Holiday Pictures or Blockbuster Movies? Insights into Copyright Infringement in User Uploads to One-Click File Hosters

Tobias Lauinger[1], Kaan Onarlioglu[1], Abdelberi Chaabane[2], Engin Kirda[1], William Robertson[1], and Mohamed Ali Kaafar[2,3]

[1] Northeastern University, Boston, USA
[2] INRIA, Grenoble, France
[3] NICTA, Sydney, Australia

Abstract. According to copyright holders, One-Click Hosters (OCHs) such as Megaupload are frequently used to host and distribute copyright infringing content. This has spurred numerous initiatives by legislators, law enforcement and content producers. Due to a lack of representative data sets that properly capture private uses of OCHs (such as sharing holiday pictures among friends), to date, there are no reliable estimates of the proportion of legitimate and infringing files being uploaded to OCHs. This situation leaves the field to the partisan arguments brought forward by copyright owners and OCHs. In this paper, we provide empirical data about the uses and misuses of OCHs by analysing six large data sets containing file metadata that we extracted from a range of popular OCHs. We assess the status of these files with regard to copyright infringement and show that at least 26 % to 79 % of them are potentially infringing. Perhaps surprising after the shutdown by the FBI for alleged copyright infringement, we found Megaupload to have the second highest proportion of legitimate files in our study.

Keywords: Abuse, illicit file sharing, one-click hosting, upload analysis

1 Introduction

One-Click Hosters (OCHs) are web-based file hosting services that allow users to upload and share large files. When a file is uploaded, the OCH generates a unique download link for the file. Each file remains private until the corresponding download link is communicated to third parties; this is why OCHs are sometimes also referred to as cyberlockers.

Similar to other file sharing platforms such as peer-to-peer (P2P) systems, OCHs are being (mis)used by certain groups of users to illegally distribute copyrighted commercial content. These users upload the latest movies, TV shows, music, ebooks, and software to OCHs and publish the corresponding links on public web sites (so-called *referral* or *indexing sites*) for everyone to download. On this account, copyright owners accuse several OCHs of being "rogue" sites that

facilitate or even profit from copyright infringement [19]. Lawsuits are pending against several OCHs, such as the criminal indictment against Megaupload[4] that led to the shutdown of the site in January 2012. In their defence, the OCHs regularly point out that their terms of service forbid uploading copyright infringing material [4], and they claim that the most downloaded files are open source software [5], and that they host "over a billion legitimate files" [22].

To date, there is no empirical data about how many files uploaded to OCHs infringe copyright. The situation on OCHs is much more challenging to assess than on P2P-based platforms such as BitTorrent (BT) [1, 21] because OCHs do not reveal the existence of a file unless the corresponding download link is known. Download links for private files might never be published, such as when an OCH is used to store personal backups or to share holiday pictures with friends and family. Therefore, using only public data (as done in [1, 2]) likely underestimates legitimate uploads on OCHs [1]. An exception is the expert report[5] produced by Richard Waterman for the plaintiffs in the Disney v. Hotfile lawsuit. Based on internal data obtained from Hotfile, Waterman estimated that approximately 90.2 % of the daily downloads from Hotfile were highly likely infringing copyright.

While the metric of infringing *downloads* has its merits when aiming to measure the illegal distribution of copyrighted works, it is equally important to quantify the number of infringing *uploads* when studying the role of OCHs in the illegal file sharing ecosystem. In particular, the number of infringing uploads reveals what *types of content* an OCH attracts, as opposed to *how many downloaders* the uploaded content attracts. In fact, since private files are unlikely to generate many downloads, the traffic of even a modest number of popular infringing files can easily dominate the traffic of a potentially much higher number of legitimate files. Our work complements the existing body of research with a different view on copyright infringement on OCHs, and introduces infringement estimates for a range of OCHs not covered before. The Megaupload case, for instance, brought complaints in mainstream media about users who lost access to their private files when the service was shut down by the FBI [10]. We aim at estimating how many legitimate files might have been affected by this event.

Nikiforakis et al. [17] introduced a methodology to guess or predict download links of files hosted on OCHs even when a download link had never been published. While the authors used their methodology to estimate how many uploads were private and to alert users and OCHs to this privacy threat, their work was not concerned with the quantification of possible copyright infringement. In this paper, we apply the methodology by Nikiforakis et al. to collect the names of *all files* uploaded to Easyshare, Filesonic and Wupload over a duration of 48 hours, a subset of the files uploaded to Filefactory during one month, and a random sample of *all available files* on Megaupload in July 2011. These data sets are

[4] Superseding indictment, U.S. v. Kim Dotcom et al., 1:12-cr-00003-LO (E.D. Va., Feb. 16, 2012).

[5] Affidavit Declaration of Dr. Richard Waterman in Support of Plaintiffs' MSJ (Public Redacted Version), Disney Enterprises, Inc. et al v. Hotfile Corp. et al, 1:11-cv-20427 (M.D. Fla., Mar. 5, 2012), filing 325, attachment 6.

independent of whether and where download links were published; therefore, they allow us to estimate the proportion of infringing uploads globally for each OCH and unbiased by any user community. The data sets contain approximately six million file names and cover some of the largest OCHs at the time of our study.

The methodology used in this work could discover files even if they were not intended to be public. We understand that such files can contain sensitive private information. Therefore, we carefully designed a privacy-preserving measurement protocol. As a core principle, we did not download any file contents and analysed only file metadata that was provided by the OCHs' APIs. Section 4.2 contains a detailed discussion of ethical considerations pertaining to our measurements.

Using only file metadata (without downloading and opening a file) to detect whether the file might infringe copyright is a challenging task. File names can be ambiguous or obfuscated; files can be mislabelled and contain fake data or malware, and there may be cases of *fair use* where excerpts of copyrighted content are legitimately used for purposes such as educational or scientific work. While we cannot detect every instance of these cases, we designed our analysis so as to minimise their impact on our final results. Our approach is based on random sampling and manual labelling. That is, we selected representative random samples of 1,000 file names from each OCH and had each file name labelled independently by three different individuals with prior experience in file sharing research. A file name could be labelled as legitimate, infringing, or unknown (when there was not enough information in the file name to make a decision). The assessments were then merged according to a conservative consensus-based algorithm. In order to provide insights into *why* a file was labelled as probably infringing or legitimate, all 6,000 file names in the samples were additionally labelled according to nine heuristics that captured different typical aspects of the names of infringing or legitimate files. We complemented these manual efforts with five automated heuristics.

This paper presents the first detailed and independent study about the extent of potential copyright infringement in the files being uploaded to OCHs. Using a unique data set, we shed light on previously unknown aspects of a common form of abuse of popular web services. Our main findings can be summarised as follows:

- Depending on the OCH, at least 26 % to 79 % of the files appear to be infringing copyright, while we could classify only up to 14 % of the files as likely legitimate. In other words, our findings empirically support the folk wisdom that OCHs are frequently being misused for illegal file sharing.
- In our most conservative scenario, around 4.3 % of the files hosted on Megaupload were detected as legitimate. We estimate that when Megaupload was forced to shut down, more than 10 million legitimate files were taken offline.
- Large files are likely to be infringing, whereas small files are most likely legitimate. The median file size of the two categories differs by two orders of magnitude. Apparently, the ability to share very large files, which is specifically advertised by OCHs, is mainly used for infringing content.

2 Background: The OCH Ecosystem

One-Click Hosters are web-based file hosting services. They are typically implemented in a centralised fashion with thousands of servers located in computing centres [2, 16, 20]. According to previous studies, there are more than 300 OCHs [12]. Labovitz et al. [11] reported that Megaupload accounted for approximately 0.8 % of all Internet inter-domain traffic in July 2009.

There is a wide variety of use cases for OCHs. They can be used to store personal backups, to send potentially large files to friends, and to distribute content to larger user bases—including the unauthorised distribution of copyrighted works. Some OCHs financially reward the uploaders of popular content, which is controversial especially when those files infringe copyright [8, 12].

In contrast to sites such as YouTube, OCHs typically do not offer a searchable index of the hosted files. A file can be downloaded only when the corresponding download link is known. Therefore, uploaders who wish to disseminate their files post the download links on blogs, social networking sites, discussion boards, or they even submit their links to specialised search engines such as Filestube [2, 12, 15, 16]. Mahanti et al. [15] observed OCHs were receiving incoming traffic from up to 8,000 indexing sites. Single indexing sites can be very popular with users and easily rank among the 100 most popular local web sites [12].

Copyright owners are known to scan the Internet for public download links leading to infringing copies of their content and to request that the corresponding OCHs take down those links under the U.S. Digital Millennium Copyright Act (DMCA). According to the criminal indictment[6] against Megaupload, Warner Bros. had 2,500 infringing links removed from Megaupload on a daily basis in September 2009. As of 29 March 2013, the Google Transparency Report[7] refers to 1,279,396 URLs leading to the OCH Rapidgator that are suppressed from Google search results due to copyright complaints.

3 Related Work

There is a wide body of peer-reviewed research in the area of OCHs [2, 8, 12, 13, 15–17, 20]. However, only Antoniades et al. [2] specifically investigated whether the shared files were infringing copyright. They based their analysis on the 100 most recent objects published on a range of indexing sites and found that between 84 % and 100 % of these files appeared to be copyrighted. While such a methodology demonstrates the availability of infringing content on OCHs, it is less suitable for assessing the relative amount of copyright infringement. It tends to underestimate legitimate use cases that do not involve publishing the download links, such as exchanging holiday pictures and other private files, or storing backups. Later works analysed the content types of files downloaded from OCHs as seen in

[6] Superseding indictment, U.S. v. Kim Dotcom et al., 1:12-cr-00003-LO (E.D. Va., Feb. 16, 2012) at ¶ 73 zzz.

[7] `http://google.com/transparencyreport/removals/copyright/domains/?r=all-time`, retrieved 29 March 2013.

network traces gathered at university networks [16, 20] or in crawls of public indexing sites [15], but potential copyright infringement was not investigated.

Nikiforakis et al. [17] introduced a methodology to discover private files stored on OCHs by guessing the associated download links. Most OCHs use download links in the form `http://och/files/{id}/{filename}`, where the file name component is often optional. When such an OCH assigns sequential identifiers, incrementing or decrementing a known identifier yields a new valid download link. Nikiforakis et al. applied this methodology to a number of unidentified OCHs and discovered 310,735 unique files during 30 days. The authors inferred the fraction of potentially private and sensitive files and argued that private files on the affected OCHs were not as private as the OCHs claimed. In contrast to their work, we analyse uploaded files for potential copyright infringement.

In a report commissioned by NBC Universal [1], Envisional Ltd estimated the number of infringing files stored on OCHs. Using an unspecified proprietary methodology, Envisional crawled the Internet for OCH download links. They manually classified a random sample of 2,000 public download links and found 90 % of them to be infringing copyright. However, it is not clear from the report what coverage of public OCH download links Envisional achieved. In contrast, we extracted download links directly from some of the largest OCHs; therefore, our results are not biased by the fact that some download links were not found by a crawler, or not even published at all. Furthermore, we provide details about how we classified the files, making our results more traceable.

In his expert declaration in Disney v. Hotfile, Waterman outlined the methodology that led him to estimate that 90.2 % of the daily downloads from Hotfile were highly likely infringing: File data was provided by Hotfile, a sample of 1,750 files was drawn at random (weighted by the number of downloads), and each file in the sample was opened and inspected by a copyright lawyer. While Waterman's methodology estimates infringing downloads, we estimate infringing uploads, which is a complementary approach. Furthermore, we cover a wider range of OCHs, highlighting the differences in the data sets, and we provide additional insights into various metrics beyond copyright infringement.

Other studies estimated the fraction of infringing content shared using BitTorrent (BT) [1, 21]. However, OCHs and BT differ significantly from both a technical and administrative point of view, so that the results cannot be compared directly.

4 Methodology

At a high level, our methodology consists of gathering data sets with the names, sizes and optional descriptions of files uploaded to five large OCHs and a reupload service. For privacy reasons, we do not download any of these files. We manually classify a random sample of 1,000 file names per data set and complement this overall assessment of copyright infringement with fourteen manual and automated heuristics (as defined in Section 4.3) to better illustrate our manual classification.

Table 1. Overview of the file metadata sets extracted from five OCHs and the reupload service Undeadlink in 2011. For a description of how files were merged, see Section 4.1. File sizes are not available for Easy-share because they were not provided by the API.

One-Click Hoster	Easy-share (ES)	Filesonic (FS)	Wupload (WU)
Time Frame	24 h starting 27 Jul and 7 Aug 15:00 GMT		
Discovered Files	53,145	1,857,770	2,393,090
split archives or files	38.87 %	55.42 %	36.49 %
Discovered Bytes	n/a	547 TB	588 TB
Files after Merging	36,855	1,015,898	1,686,388
merged comp./incomp.	10.02 % / 1.83 %	14.89 % / 3.62 %	8.43 % / 1.44 %
Comments	all files uploaded during time period (enumerated without gaps)		
One-Click Hoster	Filefactory (FF)	Megaupload (MU)	Undeadlink (UL)
Time Frame	16 Jun to 16 Jul	16 Jun to 25 Jul	28 Apr to 5 Dec
Discovered Files	1,755,967	32,806	204,263
split archives or files	33.59 %	35.99 %	36.12 %
Discovered Bytes	264 TB	4.7 TB	114.7 TB
Files after Merging	1,287,726	-	148,400
merged comp./incomp.	7.18 % / 2.26 %	- / -	5.68 % / 6.40 %
Comments	uploaded files (enumerated with gaps)	available files (random sample)	first uploads only (reupload service)

4.1 Data Sets

We base our analysis on file metadata extracted directly from five large OCHs. Additional real-time statistics published by the reupload service Undeadlink allow us to validate our classification and heuristics.

OCHs. To obtain lists with files uploaded to OCHs, we followed the methodology introduced by Nikiforakis et al. [17] and applied it with some variations to five medium-sized and large OCHs. Filefactory, Easy-share, Filesonic and Wupload used sequential file identifiers with optional file names and were subject to *enumeration* of identifiers. Megaupload used random file identifiers and we discovered files by *guessing* identifiers. Table 1 summarises the file data sets.

All five OCHs offered APIs to access metadata and availability information about the hosted files. The APIs allowed to check between 100 and 500 identifiers in one request. For each given identifier, the API returned the availability status (available or unavailable), and if applicable the file name and size as well as an optional user-supplied description of the file. In all our experiments, we only accessed the metadata APIs. That is, we never accessed the contents of the files.

On Filefactory, we obtained a current file identifier by manually uploading a test file and extracted the identifier from the corresponding download link.

We enumerated file identifiers towards the older uploads and occasionally reset the starting point to a fresh identifier. This was necessary because we noticed *unassigned gaps in the sequential identifier space*; link identifiers appeared to be assigned in batches (possibly for load balancing over several servers). We decided to keep this data set nevertheless because of its interesting characteristics, but we caution that the results are necessarily less conclusive than for the other OCHs.

Easy-share, Filesonic and Wupload also used sequential file identifiers. However, on these OCHs, we designed our experiment in a different way: To obtain valid current file identifiers, we automatically uploaded a test file every 30 minutes. We then enumerated all file identifiers between two subsequent test uploads. Following this methodology, we discovered new files within at most one hour of their upload. Our data sets contain *all files uploaded* to the respective OCH during two contiuous 24-hour periods, and they cover business days (Wednesday to Thursday) as well as the end of the weekend (Sunday to Monday).

Megaupload used random identifiers drawn from a space of size 36^8. By randomly guessing identifiers, we discovered a valid file for every 11,275 identifiers that we tested (one hit every 23 API requests), resulting in a sample of 36,657 file names. In contrast to the data sets gathered from the OCHs with sequential file identifiers, the Megaupload data set is a *sample of all files that were available* on Megaupload's servers at the time of the experiment, independent of the original upload time. From the density of Megaupload's file identifier space, we estimate that Megaupload stored approximately 250 million files on their servers in July 2011. Extrapolating from the file sizes found in the sample, the total storage capacity in use was around 33 PB (but not accounting for potential internal de-duplication of files with identical contents). We noticed that many files were called `video.flv` or `megabox.mp3` (9.5 % and 1 % of the files, respectively). These files appeared to correspond to internal data used by Megaupload's video and music streaming services Megavideo and Megabox, respectively. As these file names do not reveal whether the file contents might be copyrighted and shared illegally, we excluded these files from the following analysis. In the remainder of the paper, we considered only the 32,806 remaining files (89.5 %) because these files represented the actual workload of the file hosting service Megaupload.

Undeadlink was a service that generated new "undead" download links for Megaupload download links submitted by uploaders. Users following such a link were redirected to a live copy of the corresponding file on Megaupload. Undeadlink monitored the availability of submitted files on Megaupload and automatically reuploaded a new copy when the original file became unavailable.

Undeadlink's web site displayed the service's (re)upload queue in real time as well as a live list of the HTTP referrers of users clicking on "undead" download links. We continually extracted this data until Undeadlink was taken offline. To construct a data set of uploaded files, we retained only the first upload (per internal link identifier) and discarded any repeated upload (due to a file becoming unavailable on Megaupload). Table 1 summarises this data set.

Because of Undeadlink's functionality and the way it was advertised, we hypothesise that Undeadlink was predominantly used to protect infringing files from DMCA takedown efforts. To back up this hypothesis, we analysed the top 50 domain names found in the live HTTP referrer list of users clicking on Undeadlink download links. Among these 50 domains (representing 98.7 % of all clicks), 78.4 % of the clicks came from known and manifestly infringing indexing sites, 17.1 % from services allowing uploaders to monetise their download links (by displaying advertisements), 4.2 % of the clicks came from various unclassified web sites, and 0.2 % originated from search engines. These numbers illustrate that the vast majority of Undeadlink's (download) click traffic was very likely infringing, and we expect similar results to hold for Undeadlink's file uploads. Thus, we can use the Undeadlink data set as a benchmark for our file classification.

Dataset Processing. When analysing the file name data sets, we observed many files with extensions such as `.part1.rar`, `.r02`, and `.003` representing parts of split archives (e.g., more than half of all files on Filesonic). Since a single split archive can consist of hundreds of parts but corresponds to at most one instance of copyright infringement, not accounting for this phenomenon can overestimate copyright infringement. For this reason, we generated new data sets by virtually "reassembling" these split files. That is, we merged the names of parts into a *complete* file name whenever we found a full sequence of increasing part numbers, where all parts had the same name prefix, archive type and size, except for the last part, which was allowed to be smaller. As an example, consider the parts `etarepsed_seviwesuoh_503-.part1.rar` (100 MB), `etarepsed_seviwesuoh_503.part2.rar` (100 MB) and `etarepsed_seviwesuoh_503.part3.rar` (73 MB), which would be merged into a single "virtual" file name `etarepsed_seviwesuoh_503.rar` (273 MB). When parts were missing, we merged these file names nevertheless and marked them as *incomplete*. In the remainder of the paper, we always use the "reassembled" data set, and we either include or exclude the names of incomplete files depending on the context. The labelled samples, for instance, include the names of incomplete archives. Table 1 shows the size of the data sets before and after merging file names corresponding to split archives, and the fraction of files in the merged data set that were "reassembled" successfully. On Filesonic, the initial 55.42 % of split archive files account for only 18.51 % of the file names when merged.

4.2 Ethical Considerations

The purpose of this study is to estimate the proportion of files related to illegal file sharing on OCHs. In designing our measurement setup, we needed to find a balance between our interest in accurate data, and the users' interest in privacy. In order to make our data sets most accurate, we would need to download and inspect the contents of all uploaded files, including those that were never intended to be public and might contain sensitive information. On the other hand, fully excluding any risk of privacy violation would impose using only public data

sources. However, using only published download links would make it unfeasible to quantify the percentage of legitimate content. Such content (including family pictures or school work) is less likely to have public download links than material such as infringing copies of full-length Hollywood movies. Furthermore, even public or semi-public download links such as those found in "private" file sharing communities are not necessarily indexed by search engines, which makes it unfeasible to gather a representative sample even of public download links.

The compromise that we followed for this work was to extract from OCHs the metadata of all files, including private ones, but not to download the files themselves. The metadata we used consisted of the file identifier assigned by the OCH and the corresponding file name, file size, and an optional description of the file that the uploader could supply. The data we gathered and used contains no unique user identifiers, IP addresses or other personally identifiable information. Consequently, identifying uploaders would have been possible only in exceptional cases (by using URLs or user names supplied by the uploaders in the file name or description fields), but at no point did we attempt to do so. Furthermore, we separated the collection of the data set from its analysis, so that the researchers who labelled the file metadata had no access to the files' download links. Therefore, we consider our data sets to be anonymous and preservative of users' privacy.

The analysis that we carried out was purely passive; the only risk for users would have been a privacy breach by disclosing or otherwise misusing the data that we gathered. We handled the data set in a confidential way and disclosed only aggregate statistics as well as single, uncritical file names in order to illustrate our labelling methodology. Note, furthermore, that the methodology we used to gather our data sets was published by Nikiforakis et al. in February 2011 and was shown to be used by third parties for unknown (and potentially nefarious) purposes [17]. Therefore, the additional privacy risk induced by our data collection is negligible compared to the existing privacy threats.

4.3 Analysis Approach

In order to determine the legitimacy or potential copyright infringement of uploaded files, we chose a random sampling and manual labelling approach. From each of the six data sets, we selected 1,000 file names at random. According to standard theory about confidence intervals for proportions (Equation 1, e.g. Chapter 13.9.2 in [7]), for a sample size of $n = 1000$, the actual proportion in the full data set will lie in an interval of ± 0.03 around the proportion p observed in the sample with 95 % probability ($\alpha = 0.05$) in the worst case (i.e., $p = 0.5$). The implication is that our samples allow us to estimate with high confidence the proportion of infringing files in the full data sets.

$$p \pm z_{1-\alpha/2}\sqrt{\frac{p(1-p)}{n}} \quad \text{with } np \geq 10,\ n(1-p) \geq 10 \text{ and } z_{0.975} = 1.96\ . \quad (1)$$

A precondition for this extrapolation is that we accurately label the samples. Since we cannot verify the accuracy of our labelling process, we designed a protocol

Table 2. The manual heuristics for file names and descriptions. Many examples given in the table satisfy several heuristics; a few names were shortened (`[...]`).

ID	Description	Examples from MU: `file name` (file description)
I1	warez-like name	`Oceans.Thirteen.2007.1080p.BluRay.x264-HDEX.part06.rar`
I2	uploader name	`Kyle.Xy.S01e10.Dvdrip.Dual.Audio.[By.Mixel].avi.002`
I3	indexing site URL	`megauploadz.com.hr9rgp6jr9ixpuvq7wnq2v0kspnh9r.avi`
I4	commercial name	`South.Park.S13E13.avi`, `Lady Gaga - Just Dance.mp3`
I5	file sharing keyw.	`Acrobat.9.Pro.Cracked.rar` (AcroPro crack)
I6	obfuscated name	`042e2239101007.part09.rar`, `----------.rar`, `[...]C€l!n€ D!0n (1998-FRA) - @µ C0€µr Dµ St@d€.BGL`
L1	free/shareware	`Alcohol120_trial_1.9.7.6221.exe`, `ubuntu-11.04-desk[...]`
L2	unsuspicious ext.	`Cover letter .doc`, `crashreporter.ini`, `favicon_.ico`
L3	name or descr. suggesting personal content	`Jura2008.zip` (Photos Toussaint 2008), `DSC00318.JPG`, `IMG_0366.JPG`, `MOV00026.3GP`, `William Shakespeare.pptx`, `Lottery Number Picker (Uses Random and Array).zip`

that required each sample to be labelled independently by three researchers. We then merged the results into a single assessment by applying either a consensus or majority approach. We decided not to crowdsource the labelling task in order to avoid issues with training and data confidentiality.

In the *overall assessment*, each file in the samples was labelled according to the intuition and experience of the researcher as being either potentially *infringing*, *legitimate*, or as *unknown* if the file name was too ambiguous to make an informed decision. We complemented our data sample by having each researcher label the file names according to nine additional binary heuristics as summarised in Table 2. The purpose of these heuristics is not to build an automated classification tool; in fact, many of the heuristics are difficult to compute automatically and could be easily circumvented by uploaders if they had a reason to do so. Rather, we use these heuristics to provide insights into *why* a file was classified as potentially infringing. Six of the heuristics indicate possible copyright infringement, while three heuristics cover content that appears to be legitimate.

Heuristics suggesting infringing content (I*)

I1. *Warez scene title or release group name:* The file name follows the conventions of the Warez scene [18] or related milieux. Often uses periods instead of spaces and includes quality attributes and the name of the release group.
I2. *Uploader name:* The file name/description contains the pseudonym of the uploader. Occurs on discussion boards to increase the prestige of the uploader.
I3. *URL of indexing site:* The file name/description contains the URL of an indexing site. Often used as an advertisement vector and to "tag" the uploads.
I4. *File name or description contains the name of commercially exploited copyrighted content:* The file name or description suggests that the file contains a *specific* piece of content that is normally sold or rented, such as an episode

of a TV show `Lost.S04E02.part1.rar`, or music by Michael Jackson, *and* there is no indication of any fair use case, such as *essay*, *extract*, or *trailer*.

I5. *Keywords typical for file sharing:* The file name or description contains file sharing jargon such as *DVDrip*, *screener*, *keygen* or *crack*, but also season/episode indications such as *S03E09* for TV shows. While serial number generators or cracks might not infringe copyright, we include them here because their most likely intent is to enable unauthorised use of software.

I6. *Obfuscated file name:* The file name is seemingly random (and unlikely to be an abbreviation). Such random names have been observed on indexing sites. Also includes human-readable file names with some characters replaced, such as @ instead of *a*, which may be an attempt to circumvent simple keyword-based file name filters, e.g. Céline Dion's concert *Au cœur du stade* in Table 2. Also covers contradictory file extensions such as `.part1.rar.jpg`.

Heuristics suggesting legitimate content (L*)

L1. *Freeware, shareware (without crack), and abandonware:* The file name suggests freeware (such as a free Linux distribution), abandonware (such as old console games that are not commercialised any more), shareware, or evaluation versions of commercial software *without* a crack, serial number generator, and not labelled as infringing "full" version.

L2. *Unsuspicious file extensions:* File extensions not typically used in an illegal file-sharing context. Includes extensions for documents (`.doc`, `.odp`, `.pps`, `.xls`, `.html`, `.psd`, `.jpg` etc.), but excludes "ambiguous" extensions such as `.pdf` (sometimes infringing ebooks).

L3. *Personal and small-scale commercial content:* Files likely produced in a personal context (holiday pictures, home movies, archives of such content, and files following known naming schemes of photo cameras and mobile phones). The file name and description must be specific enough to provide confidence that the contents are indeed legitimate. Does not cover `back-up.rar` or `pictures.rar` (sometimes used to conceal copyrighted content), but does cover `pictures-california-holidays.rar` (lower probability of mislabelling). Also includes content that might not be intended to be shared on OCHs, but that is not typical either for the large-scale copyright infringement we aim to characterise, such as source code, lecture slides, or research papers.

In addition to the manually labelled heuristics, we applied five automated heuristics to the random samples. They correspond to aspects of potentially copyrighted files that can be computed in an automated way.

Automated heuristics (A*)

A1. *Split files:* The file is split into several parts (see Section 4.1). Often used to bypass file size restrictions for free users on OCHs or to allow parallel downloads, but also a tradition in the Warez scene.

A2. *Duplicate files:* The same file has been uploaded several times to the same OCH. Applies if a file with the same name and size (except for Easy-share) is found in the corresponding *full* data set. Unlikely for personal content.
A3. *Public link:* Google returns at least one result when searching for the file name (exact match).
A4. *DMCA takedown notice:* Google reports that at least one search result could not be displayed because they received a DMCA takedown notice from a copyright holder (when searching for the file name).
A5. *Hit in database of infringing file names:* File name found in a database of 3.4 million download links extracted from more than ten known infringing indexing sites in prior work [12, 13].

By definition, heuristics are not exact; we do not treat them as accurate indicators of copyright infringement. Rather, we use them to illustrate characteristics of potentially infringing files. We exclusively rely on the independent overall assessment of the three researchers to classify a file as infringing or legitimate.

4.4 Limitations

Motivated by privacy concerns, the choices that we made when designing our experiments induce inherent limitations on the results presented in this paper.

Our choice not to download any files because of ethical considerations means that we cannot evaluate the correctness of our classification. This is an issue especially for *mislabelled files* that do not contain what their file name suggests, or files with *obfuscated file names* where the name reveals nothing concrete about the files' contents. Furthermore, fair use may not be discernible from the file metadata alone. While we acknowledge that our results cannot be exact (this would be difficult to achieve even with access to the files' contents), we are confident that our results reflect the general trends of illegal file sharing occurring on OCHs. To make our file classification methodology more transparent, we defined a set of heuristics. In order to reduce personal bias, the file metadata samples were labelled independently by three researchers and the results were merged using a conservative consensus algorithm.

For a separate study, we conducted an experiment to estimate the proportion of polluted content on two popular indexing sites that allowed anonymous posts. File pollution can occur due to intentionally or unintentionally mislabelled files. We found that more than 93 % of the indexed files were authentic [13]. We do not claim that these findings can be extrapolated to the data sets used in this paper. There are reports about malware being hosted on OCHs [9], for instance. Yet, in contrast to P2P [3, 14], copyright owners do not appear to upload fake files to OCHs because they can use DMCA takedown notices to *remove* infringing files, which we assume to be more effective than *adding* fake files.

5 Analysis

Ideally, the classification result of our file name labelling should be a binary label, either *legitimate* or *infringing*. In practice, however, it is very challenging to make

Table 3. Consensus among the three labellers for the overall assessment and heuristics.

	Frequency of Consensus (%)					
Heuristic	FF	ES	FS	WU	MU	UL
Overall Assessment	57	79	77	86	56	84
I1 *Warez Name*	95	92	88	81	90	85
I2 *Uploader Name*	99	98	97	98	91	94
I3 *Indexing URL*	98	99	97	95	91	96
I4 *Commercial*	73	82	75	76	72	72
I5 *Keywords*	94	72	87	78	87	77
I6 *Obfuscated*	98	96	98	98	96	99
L1 *Freeware*	98	99	99	100	98	100
L2 *Legit. Extension*	97	98	100	100	98	100
L3 *Personal*	85	97	93	99	92	100

a binary decision for each file, especially when the file contents are not available as in our study. In the following, we explain how our conservative approach is responsible for a relatively large fraction of files with *unknown* label on some OCHs, and we present the overall assessment results obtained by merging the classifications of the three labellers. Subsequently, we analyse the individual heuristic indicators to gain more confidence in our overall labels, and we provide further insights into some characteristics of files uploaded to OCHs.

5.1 Consensus Merging and Unknown Labels

To merge the independent labelling results of the three researchers, we applied a consensus algorithm. That is, we conservatively assumed that a heuristic did not apply (or that the OVERALL assessment was *unknown*) unless all three researchers agreed. According to Table 3, a consensus in the OVERALL assessment was reached for a little more than half of the files in the Filefactory and Megaupload samples. As a corollary, the remaining file names were automatically classified as *unknown* (in addition to those already classified as *unknown* by all three researchers because of ambiguous file names). This was partially due to Filefactory and Megaupload hosting the largest fraction of files named in foreign languages and coming from cultural backgrounds that the researchers were not familiar with. These OCHs also hosted the largest detected fraction of *legitimate* files. In our experience, such files were generally more difficult to classify than large-scale commercial content because the situation was often more ambiguous, leading one researcher to label a file as *legitimate* while the others marked it as *unknown*. Other OCHs exhibited a less ambiguous workload. The "benchmark" data set Undeadlink, for instance, was labelled with a 16.3 % dissent rate plus 4.2 % consensually *unknown* files, resulting in 20.5 % *unknown* files for OVERALL. Across all OCHs, pornography was frequently classified as *unknown*, especially when the file name contained

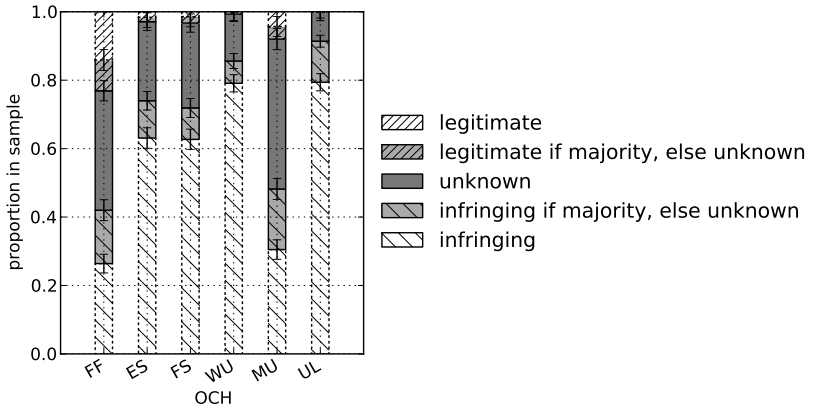


Fig. 1. The file name classification results for the six samples. The area shaded in dark grey corresponds to files with *unknown* classification. If only a **majority** among the three labellers is required for classification, the entire hatched area above corresponds to the proportion of *legitimate* files, whereas the hatched area below corresponds to files classified as *infringing*. In the more conservative case requiring **consensus** between the three labellers, the areas shaded in light grey become *unknown*. The plot shows 95 % confidence intervals. The real-world ratio between infringing and legitimate files is likely to lie in the *unknown* area (plus confidence intervals).

a scene number as in `my-sexy-kittens-29-scene1.mp4`, because it remained unclear whether it was an infringing copy or public advertisement material.

The situation for the individual heuristics was similar, except that all decisions were binary and did not permit an *unknown* value. Obfuscated file names (I6) were difficult to classify because it was often unclear whether a file name was random or an unrecognised (but meaningful) abbreviation. For shareware, it was often impossible to distinguish between a cracked version and a legitimate evaluation copy. The degree of consensus is lowest for I4 (commercial content) because it was the heuristic where the most non-trivial decisions had to be made. Other heuristics such as L1 (freeware) clearly did not apply to most files. The few realistic candidates for freeware often led to disagreements, but their number was small compared to the overall size of the data sets.

5.2 Overall File Classification

We were able to detect significant proportions of legitimate uploads only for Filefactory and Megaupload. Figure 1 shows that for the remaining OCHs, even if we assumed all *unknown* files to be legitimate, we would still estimate more than half of all uploads to be infringing. One possible explanation for this effect is that Filefactory and Megaupload were the oldest OCHs in our data sets, which might have allowed them to gain popularity with legitimate users. Wupload, in contrast, had been launched just a few months before our measurement. We estimate that

at least 79 % of the files uploaded to Wupload during our measurement infringe copyright, the highest proportion among the OCHs in our data sets. As expected in Section 4.1, Undeadlink equally exhibits a very high level of infringing files. The estimated lower bound of 4.3 % legitimate files on Megaupload might not seem very high, but compared to the overall estimate of 250 million hosted files, this still implies that the forced Megaupload shutdown resulted in at least 10.75 million legitimate files being taken offline.

Because the consensus approach might be overly conservative for some of the difficult decisions, we additionally merged the classifications of the three labellers using a majority voting algorithm: A file was labelled as *legitimate* or *infringing* when at least two of the researchers agreed. The difference between the two approaches is shown in Figure 1 through the different shades of grey. The majority strategy allows to classify more files as *legitimate* or *infringing* and thereby reduces the number of *unknown* files. However, this comes at the cost of lower confidence in the accuracy of the labels, thus we decided to retain the more conservative consensus merging for the remainder of this paper.

5.3 Heuristic Analysis

Given the OVERALL classification, we visualise in Table 4 the probability of each heuristic. The heuristics for commercial content (I4) and file sharing keywords (I5) apply frequently to the files classified as *infringing*, e.g. I4 applies to 80 % of the *infringing* files on Undeadlink, but only very rarely to files classified as *legitimate* or *unknown*. Similar results hold for legitimate file extensions (L2) and personal content (L3), which apply almost exclusively to files classified as *legitimate*. All three labellers classified `.jpg` as a potentially legitimate file extension, which was fairly frequent on Filefactory. However, not all `.jpg` files were eventually labelled as *legitimate* because some of them contained the names of models, for instance, leading to a relatively high number of *unknown* files with legitimate extensions. All in all, the heuristics apply to the file classifications in a consistent manner, which increases our confidence that the OVERALL classification is reasonable.

Among the automated heuristics, *infringing* files were split more frequently than *legitimate* files. Even though most *infringing* files were uploaded multiple times, there were non-negligible numbers of *legitimate* files that were duplicates as well. Surprisingly, there was a generally low number of DMCA takedown notices or hits in our database of infringing files for file names of all classifications. Heuristic A3 (public links) appears to be a poor indicator for infringement as it applies to *legitimate* files as much as to *infringing* files. This supports our opinion that automated classifiers not based on "curated" file name, checksum or provenance blacklists are likely to suffer from high false positive rates.

5.4 File Extensions

We analysed the file extensions being used in the full reassembled data sets (including incomplete files). Table 5 shows the five most frequent file extensions and the associated file extension entropy per data set. Some OCHs exhibit a

Table 4. Manual and automated file classification results with consensus merging for the manual heuristics. Given is p(classification) for OVERALL and p(heuristic | classification) for each heuristic, where the classification is *legitimate/infringing/unknown*. The results are coded in a greyscale from 0 % ( ) to 100 % (■). Due to the low number of *legitimate* files, the conditional probabilities $p(\cdot \mid \text{legitimate})$ for OCHs other than FF and MU are based on too few examples to be considered exact (e.g., L1 on WU, or A5 on ES and FS). File names labelled as *infringing* frequently contained the name of commercial software (I4) or were duplicates (A2); file names classified as *legitimate* often used a legitimate file extension (L2) or referred to personal content (L3).

	Conditional Heuristic % with Consensus (legit./infr./unknown)					
Heuristic	FF	ES	FS	WU	MU	UL
Overall	14/26/60	1.6/63/35	1.4/63/36	0.1/79/21	4.3/31/65	0.1/79/21
I1 *Warez Name*						
I2 *Uploader Name*						
I3 *Indexing URL*						
I4 *Commercial*						
I5 *Keywords*						
I6 *Obfuscated*						
L1 *Freeware*						
L2 *Legit. Ext.*						
L3 *Personal*						
A1 *Split File*						
A2 *Duplicates*					n/a	
A3 *Public Link*						
A4 *DMCA Notice*						
A5 *In Infr. DB*						

more uniform file type workload than others, with their file extension distribution being more heavily skewed toward `.rar` archives, `.avi` movies and `.mp3` audio files. This observation is captured by a lower file extension entropy and appears to be correlated with a higher estimated proportion of copyright infringement as reported in Table 4. A higher diversity in uploaded file types appears to be a characteristic of the OCHs hosting a higher proportion of legitimate files.

5.5 File Size Distribution

Files classified as *legitimate* tend to be two orders of magnitude smaller than *infringing* files. The median file sizes on Megaupload are 2.37 MB vs. 171.74 MB, and on Filefactory 1.32 MB vs. 150.69 MB. The median size of *unknown* files is 36.23 MB on Megaupload and 6.98 MB on Filefactory, suggesting that both legitimate and infringing files were labelled as *unknown*. Recall that file size was not used as a classification criterion. Incomplete archives were excluded from this

Table 5. The most frequent file extensions from the full data sets (out of more than 1,000 different extensions).

	FF		ES		FS		WU		MU		UL	
Rank	Ext.	%	Ext.	%	Ext.	%	Ext.	%	Ext.	%	Ext.	%
1	`rar`	23.9	`rar`	45.7	`rar`	57.5	`rar`	61.5	`rar`	46.5	`avi`	66.4
2	`jpg`	18.1	`mp3`	20.2	`avi`	14.6	`avi`	15.3	`avi`	13.2	`rar`	16.9
3	`mp3`	8.3	`avi`	8.8	`jpg`	5.3	`mp3`	6.3	`zip`	6.3	`mkv`	5.8
4	`avi`	7.9	`wmv`	6.0	`wmv`	5.1	`zip`	5.5	`mp3`	4.9	`xtm`	2.9
5	`pdf`	5.7	`zip`	3.8	`zip`	4.0	`wmv`	3.3	`7z`	4.8	`mp4`	2.8
Entropy	4.28 bits		2.83 bits		2.52 bits		2.14 bits		3.37 bits		1.80 bits	

analysis because their file size was not available. Figure 2 shows this data from a different point of view. It plots, for a varying upper file size limit, the fraction of files classified as *legitimate*, *infringing*, and *unknown*, respectively. Smaller files are much more likely to be classified as *legitimate* than larger files. The capability of storing files larger than a few hundred MB, which is specifically advertised by OCHs, appears to be mainly used for infringing activities.

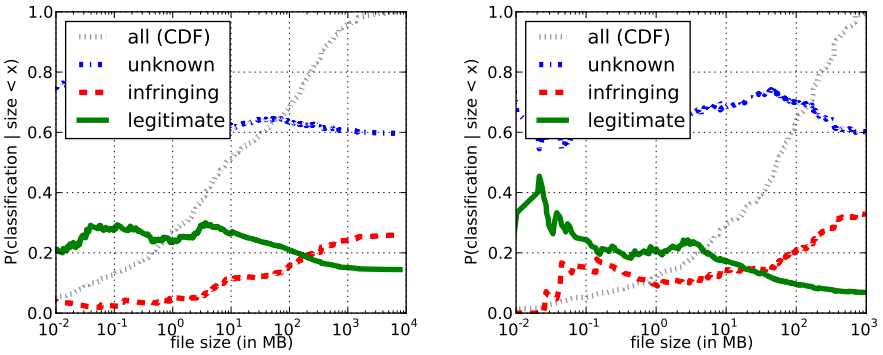
5.6 Indexing Site URLs

Some uploaders add an URL to the names or descriptions of the files that they upload in order to advertise their sites. Attempts at automatically extracting URLs from file names generated too many false positives; `.PL`, for instance, can stand for both a top-level domain and the language of a movie. Instead, we manually extracted all URLs contained in the file names of the labelled samples and verified that they were indeed indexing sites. Subsequently, we looked up these URLs in the full data sets (including incomplete archives).

Table 6 lists the three most frequent URLs from each data set together with the language of the respective web site. These sites include Warez boards and blogs, span many different languages and offer varying types of content. The most active site `noor7.us` uploaded 7,070 files to Wupload within only 48 hours.

We can estimate how many files these sites had currently available on Megaupload at the time of the measurement. `megauploadforum.net`, for instance, is responsible for at least 123 files in the labelled sample (0.37 % of the full data set). By extrapolation, we estimate that the site had between 655,516 and 1,023,603 files tagged with the site's URL stored on Megaupload's servers at the time of our experiment (a 99 % confidence interval).

However, these numbers are relatively modest when taking into account that OCHs such as Filesonic and Wupload, which were less popular than Megaupload during our measurements, received around one million uploads every day. There must have been many more (and potentially more active) actors who uploaded to Megaupload, but they are not distinguishable in our data set because they did not tag their uploads.



(a) Filefactory (FF) (b) Megaupload (MU)

Fig. 2. Overall classification as a function of the file size. The ■, ■ and ■ curves correspond to the fraction of *legitimate*, *infringing* and *unknown* labels among all files smaller than the current value on the x axis. For comparison, the ■ curve shows the file size CDF. Smaller files were more likely to be classified as *legitimate* whereas larger files were more likely classified as *infringing*. On FF, for instance, the point of an equal share of *legitimate* and *infringing* labels is for an upper file size limit of 200 MB.

6 Discussion

Our analysis provides approximated lower bounds for the proportion of legitimate and infringing files hosted on a range of OCHs. While these results suggest significant levels of copyright infringement on each of the OCHs, the question of whether the OCHs are actually *responsible* for these user uploads is a very different problem that we are not attempting to address in this paper.

We stress that our analysis does not aim at labelling one OCH as more compliant than another. Direct comparisons can be challenging because of subtle differences in how we collected our data. Furthermore, we did not specifically investigate which anti-abuse systems the OCHs had in place.

The present methodology was developed to estimate the prevalence of infringing uploads *after the fact*. It worked well with our data sets because of the relatively high numbers of rather explicit file names. This makes our methodology a bad fit for active upload filters: Many of the heuristics are trivial to circumvent for uploaders who have a reason to do so. Moreover, most of our attempts at automating the heuristics resulted in too many false positives, which ultimately forced us to resort to manual labelling.

There are known techniques that OCHs have at hand to limit abuse and copyright infringement on their systems. Blacklists based on file hashes are more promising than approaches using file names: An uploader would need to repack a file in order to circumvent a hash blacklist instead of simply renaming it. Furthermore, hash blacklists limit false positives, and OCHs could conveniently block access to *all* files with the same contents as soon as a complaint is received

Table 6. The most frequent URLs in the full data sets, seeded by the URLs found in the samples, together with the language of the web site. For MU, we also give the percentage of these files in our large random sample, which hints at how many files these sites have uploaded in MU's lifetime.

	Filefactory (FF)			Easy-share (ES)		
#	URL	Lang.	#/30 d	URL	Lang.	#/48 h
1	`myegy.com`	ar	4093	`electro-maniacs.net`	en	439
2	`0daymusic.org`	en	3656	`x-cornerz.com`	en	301
3	`mazika2day.com`	ar	2922	`pornlove.org`	n/a	275
	Filesonic (FS)			Wupload (WU)		
#	URL	Lang.	#/48 h	URL	Lang.	#/48 h
1	`hornyblog.org`	en	5126	`noor7.us`	en	7070
2	`4bookholic.com`	n/a	2010	`asiandramadownloads.com`	en	6100
3	`1-link.org`	en	1880	`hornyblog.org`	n/a	5093
	Megaupload (MU)			Undeadlink (UL)		
#	URL	Lang.	# (%)	URL	Lang.	#/7 m
1	`megauploadforum.net`	en	123 (.37)	`megaupload-download.net`	fr	2939
2	`x1949x.com`	zh	104 (.32)	`liensplv.com`	fr	1163
3	`hdtvshek.net`	ru	55 (.17)	`univers-anime.com`	fr	968

for one of them. Rapidshare recently took a more drastic measure by restricting the allowed download traffic per uploader [6], effectively precluding the use of its service for public sharing of popular content, infringing or not.

7 Conclusion

We conducted the first large-scale study that quantified copyright infringement in user uploads across five OCHs. Our results draw a mixed picture of both legitimate and infringing uses of OCHs. We classified 26 % to 79 % of the uploaded files as infringing copyright, with potentially more infringing files that we were not able to detect with our conservative and privacy-preserving methodology.

Overall, we were not able to classify between 21 % and 60 % of the files uploaded to the OCHs. That is, we do not know how many of these unclassified files are legitimate or potentially infringing. In the case of Megaupload, for instance, our methodology estimates the percentage of legitimate files as *at least* 4.3 % and *at most* 69.3 %, whereas potentially infringing files account for at least 31 % and at most 96 %. A goal for future work may be to provide a more precise estimation of the ratio between legitimate and infringing files. However, it remains unclear how this can be achieved in a privacy-preserving manner.

In our most conservative scenario, 4.3 % of the files hosted on Megaupload were detected as legitimate, which corresponds to approximately 10.75 million files. This quantity may appear relatively small compared to the 77.5 million files that we classified as potentially infringing, and even smaller compared to all the files we were not able to classify at all, yet it is quite large in absolute terms. It confirms the widely reported complaints of users who lost access to their files as a side-effect when Megaupload was forced to shut down.

Acknowledgements. This work was partially supported by Secure Business Austria, the NSF grant CNS-1116777, and the French ANR projects Aresa2 and PFlower. Engin Kirda thanks Sy and Laurie Sternberg for their generous support.

References

1. An estimate of infringing use of the Internet. Tech. rep., Envisional Ltd (Jan 2011), http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf
2. Antoniades, D., Markatos, E., Dovrolis, C.: One-click hosting services: A file-sharing hideout. In: IMC '09. ACM (Nov 2009)
3. Cuevas, R., Kryczka, M., Cuevas, A., Kaune, S., Guerrero, C., Rejaie, R.: Is content publishing in BitTorrent altruistic or profit-driven? In: Co-NEXT '10 (Nov 2010)
4. enigmax: Hotfile goes to war against copyright infringers. http://torrentfreak.com/hotfile-goes-to-war-against-copyright-infringers-110219/ (Feb 2011)
5. Ernesto: Hotfile's most downloaded files are open source software. http://torrentfreak.com/hotfiles-most-donwloaded-files-are-open-source-software-120411/ (Apr 2012)
6. Ernesto: Rapidshare limits public download traffic to drive away pirates. http://torrentfreak.com/rapidshare-limits-public-download-traffic-to-drive-away-pirates-121108/ (Nov 2012)
7. Jain, R.: The art of computer systems performance analysis: Techniques for experimental design, measurements, simulation, and modeling. Wiley (Apr 1991)
8. Jelveh, Z., Ross, K.: Profiting from filesharing: A measurement study of economic incentives in cyberlockers. In: P2P '12. IEEE (Sep 2012)
9. Kammerstetter, M., Platzer, C., Wondracek, G.: Vanity, cracks and malware: Insights into the anti-copy protection ecosystem. In: CCS '12. ACM (Oct 2012)
10. Kravets, D.: Feds tell Megaupload users to forget about their data. http://www.wired.com/threatlevel/2012/06/feds-megaupload-data/ (Jun 2012)
11. Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., Jahanian, F.: Internet inter-domain traffic. In: SIGCOMM '10. ACM (Aug 2010)
12. Lauinger, T., Kirda, E., Michiardi, P.: Paying for piracy? An analysis of one-click hosters' controversial reward schemes. In: RAID '12. Springer Verlag (Sep 2012)
13. Lauinger, T., Szydlowski, M., Onarlioglu, K., Wondracek, G., Kirda, E., Kruegel, C.: Clickonomics: Determining the effect of anti-piracy measures for one-click hosting. In: NDSS '13. Internet Society (Feb 2013)
14. Liang, J., Kumar, R., Xi, Y., Ross, K.: Pollution in P2P file sharing systems. In: INFOCOM '05. IEEE (Mar 2005)

15. Mahanti, A., Carlsson, N., Williamson, C.: Content sharing dynamics in the global file hosting landscape. In: MASCOTS '12. pp. 219–228. IEEE (Aug 2012)
16. Mahanti, A., Williamson, C., Carlsson, N., Arlitt, M., Mahanti, A.: Characterizing the file hosting ecosystem: A view from the edge. Performance Evaluation 68(11), 1085–1102 (Nov 2011)
17. Nikiforakis, N., Balduzzi, M., Acker, S.V., Joosen, W., Balzarotti, D.: Exposing the lack of privacy in file hosting services. In: LEET '11. Usenix (Mar 2011)
18. Rehn, A.: The politics of contraband: The honor economies of the warez scene. Journal of Socio-Economics 33(3), 359–374 (2004)
19. Sandoval, G.: MPAA wants more criminal cases brought against 'rogue' sites. http://news.cnet.com/8301-31001_3-57407346-261/mpaa-wants-more-criminal-cases-brought-against-rogue-sites/ (Mar 2012)
20. Sanjuàs-Cuxart, J., Barlet-Ros, P., Solé-Pareta, J.: Measurement based analysis of one-click file hosting services. Journal of Network and Systems Management (May 2011)
21. Watters, P.A., Layton, R., Dazeley, R.: How much material on BitTorrent is infringing content? A case study. Information Security Technical Report 16(2), 79–87 (May 2011)
22. Wilson, D.: Exclusive: Megaupload issues response to RIAA over Mastercard cutoff. http://www.zeropaid.com/news/91680/exclusive-megaupload-issues-response-to-riaa-over-mastercard-cutoff/ (Dec 2010)