ can be defined in terms of $g(x|t)$ and $h(t)$ according to equation 10.

$$\begin{aligned} g(x) &= \int_{-\infty}^{+\infty} g(x|t) \cdot h(t)\,dt = \\ &= \begin{cases} \dfrac{1}{6} e^{(45-x)\gamma}\left(\ln(1-e^{33\gamma}) - \ln(1-e^{27\gamma}) - 6\gamma\right), & 18 < x < 45 \\ -\dfrac{1}{6} e^{(45-x)\gamma}\left(\ln\left(1-e^{(45-x)\gamma}\right) - \ln(1-e^{33\gamma}) - 12\gamma + x\gamma\right), & 12 < x < 18 \\ 0, & \text{otherwise} \end{cases} \end{aligned} \tag{10}$$

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

Given the above, the mean time that a leecher remains in the tracker lists after failure, which is the parameter we want to estimate, t_f, is simply the mean of the G distribution and can be calculated using equation 11.

$$t_f = \int_{-\infty}^{+\infty} x \cdot g(x)\, dx \qquad (11)$$

In order to calculate t_f we need to know the value of the parameter γ, the average leecher failure rate. We calculate γ as the average of the ratio between the number of peers removed at cleanup and the number of peers in the swarm right before cleanup, calculated over all observed cleanup times for the 500 swarms we monitored at short time intervals (see section 2.2.1). Estimates for this average are presented in the first column of table 5. Using these estimates as the average leecher failure rate (in fraction of leechers per minute), we obtain the estimates for t_f presented in table 5. The estimates indicate that both leechers and seeders will remain in the tracker lists on average about 29 minutes and 13 seconds after they have failed.

Table 5. Estimates for the average leecher/seeder failure rate and for the mean time that leechers/seeders remain in the respective tracker list after failure.

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

2.2.3 Putting it all together

We estimate the fraction of leechers reported by the tracker that have already failed using equation 6. The three inputs to that equation, as well as the estimated fraction of failed leechers (and seeders) are presented in table 6. We find that roughly 30% of leechers (and seeders) that the tracker reports as being active in the swarm at any given moment have already failed, and that this percentage is fairly constant across all swarm sizes. It is therefore important, for accuracy of our estimates of the number of copies of content transferred by the leechers in a swarm, to take failed peers into account.

Table 6. Estimates for L_{failed}/L_{all}, the ratio of the average number of failed leechers to the average number of leechers reported by the tracker.

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

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

This section describes how we estimate the number of bytes of content shared in each swarm, which is the final term in the calculation of the number of copies transferred per swarm in equation 1. The section also describes how swarms are categorized by type of content, which allows them to be aggregated in order to calculate how many copies of content of each type are transferred.

Both number of bytes and the information necessary for categorization of swarms are obtained from each swarm's ".torrent" file. For trackers, each BitTorrent swarm is identified by one info-hash, which is a unique digest of the content shared in the swarm. Users of BitTorrent find the content they wish by searching ".torrent" files, which map a qualitative description of the content being shared in a swarm to the swarm's info-hash.

We obtained ".torrent" files for the swarms whose information we collected from PublicBT by searching multiple torrent index sites[13] using the swarms' info-hashes. Obtained ".torrent" files were parsed to extract the relevant information, which includes the title of the torrent and the total number of bytes shared.

Swarms for which it was possible to obtain ".torrent" files are categorized by type and other characteristics of content in a second stage of processing. In this stage, we parsed the title of the torrent to extract content characteristics such as the actual title of the content (the title of the movie, for instance) and keywords typically included in torrent titles that indicate technical characteristics of the content, such as the type of content (song, movie, TV show, software, adult content, book, etc.), encoding (mp3, aac, divx, ogg, mkv, etc.) or quality (128kbps, 256kbps, 480p, 720p, 1080p, etc.).

Since we do not actually download content, our estimates include transfers from swarms whose content matches matches the metadata and transfers from swarms whose content does not match the metadata. Swarms containing the latter do exist, sometimes as a means of frustrating users who are trying to obtain

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

copyrighted media illegally, and sometimes to spread malware. Today's most popular index websites sport rating systems that allow users to quickly identify swarms in which fake content is shared, so the popularity of these swarms is typically short-lived. Thus, the fraction of transferred copies containing fake content is likely to be small. Moreover, we also count each transfer even if a user transfers the same content more than once. For example, a user may download the low-resolution version of a movie initially, and then later download a higher-quality version when it becomes available. Each of these transfers may constitute a copyright violation, but they clearly do not all represent lost sales.

3 Results

This section presents our estimates of the amount of content made available and transferred using BitTorrent, and characterizes various aspects of that content. We start by estimating the amount of content made available in BitTorrent broken down by types of media as a way of assessing content supply in BitTorrent (section 3.1). Next we estimate how many copies of content are effectively transferred per day, a figure that had not been well characterized before, and that is relevant when considering BitTorrent from a copyright infringement perspective (section 3.2). In section 3.3 we estimate the amount of content transferred in BitTorrent that would not result in copyright violations, and in section 3.4 we compare our estimates of number of copies of copyrighted content transferred using BitTorrent to legal sales figures for music and movies to put the amount of copyright infringement in BitTorrent into perspective. In section 3.5 we discuss how much revenue is not realized due to illegal transfers of music and movies and in section 3.6 we look at the relative distribution of popularity of content transferred using BitTorrent as a way to understand whether users seek popular content or less mainstream media. Finally, section 3.7 examines what characteristics of content BitTorrent users prefer, which can be useful for those seeking to provide legal alternatives to P2P and can also influence the performance of technology for detection of transfers of copyrighted content.

3.1 Content Supplied in BitTorrent

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

In the data we collected from the largest public BitTorrent trackers during 106[14] days between August 2010 and February 2011, we found an average of 2.6 million swarms offering content at any moment, which added up to a total of close to 10 million swarms offering content at some point in the period. These are lower bounds on the number of bundles of content supplied in BitTorrent.

To understand which types of media are most supplied in BitTorrent we aggregate swarms by the type of content shared. We could gather ".torrent" files for 74% of detected swarms, and could infer the type of media for 52% of those, which corresponds to 39% of all detected swarms. Figure 3 presents the breakdown of the swarms for which we could infer the type of media. It shows that movies have the highest supply in BitTorrent (38.7% of swarms), followed by music albums, TV show episodes and then by software. When compared to previous estimates of supply of content in BitTorrent (Envisional 2011), we find similar percentages of movie and TV show swarms (previous estimates report 32% of Films and 13% of Television), but we find much lower percentages of adult content swarms (previous estimates report 36% of swarms sharing Pornography) and much higher percentages of music and software swarms (previous estimates report 3% of Music and 4% of Software swarms). Nevertheless, our results qualitatively confirm previous estimates that indicated video as the most supplied type of content in BitTorrent.

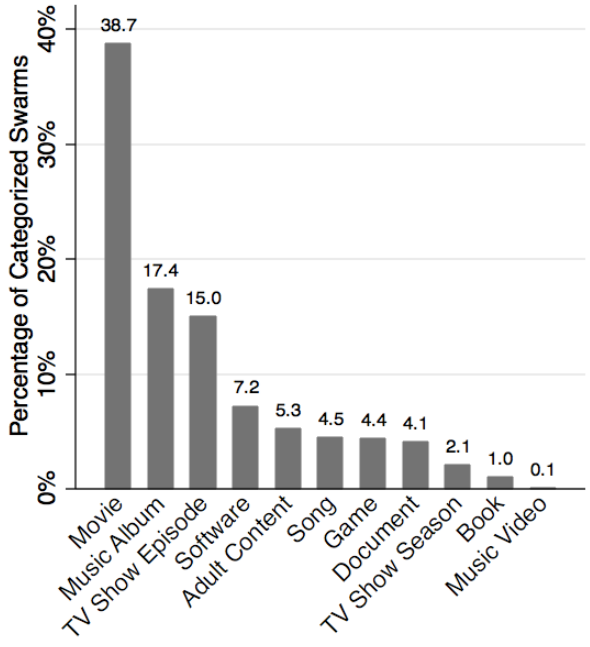


Figure 3. Breakdown of supply of content in BitTorrent by percentage of swarms sharing content of different media types.

[14] Although we collected 115 days of monitoring data, the last 9 days leading to when OpenBitTorrent migrated from TCP to UDP connections are excluded from analysis. The OpenBitTorrent server did not always respond to data requests during this 9-day period, so the data gathered then may be less accurate.

3.2 Content Transferred using BitTorrent

In this section we estimate the number of copies of content transferred using BitTorrent and break that number down by the type of media transferred. To calculate the overall number of copies of content transferred in each swarm during each day, we estimate the instantaneous number of copies transferred per unit of time at each monitoring point according to the methodology described in section 2. We then interpolate such figures for each swarm over the monitoring points during the day.

Estimates of the average number of copies of content transferred per day in all swarms for which we could find ".torrent" files (74% of all monitored swarms) are presented in figure 4 for the different scenarios of leecher connection technology[15]. We consider that the OECD scenario provides the most accurate estimate because it represents the breakdown of connection technologies of a wide range of countries with a high penetration of broadband Internet, which is likely more representative of the breakdown of connection technologies for BitTorrent users worldwide than the other scenarios. We will use OECD estimates in the remainder of this article. By that account, the swarms with torrent information that we monitored, and that contained more than 1024 bytes of content, transferred over 380 million copies of content on average per day.

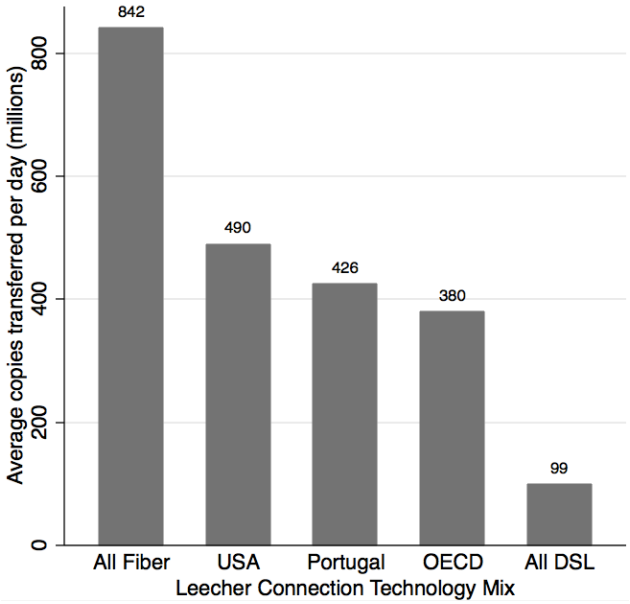


Figure 4. Estimates of overall number of copies of content transferred per day by all monitored swarms with torrent information that shared amounts of content greater than 1024 bytes using the different scenarios of leecher connection technology mixes.

[15] The "All Fiber" and "All DSL" scenarios and not realistic because they imply respectively that all leechers have a fast fiber connection or a slow DSL connection. They serve as rough boundaries for our estimates. Intermediate scenarios, which represent the breakdown of Internet connections in actual geographical areas, yield realistic estimates that are much closer together.

We break down the number of transferred copies of content by the media type present in each swarm (for the 39% of all detected swarms for which we could find the ".torrent" file and could infer the type of media). Such breakdowns are presented in figure 5, which shows that movies are the type of media with more transferred copies, accounting for over one fourth of all transferred copies, followed by individual songs and software.

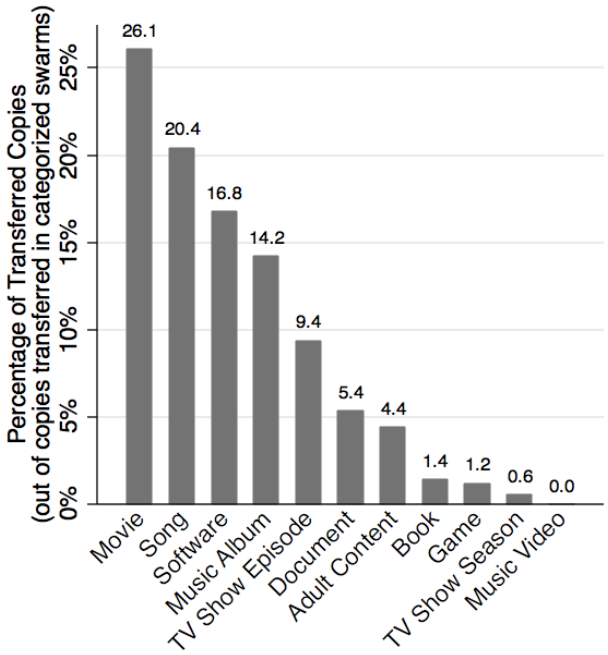


Figure 5. Breakdown of percentage of copies transferred using BitTorrent by type of media.

A comparison of the breakdowns of content supplied in BitTorrent (figure 3 in the previous section) versus breakdown of actual number of transferred copies (figure 5 above) shows the discrepancy between estimating BitTorrent activity based on supply versus estimating the actual number of transferred copies. Movies account for close to 40% of all swarms supplying content, but the percentage of transferred copies reaches only 26.1%. Music albums and TV show episodes also have higher shares of swarms supplying content than of actual transferred copies[16]. On the other hand, individual songs and software, which are only respectively the sixth and fourth most supplied types of content, are the second and third most downloaded types in BitTorrent, coming close behind movies at 20.4% and 16.8% of all transferred copies. This difference demonstrates an important limitation in previous studies (Envisional 2011; Layton and Watters 2010) that estimated BitTorrent activity by looking only at breakdown of swarms or number of

[16] The difference between supply and actual transferred copies comes in part because of the difference in number of bytes of content shared in each swarm. Video swarms we observed contained on average about 8 times more bytes than software swarms, meaning that for similar number of peers connected to swarms, copies of software are transferred on average 8 times faster than copies of video.

peers connected to each swarm. Such are estimates of supply of content in BitTorrent, which may be useful for some purposes. However, as we demonstrated, estimates of supply are not representative of the number of copies transferred, and as such are not informative for purposes of estimating the types of content whose copyright is most often violated in BitTorrent or the types of content for which there is greater economic impact from illegal transfers.

3.3 Content that can be Legally Transferred using BitTorrent

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

The index websites that we searched publicly declare to actively filter out content that cannot be legally transferred using BitTorrent. Mininova.org[17], legittorrents.com and youtorrent.com are general-purpose index websites that filter out copyrighted content, Linuxtracker.com specializes in indexing torrents for swarms containing the Linux OS, and Clearbits.com specializes in hosting and distribution of open licensed media. While there are likely more swarms sharing content that can be legally transferred using BitTorrent than those found in the above websites, looking at content indexed by these websites shows us how much BitTorrent activity comes from transfers of content actively promoted as legal.

Out of the close to 10 million swarms detected in our monitoring, we found 13,231 swarms whose torrents were indexed by the websites mentioned above. As table 7 shows, such swarms correspond to 0.16% of all detected swarms and the number of copies of content transferred represents 0.55% of overall transferred copies of content. Hence, despite the effort from these indexes of legal content to promote legal transfers in BitTorrent, the number of transferred copies from swarms that they index is close to insignificant when compared to transferred copies of titles indexed by other general-purpose indexes, for which the majority of indexed titles is likely content whose BitTorrent transfers are unlawful.

[17] Mininova.org was in the past one the largest BitTorrent index websites, hosting torrents for copyrighted content as well as content that could legally be transferred using BitTorrent. In late 2009, after a court order, Mininova started to actively filter out torrents "if there is reasonable doubt that the actual content contains copyrighted works" (http://blog.mininova.org/articles/2010/12/10/brein-mininova-settlement-reached-lawsuit-ended/)

Table 7. Percentage of swarms and percentage of transferred copies found in indexes specializing in legal content (sharing content that can be legally transferred using BitTorrent)..

	Content found in indexes that specialize in legal content
Percentage of swarms	0.16%
Percentage of transferred copies	0.55%

Figure 6 shows a breakdown of the types of content available in indexes of legal content, and the number of copies transferred. These websites make an effort to promote legal audio and video, which are also the types of content that are most downloaded. This is similar to what we observe when considering all monitored swarms, where a significant share of content is made available illegally, and where audio and video are among the types of content with the greatest supply (figure 3) and also the greatest number of transferred copies (figure 5). Besides audio and video, the third most transferred type of legal content is documents, for which the percentage of transferred copies is disproportionately high when compared to the supply of legal documents, or when compared to transfers of documents from all monitored swarms (as shown in figure 5).

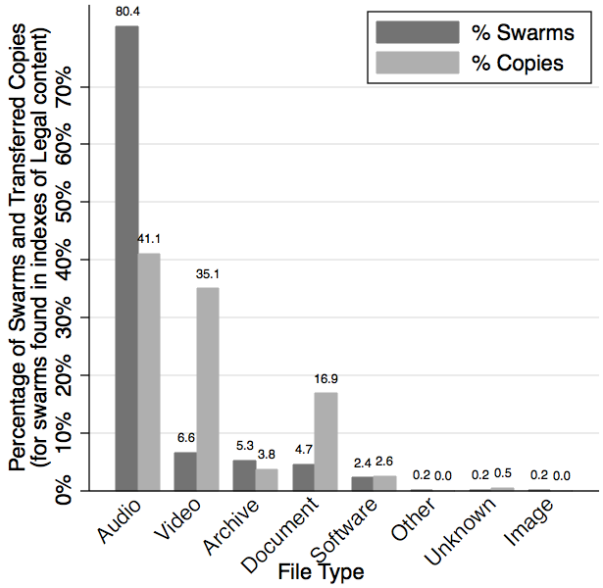


Figure 6. Breakdown by type of file of supply and number of copies transferred from swarms detected sharing content that can be legally transferred using BitTorrent.

3.4 Comparison of Transfers of Copyrighted Content to Legal Sales

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

compare worldwide legal sales for each title in the top 10 bestselling songs, albums and movies in theatres worldwide and DVDs in the U.S. to copies transferred in the swarms sharing each of those titles in BitTorrent.

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

In table 8 we compare the number of copies transferred per day using BitTorrent to daily worldwide averages for 2010 of movie theatre admissions, DVD and Blu-ray sales and rentals, online movie sales and rentals, and sales of songs and music albums. We obtained figures for the worldwide movie market from IHS Screen Digest[18]. For music, the IFPI reports that the digital music market had a trade value of $4.6 billion and represented 29% of the industry's revenue in 2010 (Moore 2011), which means that the revenue of the "physical" part of the market was about $11.3 billion. Assuming all revenue comes from music sales, that a digital song costs on average $1.2, a digital album costs $10 and a physical album costs $15[19], and that albums contain on average 10 songs, we estimate sales of 3.3 million albums per day (physical + digital), which correspond to 33.4 million songs per day (physical + digital).

Table 8. Comparison between estimated daily number of copies of content transferred using BitTorrent for the swarms whose content could be categorized and sales figures for equivalent content types.

	BitTorrent daily transferred copies (M = millions)	Worldwide Market	
		Daily Transactions (M = millions)	Transfers to Sales Ratio
BitTorrent Movies vs. legal movie transactions	**57 M** movies	**35 M** sales + rentals (all channels)	1.6
BitTorrent Movies vs. Box Office		**18.7 M** movie theatre admissions	3.1
BitTorrent Movies vs. DVD and Blu-ray		**16.1 M** DVD + Blu-ray sales + rentals	3.6
BitTorrent Movies vs. Online transactions		**0.25 M** Online movie sales + rentals	227
BitTorrent Albums vs. legal music album sales	**31 M** album bundles	**3.3 M** digital + physical albums	9.5
BitTorrent Songs vs. legal song sales	**358 M** single + bundled songs	**33.4 M** digital + physical songs	10.7

[18] IHS Screen Digest (http://screendigest.com) is a media-focused research, publishing and consulting company that collects data on worldwide movie transactions in various distribution channels.

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

For movies as well as for music, the number of copies transferred using BitTorrent is much greater than legal sales, as shown in table 8. In the case of music, assuming that each music album transferred using BitTorrent contains on average 10 songs, the overall number of songs transferred using BitTorrent (either individually or as part of an album) was 10.7 times greater than estimated worldwide sales of songs (either digital or as part of physical media). In the movie market, the number of BitTorrent transfers of movies is 3.1 times greater than the number of movie theatre admissions, 3.6 times greater than the number of DVD and Blu-ray discs sold and rented worldwide, and 227 times greater than the online market for movies. Even if all of the above are included, there are 1.6 BitTorrent transfers for each of these transactions. These comparisons are meant to show the magnitude of the illegal BitTorrent "market" versus the various channels of the legal market, but not to imply any direct substitution. It is impossible to determine whether a given illegal download using BitTorrent substitutes for a legal transaction, and if so, in which legal distribution channel. For example, movies are distributed through different channels sequentially over time, typically first in movie theatres, then DVD/Blu-ray, and then eventually on TV. We cannot tell which of these legal channels would be affected by illegal downloads.

Focusing on the most popular titles in terms of legal sales during the period we monitored, we are able to establish a title to title comparison of worldwide sales to copies transferred using BitTorrent for the worldwide top 10 selling music singles[20], music albums[21], and top 10 box-office grossing movies[22], and for the U.S. top 10 selling DVDs[23]. Figures for BitTorrent transfers for each song, album or movie were obtained by adding up the number of transferred copies in all swarms whose torrent name or file names match that title. Figures for sales were obtained by merging weekly sales data for each type of media in the same weeks for which we collected BitTorrent data. This comparison is presented in tables 9 through 11. The figures do not imply a direct competition between BitTorrent and sales of the particular type of media for each title; they simply show that BitTorrent transfers greatly exceed legal sales for the vast majority of the top-10 titles in each of the media types considered. The tables show that sales ranks are typically higher than BitTorrent ranks for the top 10 sales titles. This means that choosing the top sales titles to compare to BitTorrent transfers will yield smaller transfers to sales ratios than those that would be obtained if the comparison were done for the top transferred titles.

Considering music titles, both singles in table 9 and albums in table 10, we can see that BitTorrent transfers exceed sales by over an order of magnitude for most titles. In the particular case of music albums, we observe a large variation of the transfers to sales ratio between titles. One possible

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

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

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

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

explanation for this variation comes from the nature of the media transferred and the demographics it typically appeals to. Clearly, the transfers to sales ratio is greater for music albums of pop artists (e.g., Lady Gaga, Rihanna, Justin Bieber) whose music caters to a teenager and young adult audience that is typically Internet-savvy as well. In comparison, albums that perform well in sales but not so well in BitTorrent are those that typically cater to an older audience (e.g., Susan Boyle), who may not know how to transfer content from P2P, may not be willing to do it because they know it is illegal, or may have higher willingness to pay for legal content. This hypothesis is also corroborated by the figures comparing DVD sales to movie transfers in table 11. In this case, titles that perform worse in BitTorrent when compared to legal sales are mostly content destined for children (e.g, “Tinker Bell”, “Toy story 3”), whose parents likely belong to the older audience that prefers to purchase content instead of transferring it from P2P.

Table 9. Comparison of worldwide sales of music singles to number of copies transferred using BitTorrent for the top 10 most sold music singles during the monitoring period (sales and transfers in thousands).

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

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

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

Relevant implications for business and enforcement can be drawn if the difference in ratios of BitTorrent transfers to sales is indeed the result of different demographics having different propensity to transfer content from BitTorrent. It may be possible to predict which titles in copyright holders’ catalogs are more likely targets of illegal sharing, and thus estimate the extent to which sales of those titles will be affected by online copyright violations. Titles with higher transfers to sales ratios are those likely to appeal to teenagers and young adults, an important segment of the population whose members are typically avid consumers of media, but who, at the same time, may have less willingness to pay or disposable income

to purchase such media. This segment of the population has in P2P a free, yet illegal, alternative, which they seem to be taking advantage of. Copyright holders can use this information to try to drive those consumers away from P2P, either by deploying selective enforcement focusing on the titles that typically appeal to those demographics, or by further investigating which factors drive such consumers away from purchasing content in order to devise more compelling legal alternatives.

Table 11. Comparison of U.S. sales to number of copies transferred using BitTorrent for the top 10 most sold DVDs during the monitoring period (sales and transfers in thousands).

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

Table 12 shows ratios of BitTorrent transfers to worldwide box-office ticket sales for the top 10 movies in terms of theatre admittance worldwide during our monitoring period. These ratios vary widely. One factor that seems to be of importance is DVD release date: movies whose DVD was released in our monitoring period have higher average transfers to sales ratios. One possible explanation is that BitTorrent users may obtain the higher-quality DVD-rip copies around the time of the DVD release, maybe instead of obtaining the lower-quality cam rips[24] that are available between the theatrical release and the DVD release, or perhaps to substitute for a lower-quality cam rip they had obtained previously. Another possible explanation is that the marketing boost that happens around the DVD release may work for BitTorrent as well as it works for legal sales.

Table 12. Comparison of estimated worldwide box-office ticket sales to number of copies transferred using BitTorrent for the top 10 box-office movies during the monitoring period (sales and transfers in thousands).

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

[24] See Appendix A.

3.5 Revenue not Realized due to Illegal Transfers using BitTorrent

This section discusses the revenue that is not realized by copyright holders due to transfers of music and movies using BitTorrent, which is among the more important and more controversial unanswered questions regarding the impact of unauthorized transfers of copyrighted content (GAO 2010). Existing estimates of revenue not realized due to competition with P2P typically assume that the impact of illegal transfers is linear with the number of illegal copies transferred from the various content categories, weighted by the relative retail price of content in these categories (e.g., Siwek 2007; Tera Consultants 2010). Thus, if the analysts behind these estimates are correct, the revenues not realized by copyright holders would equal the product of this weighted average and some constant multiplier. Even assuming this hypothesis of linearity is correct, it is very hard to determine the value of the multiplier. Existing literature that attempts to estimate an average impact of illegal transfers presents contradictory results. Some studies argue that P2P reduces revenues due to substitution effects, and estimate multipliers for the case of music downloads that fall mostly within the range of 10% to 30% (Peitz and Waelbroeck 2004; Zentner 2006; Hong 2007; Rob and Waldfogel 2006; Liebowitz 2008). Other studies argue that P2P has a positive impact on sales due to marketing effects (Gopal and Bhattacharjee 2006; Andersen and Frenz 2008). Some studies argue that the impact of illegal transfers is probably different for the most popular titles versus less popular titles (Blackburn 2004), as these substitution and marketing effects depend on a given title's popularity. If this is the case, it would be more appropriate to use different multipliers for the more popular and less popular content. Furthermore, the impact of illegal transfers on sales is probably different for different user groups (e.g. in rich countries versus poor), and possibly different by content type (e.g. music versus movies).

Despite this uncertainty, there have been estimates of revenue not realized due to competition with P2P, which assume a linear impact of illegal transfers and fixed multiplier values. Examples of such estimates are studies by Siwek (2007), which deals with revenue not realized by the music industry and is widely cited by the industry, and by Tera Consultants (2010), which deals with impact on the music, movie and TV industries. Such studies have used multiplier values of 10% and 20% in the case of P2P transfers of music (i.e., they assumed that each copy transferred using P2P resulted in lost revenue equivalent to 10% or 20% of the average retail price of a copy of the same type of content), and 5% and 10% in the case of movies.

We don't know the right value for the multipliers for different types of content. For that reason, figure 7 shows revenues not realized for movies and music in 2010 as a function of these multipliers, using the number of copies of content of each type transferred that we calculated in section 3.4. In underlying calculations, we assume average music prices of $1.2 per song and $10 per album. As discussed in

section 3.4, it is impossible to determine which legal channel is affected by P2P, so we consider two possibilities for movies. In one case, we assume P2P competes only with box office sales, so movie theater revenues are affected, and the average price per title is $7.89 (which was the average movie ticket price in the USA in 2010). In the other case, we assume P2P competes only with purchases, so retailers rather than theaters are affected, and the average price per title is $15 (the average price of DVDs[25]). Reality is probably somewhere between these two cases. Horizontal lines in the figure represent the 2010 estimated revenues of the music industry and of the movie industry (in the latter case broken down by revenues from box-office ticket sales[26] and from home entertainment sales[27]).

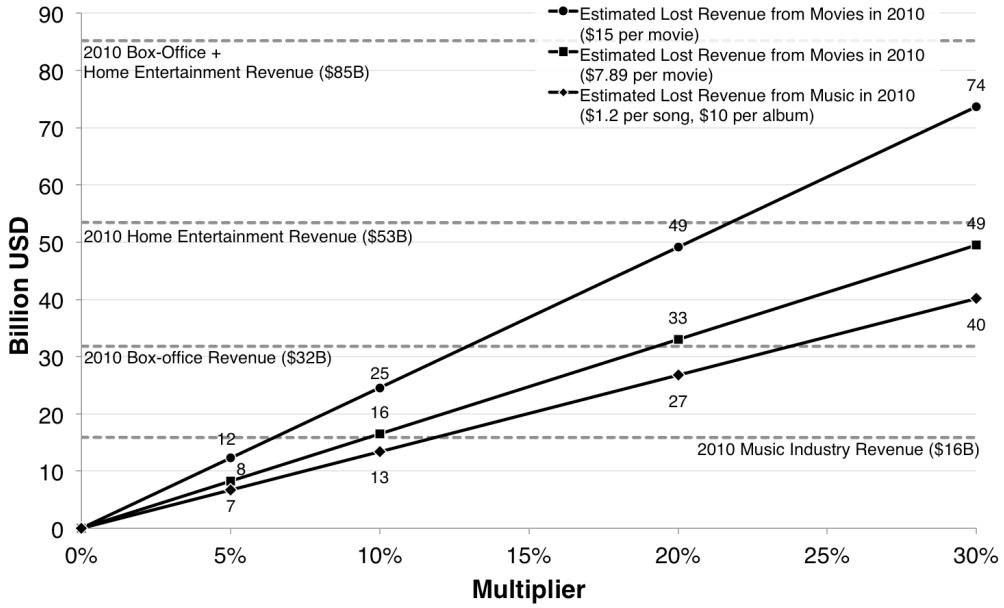


Figure 7. Revenue not realized due to illegal transfers of movies and music as a function of the multiplier.

As an example, if the multiplier were 10% for music, which corresponds to the lowest value used in studies described above that have been cited by the music industry, then this would mean that revenues lost to P2P in 2010 were $13 billion, which is equivalent to 84% of the revenues actually realized by the music industry in that year. However, there have been arguments that a 10% multiplier is too high[28]. If instead the multiplier is 5% for music, then revenues lost to P2P by the music industry in 2010 would still

[25] Calculated using the breakdown presented in http://hd.engadget.com/2008/12/31/on-average-consumers-pay-10-more-for-blu-ray-discs-than-dvd

[26] Obtained from http://www.mpaa.org/Resources/93bbeb16-0e4d-4b7e-b085-3f41c459f9ac.pdf

[27] Obtained from estimates by StrategyAnalytics available at http://www.strategyanalytics.com/default.aspx?mod=pressreleaseviewer&a0=4908

[28] For instance, in an oral interview at the Intellectual Property Breakfast Club, Stephen Siwek, the author of the most cited estimate of lost revenue by the music industry (Siwek 2007), asserted that current estimates should use multipliers smaller than 10% (http://broadbandbreakfast.com/2011/04/intellectual-property-breakfast-club-tackles-the-costs-of-global-piracy/).

be equivalent to close to half of the revenue realized by the industry in that year. In the case of movies, if the multiplier were 5% for example, then lost revenues due to P2P in 2010 would represent between $8 and $12 billion, or 10% to 14% of the industry's revenue from box-office + home entertainment in that year. Even with a 10% multiplier, the impact on the movie industry as a percentage of total revenue is between 19% and 29%, which is smaller than the calculations above for the music industry. However, this is still a significant amount of money in absolute terms; even a multiplier of 1% would yield billions of dollars in revenue not realized, although a relatively small fraction of the industry's revenues overall.

While estimating the impact of illegal transfers on revenues is highly controversial and inherently difficult to do accurately, it is also a very important input to policymaking. Using our estimates of number of copies of copyrighted content transferred using BitTorrent, linear models such as those used in existing literature indicate that, with any multiplier value within the range people have been considering, illegal transfers of copyrighted content would have a significant effect on the revenue of copyright industries. Moreover, at present, it appears that music industry revenues are affected by P2P more than movie industry revenues when measured as a percentage of overall revenues, although the movie industry must also contend with video streaming services that violate copyright law.

3.6 Distribution of Popularity of Transferred Content

In this section we estimate the distribution of popularity of top titles from different types of media transferred using BitTorrent and compare it to that of content sold in legal outlets, to better understand the preferences of users and how these preferences differ between what they can obtain for free and what they pay for. We estimate the popularity of the top 1000 titles of Songs, Movies, Music Albums, TV Show seasons and TV Show episodes transferred in BitTorrent, where popularity is defined as the share of transferred copies of each title (sum of copies transferred in all the swarms that share the title) out of all transferred copies in all swarms sharing the respective type of media.

We find that most BitTorrent transfers of media concentrate in a small number of popular titles, especially in the case of movies and songs. Figure 8 presents the cumulative distribution of popularity of the top 1000 titles transferred using BitTorrent for different types of media, and shows that the 1000 most popular titles in BitTorrent account for more than 50% of transferred copies for all media types except music albums. In the particular case of single songs and movies, it takes respectively the top 38 and top 117 titles to account for half of all transferred copies. Thus, the content preferences of users are highly concentrated. It is particularly surprising that a mere 38 songs could account for half of the transfers worldwide given the tremendous number of songs that are available.

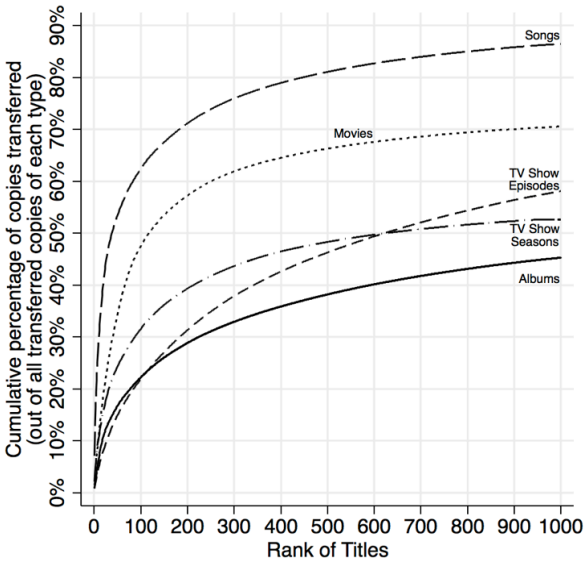


Figure 8. Cumulative distribution of the percentage of copies transferred of the top 1000 titles of Songs, Albums, Movies, TV Show Episodes and TV Show Seasons found in BitTorrent, out of all transferred copies of each type.

In the particular case of movies, as figure 9 shows, the distribution of transferred copies for the 200 most transferred movies in BitTorrent is quite similar to the distribution of worldwide number of sales and rentals (both in DVD and in Blu-ray format) for the top 200 movies most sold/rented worldwide during 2010 (out of movies released in DVD format in 2010)[29]. Hence, the concentration of users' preferences around the most popular titles is similar in BitTorrent transfers and in legal transactions of movies in physical media.

Both figures show that in BitTorrent, a lot of the copies of music and movies transferred (and consequently copyright violations, and possibly impact on revenues of copyright holders) come from a small number of titles, and in the case of movies, the concentration of transfers around the most popular titles is similar to that attained by legal sales and rentals of movies in DVD and Blu-ray media. Furthermore, the specific very popular titles that account for the bulk of BitTorrent transfers are not only popular in BitTorrent, they are also among the most popular titles in terms of legal sales. Looking at users' preferences for specific titles we find that titles that rank high in terms of worldwide sales also rank high in terms of BitTorrent transfers. This is visible in the comparison between sales rank and BitTorrent transfers rank for the 10 top selling singles, albums and box-office movies worldwide and for the 10 top selling DVDs in the U.S. presented in tables 9 through 12 (section 3.4). The tables show that titles in the sales top 10 also rank high in terms of Bittorrent transfers, most of them being part of the BitTorrent top 50. Hence, BitTorrent serves as a source of popular content that is widely available for sale in legal

[29] Data on worldwide movie sales and rentals in DVD and Blu-ray format was obtained from IHS Screen Digest (http://www.screendigest.com/).

outlets, not catering only to those seeking titles that can't be easily found for sale. Such BitTorrent transfers of widely available popular content are likely to displace more potential sales than those of content that is hard to find or even unavailable in legal outlets, and thus they are expected to have a large impact on revenues of copyright holders.

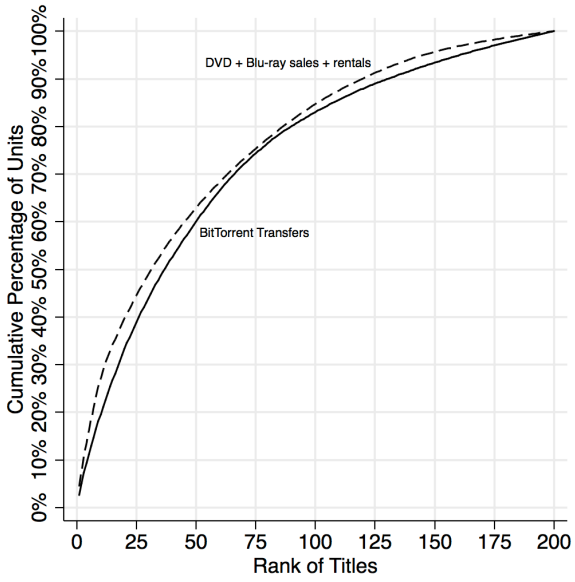


Figure 9. Comparison of the distribution of popularity of the 200 most transferred movies in BitTorrent to the distribution of popularity of the 200 movies with the highest number of sales and rentals both in DVD and Blu-ray format in 2010 (out of movies released in DVD format in 2010).

A direct consequence of the distribution we find for BitTorrent transfers is that a large inventory is not that important a factor when trying to compete with P2P transfers. In fact, the diversity of offered titles seems to be minor when compared to the importance of selecting which really popular titles to offer, especially in the case of music and movies for which it takes only 38 or 117 titles to capture 50% of the "market".

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

These results differ from results we obtained from monitoring a university campus in the U.S. published in a previous study (Mateus and Peha 2011), which, despite finding a significant share of transfers taken up

by popular titles, also found heavy tails in the distribution of popularity of music and video titles. Somewhat surprisingly, it would appear that user preferences for content are more diverse and more diffuse within a single U.S. university than within the entire worldwide population of Internet users. This could mean that U.S. college students are qualitatively different from other groups of P2P users. Alternatively, it could mean that any specific demographic group in any nation would have a variety of content interests, but most of these groups happen to share interest in a small number of audiovisual blockbusters. This difference between users in a U.S. college and users worldwide has many implications. First, it implies that one needs to be especially careful about the conclusions that can be drawn from existing studies that looked at limited pools of users. Second, it implies that if one is creating a store to sell digital content legally, the importance of having a large variety of content depends on intended audience. Third, it implies that determining the economic impact on copyright-holders based on P2P activity is even more complicated. Some groups of P2P users may be more inclined than others to purchase content at full retail price if P2P were not an option, and it is possible that these different groups have very different content preferences.

3.7 Technical Characteristics of Transferred Content

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

By observing the file types shared in swarms for each type of media shared in BitTorrent we find that there is a preferred file type for each type of media, which in most cases accounts for more than three quarters of all transferred copies of content of that type of media. Table 13 shows this by presenting, for each media type, the main file type transferred, the percentage of swarms that contain that file type and the percentage of copies transferred from those swarms. For each type of media, the preferred file format in BitTorrent coincides with the file type that is generally most well known, widespread, and widely supported in terms of hardware and software readers/decoders (*mp3* for music, *avi* for video, Windows executable files for software and *pdf* for documents and books). The second most transferred type of file, for most media types, corresponds to archives. This has implications for copyright enforcement using deep packet inspection (DPI). On one side, it implies that, nowadays, content recognition technology needs only to be able to decode a small set of formats to be able to access the media transferred inside

most files shared in BitTorrent. On the other side, it shows the already significant share of content transferred in BitTorrent that DPI cannot detect because it is transferred inside archives[30]. Using DPI for enforcement may lead to P2P users' behavior changes, in particular, enforcing copyrights for only a small number of file types may lead to users switching to more obscure file types that are not being enforced or to archived content, which in turn will increase the amount of content that cannot be identified by DPI.

Table 13. Preferred file types supplied and transferred for each media type.

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

We looked at method of digitalization of movies and TV shows shared using BitTorrent, and at resolution of movies, TV shows, songs and music albums. To do so, we examined tags indicating the method of digitalization[31], video resolution and audio bit-rate of content found in detected video and music swarms and broke down both the number of swarms and the number of transferred copies by the different categories for each of those variables. Such breakdowns show that high quality copies of movies, TV shows and music are supplied in BitTorrent, and that users transfer preferentially the high quality copies. It is only natural that users prefer the highest quality when there is no difference of price between different qualities of the same content. Considering that the cost of obtaining content from BitTorrent is a function of the number of bytes transferred and of the time spent transferring those bytes, and that many fixed broadband Internet connections use flat rate plans where number of transferred bytes does not influence price, then users' preference for higher quality also shows that they are not sensitive to the time spent in the transfer. For business, the direct consequence of the availability and preference for high quality content in BitTorrent is that those providing legal alternatives can no longer use quality as a differentiating factor to attract customers away from the free but illegal BitTorrent transfers.

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

[31] Refer to Appendix A for a list of tags indicating methods of digitalization of video content shared in P2P and their respective meanings.

In the case of movies, we find that most swarms contain high quality DVD and Blue-ray Rips and that those are the types of most movie copies transferred. Figure 10.a shows a breakdown of the 56% of swarms sharing movies that contained information about the digitalization method. It shows that over 70% of movie swarms contain high quality formats, and those account for close to 70% of transferred copies. The percentage of swarms offering content digitalized prior to DVD release (Cam Rip, Telesync, and Telecine) is smaller, but it is about 15% of swarms and transferred copies. Concerning resolution, the breakdown of the 47% of movie swarms that had such information is portrayed in figure 10.b. DVD quality content accounts for close to 90% of movie swarms and transferred copies, with the remaining swarms containing higher definition content.

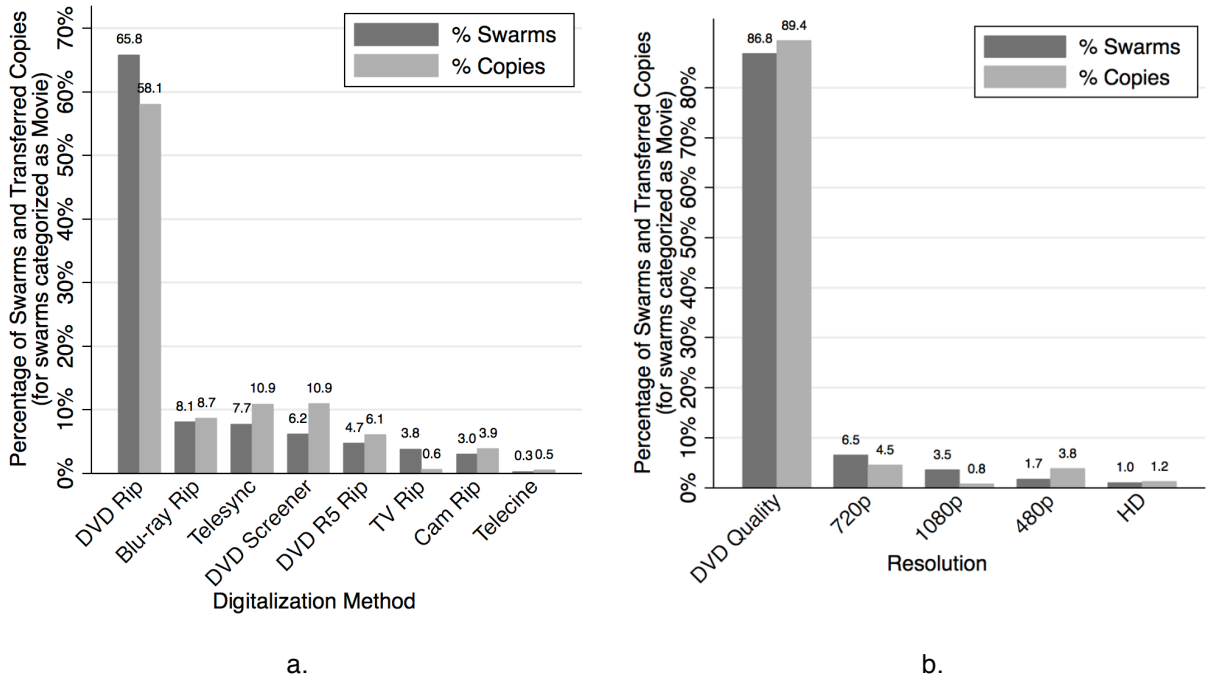


Figure 10. Breakdown of movie swarms and of transferred movie copies by methods of digitalization and resolution. a) Breakdown by method of digitalization. b) Breakdown by resolution.

High quality content is also prevalent in supply and consumption of TV content. However, consumption patterns are different for single episodes or whole TV show seasons. We found information on digitalization method in 21% of the TV show episode swarms and in 24% of the TV show season swarms. Resolution information could be found in 48% of TV show episode swarms and in 37% of TV show season swarms. Both supplied and transferred TV show episodes are high quality content. Most swarms and most transferred copies are TV Rips, obtained from digitally recoding the episode as it is airing, as shown in figure 11.a. Single episodes extracted from DVD rips are the second digitalization method with most swarms and most transferred copies. In terms of resolution, as shown in figure 11.b, most swarms contain copies in HDTV resolution, and that is the preferred resolution in transferred copies as well.

When it comes to full seasons of TV shows, most content supplied is from DVD Rips, and those DVD Rips account for close to three-quarters of transferred copies. This is portrayed in figure 12.a, which also shows a higher percentage of Blu-ray Rips in the case of full seasons than in the case of single episodes, both in supply and consumption of content. As for resolution of transferred content, as shown in figure 12.b, the higher share of swarms and transfers are DVD quality, but the share of high resolution content, in particular 720p and 1080p content, accounts for more than a quarter of swarms and transferred copies.

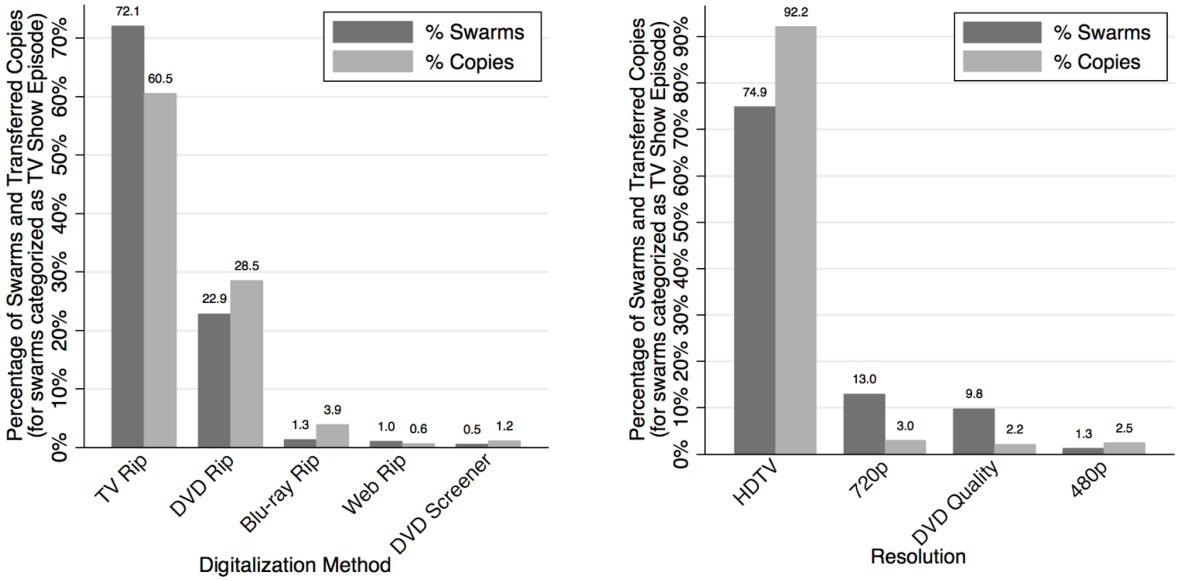


Figure 11. Breakdown of TV show episode swarms and of transferred TV show episode copies by methods of digitalization and resolution. a) Breakdown by method of digitalization. b) Breakdown by resolution.

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

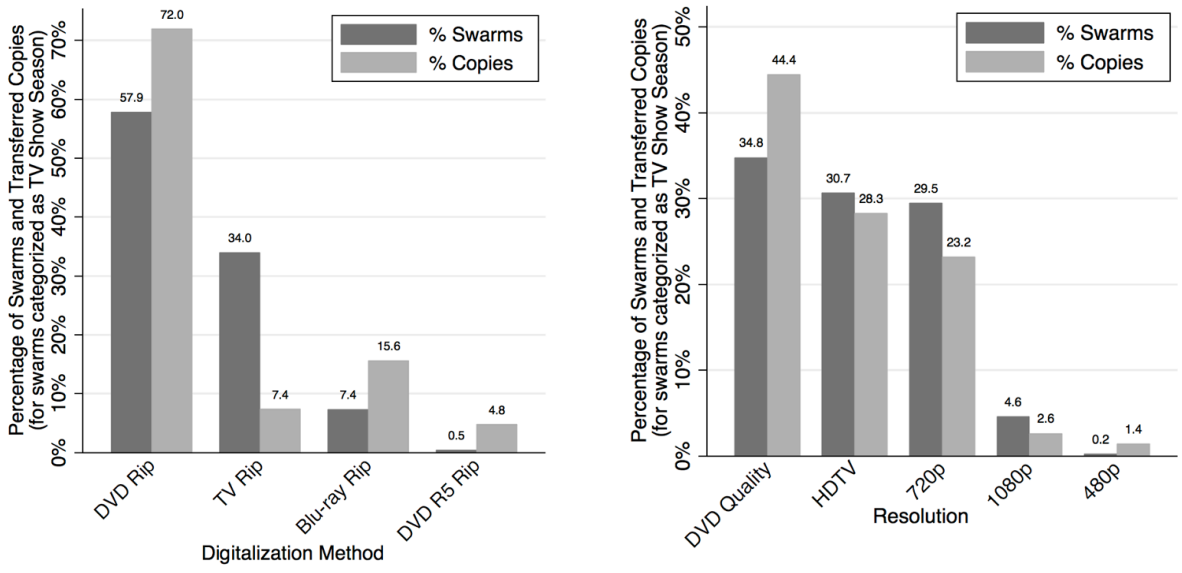
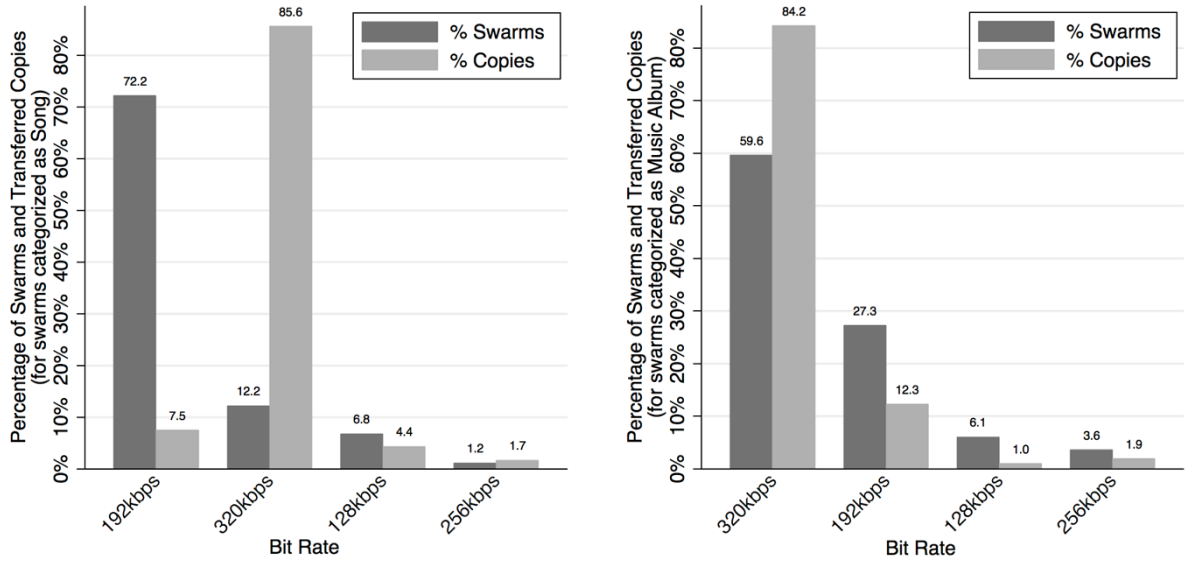


Figure 12. Breakdown of TV Show season bundle swarms and of transferred TV show season bundle copies by methods of digitalization and resolution. a) Breakdown by method of digitalization. b) Breakdown by resolution.

Finally, concerning music, we find that higher-quality content is also preferred to lower quality content. However, since only 6% of song swarms and only 8% of music album swarms contained information on bit-rate, conclusions should be carefully drawn from these data. As shown in figures 13.a and 13.b, most single songs with bit rate information were supplied at a bit rate of 192kbps (higher quality than a regular songs sold in iTunes, which is 128kbps) while most album bundles are supplied at a bit rate of 320 kbps. However, when it comes to consumption, the majority of copies transferred are of the high-end 320kbps media, both for single songs and for album bundles. One possible explanation for this fact concerns available download bandwidth. If the size of transferred songs was a concern in the past due to bandwidth limitations, it is no longer a concern nowadays for most users, who prefer to obtain the higher-quality versions of the content. However, the fact that we also observe a very low supply and number of transfers of lossless music (wav, flac, ape), indicates that users prefer most popular formats in high quality, perhaps due to the convenience allowed by the widespread support for those formats from music playing software and portable music players.



a. b.

Figure 13. Breakdown of song and album bundle swarms and transferred songs and album bundles by bit rate. a) Songs. b) Album bundles.

4 Conclusions and Policy Implications

Using data collected over 106 days between August 2010 and February 2011 from the most popular public BitTorrent tracker, we find an average of 2.6 million BitTorrent swarms offering content at any moment, from which we estimate that a lower bound of 380 million copies of content with more than 1024 bytes are transferred on average per day.

Breaking down the number of swarms and number of transferred copies by type of content shows that the type of content most supplied in BitTorrent is movies, i.e., the type of content shared in the highest number of swarms, but the type of content with the highest number of transferred copies is software. Swarms sharing movies account for 38.7% of all monitored swarms, after which come music albums (with 17.4% of all swarms), TV show episodes (15% of swarms), and then software (7.2% of swarms). These results are in agreement with previous studies that indicated video as the most supplied type of content in BitTorrent. When it comes to actual transfers, movies account for 26.1% of all transferred copies, followed by individual songs (20.4% of all transferred copies), and then by software titles (16.8% of all transferred copies). When the content shared is copyrighted and is being transferred without the permission of copyright holders, then the number of transferred copies provides a more accurate approximation of the number of copyright violations performed using BitTorrent, and probably the economic impact, than the number of swarms offering content.

Differences between supply of content in BitTorrent and actual number of transfers demonstrate an important limitation in previous studies (Envisional 2011; Layton and Watters 2010) that estimated BitTorrent activity by looking only at breakdown of swarms or number of peers connected to each swarm. Such are estimates of supply of content in BitTorrent, which are useful for some purposes, but are not representative of the number of copies transferred, and as such are not informative for purposes of estimating the types of content whose copyright is most often violated in BitTorrent or the types of content for which there is greater economic impact from illegal transfers.

The majority of audio and video content made available and transferred using BitTorrent is likely copyrighted content whose transfers result in copyright violations. We reach this conclusion based on the metadata found for the swarms we monitored, most of which indicate that the shared copies were obtained by digitalizing or re-encoding copyright protected content, and based on the fact that only a very small share of the content made available and transferred using BitTorrent could be found in index websites that specialize in content that can be legally transferred using BitTorrent. Despite the effort from these index websites to promote legal transfers, the swarms they index account for 0.16% of all detected swarms, and transfers from those swarms account for only about 0.55% of all transfers.

Perhaps one of the most important questions regarding illegal transfers of copyrighted content using P2P is how it affects legal sales and the revenue of copyright holders. While we cannot estimate the exact impact of illegal transfers, we find that there are many more transfers of copyrighted movies and songs using BitTorrent than there are legal sales of movies and songs. We estimate that on average 358 million songs were transferred per day using BitTorrent, either as individual songs or as part of albums. This number is 10.7 times greater than the estimated average 33 million songs sold worldwide daily. As for movies, we estimate an average of 57 million copies transferred per day using BitTorrent, a number that is 1.6 times the daily average number of worldwide legal movie transactions (movie theatre admissions plus DVD and Blu-ray sales and rentals plus online sales and rentals). In particular, there were 3.6 BitTorrent transfers for every legal sale or rental of a DVD or Blu-ray, and 227 BitTorrent transfers for every paid download. From a legal perspective, this means that copyright law is infringed hundreds of millions of times per day around the world. From the perspective of the revenue of copyright holders, it indicates that there is likely significant revenue that is not realized due to the impact of such illegal transfers.

Estimating the impact of illegal transfers on revenues of copyright holders is highly controversial and inherently difficult to do accurately, but it is also a very important input to policymaking. Given the number of copies of copyrighted content transferred using BitTorrent presented above, linear models such as those used in existing literature (Siwek 2007; Tera Consultants 2010) indicate that, if one adopts multiplier values within the range people have been considering, the music industry is losing a substantial fraction

of the revenue it could realize in the absence of P2P. The impact of BitTorrent on the movie industry appears to be much smaller as a fraction of total industry revenues, but still billions of dollars in absolute terms. (This does not include illegal transfers of copyrighted content using other technologies that are more prevalent for video than audio, notably video streaming.)

Hence, evidence we collected from BitTorrent trackers corroborates the fact that copyright law is violated frequently using P2P and that such illegal transfers significantly impact the revenue of copyright holders. This calls for considering significant changes, which could be changes in policy, business practices, enforcement methods, technology, consumer education, or a combination of these. This paper's results alone cannot tell us exactly what approach should be followed, but they can help answer some important questions and inform policymaking and business practices.

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

Another question concerns the development of legal alternatives to P2P transfers. We find that most of the copies transferred using BitTorrent (and consequently copyright violations, and possibly impact on revenues of copyright holders) come from a very small number of extremely popular titles. This is particularly evident in the case of songs and movies, for which half of all transferred copies are realized respectively by the top 38 and top 117 titles. Furthermore, we find that the most popular titles in terms of legal sales are also popular in BitTorrent and are among the top titles that account for the bulk of BitTorrent transfers. Hence, BitTorrent is not catering only to those seeking titles that can't be easily found for sale, it serves as a source of popular content that is widely available for sale in legal outlets and whose transfers are likely to displace more potential sales than those of content that is hard to find or

even unavailable in legal outlets. For those seeking to develop legal services to compete with P2P in a global marketplace, the lesson to take from this observation is that the careful selection of which really popular titles to offer, especially in the case of music and movies, can influence a significant share of sales.

The distribution of popularity of titles transferred globally using BitTorrent that we observe here differs from what we had observed when monitoring P2P media transfers in an U.S. university campus in 2007-2008 (Mateus and Peha 2011). A comparison of both distributions shows, somewhat surprisingly, that user preferences for content seem to be more diverse and more diffuse within a single U.S. university than within the entire worldwide population of Internet users. This difference implies that one needs to be especially careful about the conclusions that can be drawn from existing studies that looked at limited pools of users, as those might not be generalizable to larger segments of the population. Furthermore, to those looking to sell digital content legally, it shows that the importance of having a large variety of content depends on intended audience.

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

Finally, our results can also offer some information regarding copyright enforcement using deep packet inspection (DPI), one of the main technologies being considered for that purpose. We find that for each type of media transferred using BitTorrent there is a preferred file type (*mp3* for music, *avi* for video, *exe* for software and *pdf* for documents and books), which in most cases accounts for more than three quarters of all transferred copies of that type of media. The main file type coincides with the file type that is generally most well known, widespread, and widely supported in terms of hardware and software readers/decoders. For most media types, the second most transferred type of file corresponds to archives (*rar*, *zip*, etc.). The implications for copyright enforcement using DPI are twofold. On one side, content recognition technology needs only to focus on a small set of formats to be able to access the media in most files shared in BitTorrent. But on the other side, there is already significant share of content transferred in BitTorrent that DPI cannot identify as copyrighted because it is transferred inside archives. Moving forward, if DPI is used for enforcement, and in particular if enforcement focuses on a small

number of file types, P2P users may start switching to more obscure file types that are not being enforced or to archived content altogether, which in turn will increase the amount of content that cannot be identified by DPI.

5 References

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

Blackburn, David. 2004. Online Piracy and Recorded Music Sales. In Harvard PhD Programme. http://www.katallaxi.se/grejer/blackburn/blackburn_fs.pdf.

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

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

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

GAO. 2010. *Intellectual Property: Observations on Efforts to Quantify the Economic Effects of Counterfeit and Pirated Goods*. Report. United States Government Accountability Office, April 12. http://www.gao.gov/products/GAO-10-423.

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

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

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

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

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

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

Mateus, Alexandre M., and Jon M. Peha. 2008. Dimensions of P2P and Digital Piracy in a University Campus. 36th Telecommunications Policy Research Conference (TPRC). http://www.ece.cmu.edu/~peha/dimensions_of_piracy.pdf.

———. 2009. Characterizing digital media exchanges in a university campus network. 37th Research Conference on Communication, Information and Internet Policy (TPRC). http://www.ece.cmu.edu/~peha/campus_media_exchanges.pdf.

———. 2011. P2P on Campus: Who, What, and How Much. *Accepted for publication in I/S: A Journal of Law and Policy for the Information Society*.

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

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

Oberholzer-Gee, Felix, and Koleman Strumpf. 2009. *File-Sharing and Copyright*. Working Paper 09-132. Harvard Business School.

Peitz, Martin, and Patrick Waelbroeck. 2004. The effect of Internet piracy on music sales: cross-section evidence. *Review of Economic Research on Copyright Issues* 1, no. 2: 71-79.

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

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

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

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

Schulze, Hendrik, and Klaus Mochalski. 2009. *iPoque Internet Study 2008/2009*. iPoque. http://www.ipoque.com/resources/internet-studies/internet-study-2008_2009.

Siwek, Stephen E. 2007. *The True Cost of Sound Recording Piracy to the U.S. Economy*. Policy Report. Institute for Policy Innovation, August. http://ipi.org/IPI/IPIPublications.nsf/PublicationLookupFullTextPDF/51CC65A1D4779E408625733E00529174/$File/SoundRecordingPiracy.pdf.

Tera Consultants. 2010. *Building a Digital Economy: The Importance of Saving Jobs In The EU's Creative Industries*. Report. March 17. http://www.iccwbo.org/uploadedFiles/BASCAP/Pages/Building%20a%20Digital%20Economy%20-%20TERA(1).pdf.

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

Appendix A Meaning of Tags Present in Video Torrent Titles

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

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

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

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

R5: “Typically high quality Telecines intended for the East European market (released in Russian language only)” or Region 5 DVDs (DVDs released in Russia soon after the theatrical release).

HDTV: “Commonly used to tag high definition TV rips.”

SCREENER / DVD-SCREENER (SCR/DVDscr): Extracted from a DVD sent to rental stores and various other places for promotional use. “Usually letterbox format but without the extras that a retail DVD would contain.” Displays a ticker that is not usually in the black bars, and will disrupt the viewing. Typically the quality is very good.

DVDRip: “A copy of the retail DVD and should be excellent quality with no markers/tickers. DVD screeners are sometimes mislabeled as DVD rips”

WORKPRINT (WP): “A copy of the film that has not been finished. It can be missing scenes, music, and quality can range from excellent to very poor. Some WPs are very different from the final print (Men In Black is missing all the aliens, and has actors in their places) and others can contain extra scenes (Jay and Silent Bob).”

BLURAY (BD, BDRIP or Blu-ray): obtained from Blu-ray discs, in high definition format and as such the best quality source commonly available.