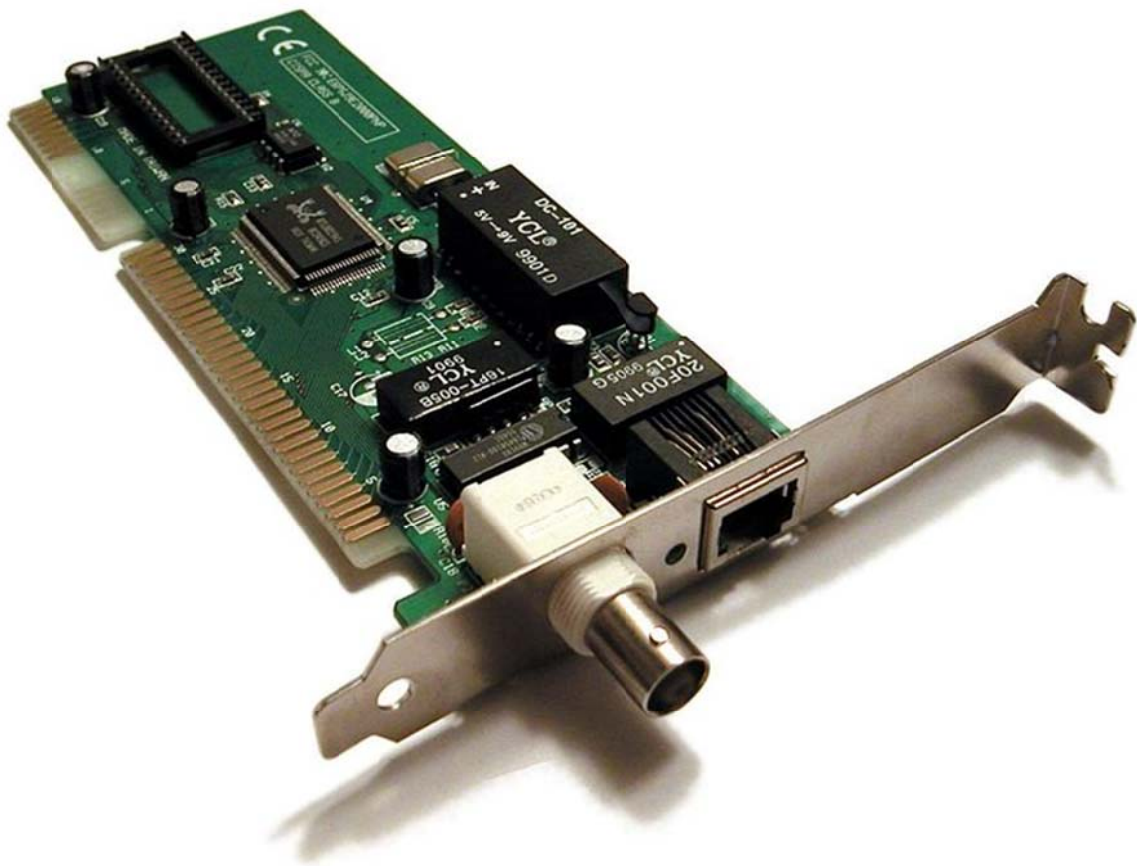




Determining Infringing Content on BitTorrent Networks: Enhancing Sampling and Detecting Fake Files



AUGUST 2011

**Robert Layton and Paul A. Watters
Internet Commerce Security Laboratory (ICSL)
University of Ballarat**

Executive Summary

At the ICSL, we have developed a methodology for systematically estimating the true extent of illegal file sharing on peer-to-peer (P2P) networks. Previous studies have found that the overwhelming majority of BitTorrent usage was for sharing copyright infringing content [1]. We have identified a clear power law relationship between torrents and downloads, meaning that a small number of very popular torrents account for a larger proportion of overall downloads. We also validated the results by cross-checking the relative proportions of categorised downloads (movies, music, software etc.) against keyword terms entered onto a popular torrent search engine site [2].

Multiple versions of many files, most notably movies, are available for download via BitTorrent. Further to that, there are millions of torrents on BitTorrent networks, released under a variety of circumstances and legalities. The focus of this report is on popular usage of BitTorrent. We look primarily at the top 1000 most downloaded and top 1000 most actively seeded torrents. To remove the effect of outliers and potentially biased numbers, we remove any torrents not occurring on at least three trackers we scraped and use the median value in our calculations. We further validate our findings by randomly sampling from all torrents collected.

In this report, we further refine and extend the methodology by introducing a second sampling technique for identifying popular and available trackers, and for identifying “fake” files which are being shared. In some cases, users may search for and attempt to download infringing content and the downloaded file is a fake, potentially containing malware. We factored this into the findings and found a large number of popular torrents appear to be faked. Further to this, we show a strong body of evidence that suggests that nearly all BitTorrent use is nefarious in nature, either faked files or copyright infringing.

The key findings of this report are:

- 50% of popular torrents appear to be faked files, either malware or incorrect files.
- Approximately 60% of popular torrents are movie based content.
- 97.9% of BitTorrent use in our samples is nefarious in nature, either faked files, copyright infringing or criminally infringing.
- 97.2% of the most popular “real” torrents (i.e. not faked files) are copyright infringing.

Table of Contents

1 . Introduction	4
2 . The BitTorrent Protocol	7
2.1 Terminology	7
3 . Methodology.....	8
3.1 Tracker Sampling.....	9
3.2 Scraping.....	9
3.3 Median Calculation and Filtering	10
3.4 Filename Determination	10
3.5 Categorisation	11
3.6 Fake Detection	11
3.6.1 Rule 1 – isoHunt voting	11
3.6.2 Rule 2 – Torrentz voting.....	12
3.6.3 Rule 3 – Manual Analysis	12
3.7 Infringement Determination.....	12
4 . Results	13
4.1 Tracker Selection.....	13
4.2 Torrent Scraping.....	14
4.3 Median Calculation	14
4.4 Filename Determination	14
4.5 Categorisation	15
4.6 Fake Detection	16
4.6.1 Categories within the real files	17
4.6.2 Category Separated Fake Detection	20
4.7 Copyright Infringement.....	21
5 . Criticisms of previous studies.....	23
6 . Discussion.....	25
6.1 Acknowledgements.....	26
7 . References.....	27

1 Introduction

The downloading of infringing content continues to be a major problem for creative industries worldwide, as the piracy of high value content drains the funds available for investing in new artistic ventures. The widespread utilisation of P2P and other technologies (such as cyberlockers) for sharing infringing content acts as a major disincentive for investment in the arts, as well as fuelling criminal enterprises that are linked to the funding of terrorism [3].

Previous studies [1, 2, 4] have attempted to determine the overall rate of infringing content being shared on P2P networks. There is nothing intrinsically “bad” about P2P technology, however its highly distributed nature has made it attractive to pirates. One reason for this is the difficulty (but not impossible) to track and trace all of the downloaders of torrents and the files to which torrent metadata is linked.

In some ways, quantifying the volume of infringing content appears to be a straightforward proposition: the centralised trackers around which P2P protocols (like BitTorrent) are built provide all of the information that is required. However, through our research program, we have identified a number of integrity and statistical issues that have required research to provide better results, including:

- *Sampling.* Making inferences about any population from a limited sample is always dangerous because of the potential for bias. Given the enormous number of hosts connected to the Internet, many of which will use BitTorrent, it is infeasible to monitor all activity. Therefore, we focus on sampling activity on the most popular trackers. We use two techniques to select samples for measurement.
- *Fake download data.* Through the BitTorrent protocol, trackers provide information on how many completed downloads have been registered with a tracker. Since the download counts are not an essential part of the tracking process, there is the potential for trackers to fake this data without negatively affecting their operation. However unless all trackers are in collusion (possible but unlikely), an effective sampling strategy based on multiple observations can overcome the problem.
- *Double counting.* During and after downloading, a client may interact with a number of trackers, so to avoid double counting of completed downloads, we take the median figure for each torrent for each tracker that has registered a download. As the breakdown point for the median is the theoretical maximum of 0.5 [12]¹, this is the fairest and most statistically reliable technique.
- *Fake content.* For whatever reason – disincentive to download or maliciousness by other users – a number of torrents do not actually contain the content which they claim to. In many cases, it may seem obvious – for example, movie torrents appearing before a film has even screened at the cinema – but nonetheless, faking is a significant issue. We present a three-stage fake detection process in this report but also indicate that there is room for significant improvement due to the “brittleness” of any rule-based system.

Determining the authenticity of files on P2P networks is an important issue from a cost and efficiency perspective. For example, users who are paying “per megabyte” for each file downloaded (and/or uploaded) would waste significant amounts of their download limit for fake files, often reaching over 700 megabytes in size [5]. There is also evidence that fake files are associated with virus

¹ The breakdown point is the percentage of values needing to be changed to change a result to an arbitrary value. The median has the highest possible breakdown value, indicating that 50% of the values need to be changed to change the median to an arbitrary value. In contrast, only one value needs to be changed to change the mean to an arbitrary value. [12]

infection [6], Trojan horse installation [7] (e.g., by requiring the user to install “video player” software) and even “trolling” where malicious users deliberately upload fake movies to frustrate the downloading process. While it is certainly plausible that fake files are uploaded by many users, the enormous popularity of P2P networking and download rates indicate that users are finding genuine material.

In response to concerns about authenticity, P2P communities (such as isoHunt.com) have built communities [8] around P2P downloading, where registered users are able to download torrents, retrieve the associated files, and then post *ratings* of these files to the website. This feature allows other users to review the ratings and decide whether or not to download that file. Since searching by movie name on the isoHunt.com site allows users to retrieve multiple torrent files, each with their own rating, users can typically discriminate between fake and authentic files. Furthermore, each registered reviewer also has a *reputation* score which can be used to further validate the scores provided by each reviewer, e.g., if one reviewer has a reputation score greater than another, then users would tend to *trust* the former reviewer’s ratings compared to the latter. The theory behind the effectiveness of these ratings schemes was developed by Tran et al [9].

As an example, consider the following search on the isoHunt.com site:

<http://isoHunt.com/torrents/jaybob?iht=-1&ihs1=11&iho1=d&age=30>

This URL searches for all torrents with the term “Jaybob” that have been uploaded in the last 30 days. Why choose the term “Jaybob” rather than a movie name? “Jaybob” is typically the top search term as reported in the isoHunt.com *zeitgeist* function [8] and therefore DVD rips originating from “Jaybob” will be likely targets for both genuine downloaders and fake torrent submitters. Executing this search on 22/03/2011 generated 31 hits, with the last 10 results (ordered by number of seeders) summarised in Table 1.

In most cases, it seems clear that the “rating” can be used to define a simple decision rule that separates fake from genuine titles. Take “Triple Dog {2010} DVDRIP. Jaybob” as an example; this is a genuine torrent where 5 people clicked “thanks” and rated it as genuine; sometimes, though, users leave comments without indicating whether the file is genuine or not. This is the case with “Legacy Black Ops {2010} DVDRIP. Jaybob”, where 6 users have left comments but no rating – this occurred because 6 users have initialised the downloading process, and the torrent was only 7 hours old.

Table 1. Fake and Genuine “Jaybob” Torrents from isoHunt.com

Torrent Name	Rating	Number of Comments
House Of Bones {2010} DVDRIP. Jaybob	+7	6
A Modern Cinderella Tale {2010} DVDRIP. Jaybob	+10	6
Twisted Path {2010} DVDRIP. Jaybob	+5	9
Triple Dog {2010} DVDRIP. Jaybob	+5	10
Legacy Black Ops {2010} DVDRIP. Jaybob	-	6
The Mechanic (2011)	-1	5
Burlesque (2011)	-2	3
New Moon	-2	3
Burlesque (2011) ENG-Jaybob	-4	4
Burlesque (2011) ENG-Jaybob	-4	4

Alternatively, looking at a fake like “Burlesque (2011) ENG-Jaybob”, 4 users flagged the file as fake, and each user also left explanatory comments, e.g., “ptolemy101” writes “Fake. rar folder with a password.”, indicating that the file might be a troll, or contain malicious code, or require payment to open the .rar format archive file. One of the site moderators “WhiteViper” also notes that some low-reputation users may flag a fake file as genuine, and that these votes are not “genuine”. Despite this, ratings appear to be generally internally consistent for each torrent, indicating a relatively high degree of consensus for each. This means that a simple consensus ensemble technique like voting [10] can be used to sum the different votes being cast by the users, without further need for weighting, although the reputation scores for each rater are certainly available. It should also be noted that the comments for each torrent in themselves comprise a rich data source, e.g., although the fields are free-form, some semi-structured data is available for processing such as a quality rating for audio and video (usually expressed in the form /A and /V respectively). Further mining this rich data source remains an interesting possibility for deeper research into the judgements being made by users.

In this report, we present the current version of our methodology, and present the results for trackers scraped in May 2011, just over 12 months since our first report [1]. We provide a consolidated description of each process stage, with sufficient detail allowing any other *bona fide* researcher to replicate our findings.

2 The BitTorrent Protocol

The BitTorrent protocol (hereafter just BitTorrent) is a peer to peer file sharing protocol, allowing files to be shared to a large number of users without a drastic impact on a single server providing the download. BitTorrent is especially popular for sharing very large files, often hundreds of megabytes or larger. It was developed in 2001 by Bram Cohen and is now maintained by his company BitTorrent, Inc. BitTorrent outlines a method for users (peers) to download a pieces of a set of files from other peers who have different pieces. Combining those pieces leads to the recreation of the original set of files. Tracking which peers have which pieces is managed by a server called a 'tracker', which is the first contact point for a new peer to join the network. Newer versions of the protocol do not need a tracker to operate, however the focus of this report is on BitTorrent trackers in use.

In the BitTorrent specification, a client is a person wanting to download a file using BitTorrent. In order to do this, they obtain a *torrent* file which contains information on how to download their file. As an example, they may use a BitTorrent search engine to find a torrent file for a movie. The torrent file itself does not contain the movie itself, only information on how to download the movie. The torrent file contains information on which files it contains information for, how many pieces are needed to recreate the file and the SHA-1 hash of each piece enabling a client to ensure that the pieces they have downloaded are valid and have not been tampered with.

Torrent files also contain information on which trackers to use for finding peers and pieces. A tracker is a server which maintains a list of clients that have connected to the tracker, remembers which files the clients are downloading and which pieces of each file they have. When a new client connects wanting a movie, the client calculates an identifier for the torrent, known as the *info hash*. The tracker searches its database for this identifier and returns all of the other clients that have the same file that is being requested. The tracker returns the IP address of each of these clients, called *peers*, to the new client. At this stage, the new client contacts each of these peers and requests to download a piece of the movie from them.

A tracker may also return a list of all torrents on its system as well as the number of times it has been downloaded and information on the total number of people currently sharing a file. This information is called a full scrape, and is the basis for the methodology and results in this report. A more detailed description of BitTorrent and the methods for performing a scrape are available in [1].

2.1 Terminology

This report uses the following terminology:

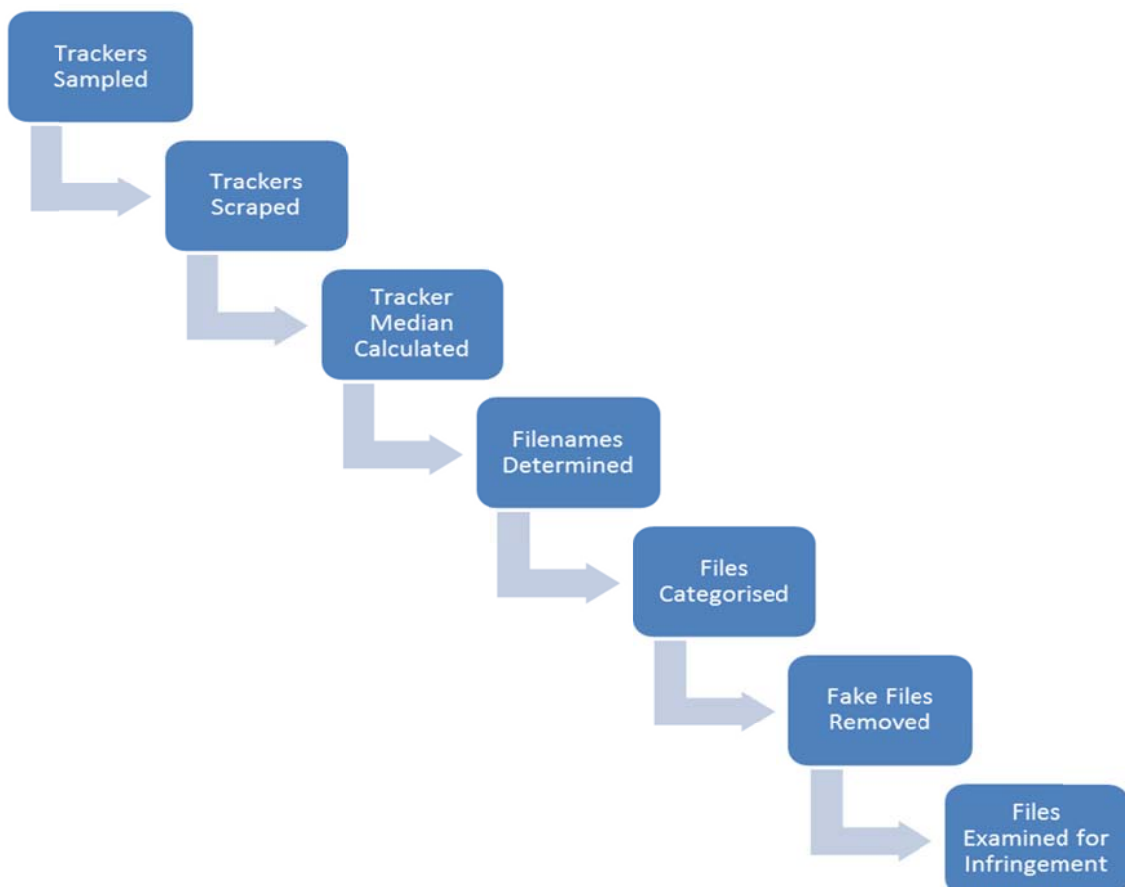
- **downloaded:** The number of times a torrent has been downloaded and registered by the tracker. This can be seen as the *all-time count* of downloads of a file.
- **complete:** The number of clients currently sharing a file who have downloaded the complete file. These clients are known as seeders, sharing the file without needing to download any more pieces of it. This can be seen as the level of *active sharing* of the torrent at the time of measurement.
- **seeder:** A single client currently sharing a single file. Used when referring to the connections counted in the complete figure mentioned above. A single client can be responsible for multiple "seeders", if they are sharing multiple files. However a single client sharing a single file on multiple trackers will still be a single seeder.
- **info hash:** The SHA-1 hash of the information section in the torrent (see [1])
- **hex hash:** A hex-encoding of the info hash, used as a human-readable version of the info hash (see [1]).

3 Methodology

Our methodology is observational in nature, meaning that we intend to observe the behaviour of downloaders in a systematic way, without intruding on or manipulating their activity. The observational process involves scraping trackers, recording the resulting scrape and analysis the resulting scrapes. This provides an objective way of understanding BitTorrent usage through the trackers, rather than relying on using a sampling of available torrents. As some torrents will be downloaded more (or less) frequently than others, this report focuses on popular usage of BitTorrent. The methodology also focuses on tracker based BitTorrent usage and does not estimate trackerless downloads of any kind, including those using Distributed Hash Tables (DHT), Peer Exchange (PEX), web seeding etc. However, given that trackers are still very widely used by BitTorrent clients, they provide one route to obtain a view on the number and relative proportions of downloads.

The methodology works in seven stages:

1. Trackers are representatively sampled to reduce bias from any one tracker (Section 3.1)
2. The tracker sample is then scraped (Section 3.2)
3. The median download figure is calculated for each torrent appearing on at least three trackers (Section 3.3)
4. Filenames are determined for each remaining torrent (Section 3.4)
5. Categorisation of the torrents is performed (Section 3.5)
6. Fake detection on the torrents performed and categorisation on only the real files was reperformed (Section 3.6)
7. The number and proportion of files infringing copyright was determined on real files (Section 3.7)



As the focus of this report is on popular usage of BitTorrent, we focus our investigations onto the lists of the top 1,000 complete and downloaded files. After performing fake file detection we update this list to be the top 1,000 complete and downloaded *real* files. To examine the impact of popular usage on categories and infringement levels, we also take a random sample of *real* torrents meeting the median criteria (Section 3.3) to compare in stages 5 and 7. We describe the design of each stage for this study below.

3.1 Tracker Sampling

To obtain the most representative results from a sample, in relation to the most popular downloads, it is important to extract data from the most popular trackers in use. In our previous studies [1, 2], we sampled the 10 most popular torrents on the website Torrentz (<http://www.torrentz.com/>). All trackers listed for each of these files was then selected for sampling. In this report, we adopted a two-pronged approach to broaden the sampling base, to reduce any bias that might have resulted from using Torrentz:

- We obtained a list of “live” trackers from the meta-tracker Trackon website (<http://www.trackon.org/>), and attempted to perform a scrape on each tracker.
- We used the isoHunt “zeitgeist” (<http://ca.isohunt.com/stats.php?mode=zg>) service to identify the Top 10 current torrent search terms, and we extracted the tracker list from the most popular search result under each term.

Thus, we have obtained a tracker list that reflects both availability (trackon) and popularity (isoHunt).

3.2 Scraping

Once the tracker sample was determined a full scrape was attempted for each tracker. This scrape was downloaded in a similar way to a normal HTTP download. If the download was interrupted, the scrape was not attempted again in that iteration. An interrupted download could still be useful, however, as it would contain valid scrape information up to the end of the downloaded portion. For example, if a scrape was interrupted after downloading 80% of the file, there would still be 80% of the scrape information available. When parsing the scrape data, the consistency of the file was not verified to ensure that information could be gathered from interrupted downloads. Rather, any valid data for each file was collected and saved into a database.

The information collected included, according to the official scrape convention in the BitTorrent specification [11]:

- the info hash
- the number of peers currently sharing the whole file, after downloading it (“complete”)
- the number of times the tracker has registered a complete download for a client using “event=complete” (“downloaded”)

For some trackers, we observed considerable variance in the reported “complete” values. In some cases, the trackers we retrieved data from indicated that all files had been “downloaded” only 10 times, even when the number of current seeders (“complete”) was in the thousands. This is plausible, but unlikely as it suggests a large number of people downloaded the file from an alternate source before adding the new tracker to seed the file.

Conversely, it is also possible that many clients who complete a download (counted in “downloaded”) then remove the torrent after downloading and do not seed the file (and thus would not be counted in “complete”). Therefore, rather than assuming one or more count to be less reliable or indicative, we present the results for both the “complete” and “downloaded” files.

3.3 Median Calculation and Filtering

As noted, there is a potential for faked figures to be returned by trackers. To compensate for this, we take the median figure of the torrent downloaded and complete figures. While it is possible that other figures such as the maximum would be higher and still valid, we take the more reliable option and use the median figure instead.

To ensure that this value is reliable, a torrent is only considered to appear on a tracker if the count (either downloaded or complete) is more than zero. Some trackers have an entry for some torrents but do not have statistics available for them and these are ignored in the calculation. After that, only torrents which appear on three or more trackers are included in the results. This requirement of available statistics on three or more trackers is called the *median criteria* in the rest of the report. All figures used in the rest of the report are medians of the downloaded or complete figure for all trackers with the given file, unless otherwise specified.

3.4 Filename Determination

As the procedure that calculates the info hash is a one-way function, we could not recreate the filename from the scrape data alone. However, by querying external data sources, it is possible to correlate the info hashes with file titles. One of the advantages we have here is that – like searching for Internet pornography – users need to search for terms of interest, and search engines thus provide a convenient means to perform reverse lookups [12]. For example, searching for the hex hash value “9064267d4a83e096e6eb14593762bc18633eda0f” returns “Avatar 2009 720p BluRay x264” as the filename. Verification of this method can be performed by downloading the returned torrent and recalculating the hex hash.

To determine the filename, we used both a BitTorrent search engine (<http://Torrentz.eu/>) and Google. The procedure started by searching the BitTorrent search engine for the hex hash. If the BitTorrent search engine had the torrent that generated this info hash, it would return the torrent, including the names of the files contained in it. We then parsed the search results to extract only the filename, and stored the resulting filename in the database.

If this procedure failed, we performed a Google search for the hex encoded info hash. If results were returned from Google, we ranked them in order of appearance. If the title of the search result (i.e., the title of the corresponding webpage found by Google) included the hex hash, it was ignored, as many websites repeat this value in their title, giving a null result. If the hex hash was not in the title, we used the title as our filename result. In some cases, the filename was “dirty”, as the title of the search result was likely to contain other information such as the name of the website linked to by Google. A full parsing of the returned results remains a significant problem for automatic parsing, and was considered out of scope for this methodology. In many of these cases, the filename still appeared and could be manually extracted.

To determine the accuracy of the filename determination procedure, the results were verified by performing a reverse lookup. To do this, we selected the top 50 seeded torrents with filenames, and a random sample of 50 torrents from the full set of named torrents, as our test set. For each of these 100 torrents, the original torrent file was searched for, using the given info hash. The torrent file was then downloaded, and the info hash re-calculated to verify that the torrent was correct. This sampling method was chosen to ensure that there were no biases between the top torrents, compared to a representative sample of the full set of named torrents.

3.5 Categorisation

After the filenames were determined, category determination was performed. Category determination was easier for some files and categories than others. For example, most movies are of the form:

```
<movie-title> (<year>) <release-group>
```

in which fields can be separated by spaces, periods or other characters. This format changes a little bit as well between release groups and sometimes is a different format altogether. Another common pattern is:

```
S<season-number>E<episode-number>
```

which is used for TV shows to indicate which episode is available. An example of this would be:

```
The.Simpsons.S10E04
```

to indicate the fourth episode of the tenth season of the TV show “The Simpsons”.

To perform automatic categorisation, we use a simple rule based system. A list of patterns, in the form of regular expressions, was listed along with the category they corresponded to. To categorise a torrent, each rule in order was applied to the file. Once a rule was triggered, which happened when the filename contained the pattern given by the regular expression, the file was assigned the category from the rule, and the matching procedure would stop. If no rule was triggered, the file was manually checked to determine the category.

To verify the results all categorisations reported were manually verified, and those without a verification had one manually assigned. In cases where the category is ambiguous, the category was listed as UNKNOWN and statistics for this “category” are also given.

3.6 Fake Detection

In addition to the infringement determination, we also assessed whether each file was fake (or not), in the top 1,000 “downloaded” and “complete” torrents. For each torrent in each sample, the following fake detection algorithm was used to determine whether a file could be classified as:

1. “real”, meaning the torrent was authentic
2. “fake”, meaning the torrent was not authentic
3. “not found”, meaning that the torrent could not be located in the isoHunt.com database, or
4. “failed”, a critical error occurred while performing the fake detection.

For each torrent, a three-stage, rule-based process was followed (isoHunt voting, Torrentz voting and manual analysis). After the fake detection is run, a new top 1000 downloaded and complete list is created composed of just real files for use in step 7 – infringement detection.

3.6.1 Rule 1 – isoHunt voting

1. A search for the torrent hash was made against the isoHunt.com database using the appropriate URL.
2. If the torrent hash was not found, then the torrent was flagged as “not found”, and Rule 2 was invoked.

3. If the torrent hash was found, then the torrent details page was retrieved. This page was parsed to retrieve the number of times the torrent was flagged by unique users as fake (F), and/or genuine (G).
4. The rating (R) of the torrent was given by the sum of F and G (F being a negative number)
5. If $R < 0$, the torrent was flagged as “fake”.
6. If $R > 0$, the torrent was flagged as “real”.
7. If $R = 0$, Rule 2 was invoked.
8. If a network exception occurred during the testing, the torrent was flagged as “failed”.

3.6.2 Rule 2 – Torrentz voting

1. A search for the torrent hash was made against the Torrentz.eu database using the appropriate URL.
2. If the torrent hash was not found, then the torrent was flagged as “not found”.
3. If the torrent hash was found, then the torrent details page was retrieved. This page was parsed to retrieve the number of times the torrent was flagged by unique users as fake (F), and/or genuine (G).
4. The rating (R) of the torrent was given by the sum of F and G (F being a negative number)
5. If $R < 0$, the torrent was flagged as “fake”.
6. If $R > 0$, the torrent was flagged as “real”.
7. If $R = 0$, Rule 3 was invoked.
8. If a network exception occurred during the testing, the torrent was flagged as “failed”.

3.6.3 Rule 3 – Manual Analysis

For the remaining files, a manual analysis was performed. This was performed by the authors who searched for the origin of the file to determine if the file was fake.

3.7 Infringement Determination

Once the top 1000 downloaded and complete real files were determined, they are manually checked to determine if they are copyright infringing or not. This determination was primarily based on the title of the file. There were two key limitations to the procedure: firstly, we took the filename at face value, and secondly, if there was any ambiguity in the filename, we erred on the side of caution, and guess that it is legal. We assert that the filenames are mostly accurate based on both their popularity and the fact that the fake file detection did not find they are faked. Files with incorrect filenames are likely to be discovered by the fake file detection in step 6. We cannot download the files to manually check them due to our legal requirements that we have – as researchers – not to infringe copyright. We counterbalance this by being extremely conservative in infringement determinations (any ambiguity leads to the categorisation of the torrent as unknown legality), and as the results indicate, this still leaves little doubt as to the overall pattern of infringement.

4 Results

The results below are provided for our original study, a follow-up study, and a method for validation.

4.1 Tracker Selection

The trackers scraped for this report are shown in Table 4.1.1, including whether or not a scrape was successfully returned, in order of the scrape size returned. 20 out of 41 trackers successfully returned a scrape.

Table 4.1.1 – Tracker Scrapes

Tracker	Usable Scrape?
artificial.intelligence.tracker.prq.to	Yes
gemini.tracker.prq.to	Yes
dances.with.wolves.tracker.prq.to	Yes
ipv4.tracker.harry.lu	Yes
a.tracker.prq.to	Yes
armaggedon.tracker.prq.to	Yes
cstv.tv.tracker.prq.to	Yes
tpb.tracker.prq.to	Yes
tracker.prq.to	Yes
exodus.1337x.org:80	Yes
tracker.ilibr.org:6969	Yes
bt.firebit.org:2710	Yes
tracker.torrent.to:2710	Yes
tracker2.istole.it:80	Yes
tracker.torrentbay.to:6969	Yes
tracker.podtropolis.com:2711	Yes
tracker.torrentbox.com:2710	Yes
papaja.v2v.cc:6970	Yes
cpleft.com:2710	Yes
tracker.zokio.net:8080	Yes
9.rarbg.com:2710	No
ck3r.org	No
amvnews.ru	No
ix3.rutracker.net	No
bt.uamedia.info	No
exodus.desync.com:6969	No
eztv.tracker.prq.to	No
ipv4.tracker.harry.lu.nyud.net	No
scrape.opensharing.org:2710	No
scrape.opensharing.ru:2710	No
scrape.xxx-tracker.com:2710	No
amplicate.appspot.com	No
anisource.spb.ru	No
antidenim.appspot.com	No
atrack.pow7.com	No
beeretracker.ru	No
btrs-atrack.appspot.com	No
dotehwerk.appspot.com	No
fordred1.appspot.com	No
jonnyfavorite1.appspot.com	No
tekfu-track.appspot.com	No

4.2 *Torrent Scraping*

Trackers might disallow scrapes because of a lack of bandwidth, or to prevent exhaustive searching against the torrents that they are tracking. A smaller tracker may wish to minimise their bandwidth usage by disabling this feature. Those trackers not returning scrapes are obviously not included in the following results.

Some gross statistics from the sample are given below for the unfiltered set. These figures are given for illustrative purposes only, as they do contain fake files and potentially faked figures reported by some trackers.

- Total torrent entries in sample: 29,186,794
- Total unique torrents in sample: 4,065,114
- Total “complete” counts including duplicates: 141,004,488
- Total “downloaded” counts including duplicates: 2,877,943,586
- Ratio of unique to total torrents: 0.1393
- Ratio of total torrents to “complete” files: 0.2070
- Ratio of total torrents to “downloaded” files: 0.0101
- Ratio of unique torrents to “complete” files: 0.0288
- Ratio of unique torrents to “downloaded” files: 0.0014

4.3 *Median Calculation*

Scraping the trackers resulted in a total sample of 29,186,794 torrents, of which 4,065,114 were unique. After filtering to include only torrents on at least 3 trackers, there were 1,413,960 unique torrents passing the criteria for inclusion noted in section 2.3 using downloads and 1,742,877 using complete figures. Within this filtered set of torrents, there was a total of 48,838,558 downloads and 11,181,570 complete (seeders). This figure excludes duplicates (as the median is used) but contains fake files. All figures from this point on include only the median filtered set and all figures use just the median downloaded or complete value for each torrent.

4.4 *Filename Determination*

We then selected the top 1,000 “complete” and “downloaded” torrents and determined their filenames. Previously, we found that the ranking of torrent popularity would follow a power law, i.e., relatively few torrents would account for the largest proportion of downloads. In this study, we found that just 1.0% of torrents (a total of 14,305) were responsible for 80% of the “downloaded” figures and 15.6% of torrents (a total of 272,420) were responsible for 80% of the “complete” figures. For the rest of the document, we use figures from the top 1000 downloaded and complete torrents. The top 1000 downloaded accounts for 37.7% of the total median download figure (18,403,222 downloads) while the top 1000 complete torrents accounts for 22.9% of the total median complete figure (2,566,126 seeders).

Out of all attempts to determine the filename of the Top Downloaded list, 979 succeeded and resulted in a filename being assigned. For the Top Complete list, 962 succeeded with an assigned filename.

Both lists of top 1000 were manually checked to ensure the accuracy of the filename validation. In both lists, a number of results were incorrect, labelled as “Torrent Search Engine” which is a returned result from a Google search rather than a filename. Another set of mistakes occurred when the hash itself was returned as the filename. In both of these cases, the entries were listed as failed filename determination attempts in the above results, but were manually categorised in the next step of the methodology.

4.5 Categorisation

The categorisation was performed using a set of manually derived rules as previously described [1, 2] followed by a manual investigation of the files with no given category. Categorisation was performed on both the top 1000 downloads and complete lists. Of these torrents, 346 were automatically categorised from the top 1000 downloaded and 576 were automatically categorised from the top 1000 complete. All categorisations from both lists were then manually verified and any torrent without a categorisation was given one manually if possible. There were a small number of categorisation errors, which are noted in the results. The percentages of files in each category are given in tables 4.5.1 and 4.5.2. Adjusted percentages are given including only torrents given a filename.

Table 4.5.1 – Category groupings of the top 1000 downloaded (Total downloads)

Category	Number	Adjusted %
Game	1	0.1%
Movie	336	34.3%
Music	1	0.1%
Porn	410	41.9%
Software	22	2.2%
TV Shows	48	4.9%
Unknown	157	16.0%
Errors	4	0.4%
Total	979	100.0%

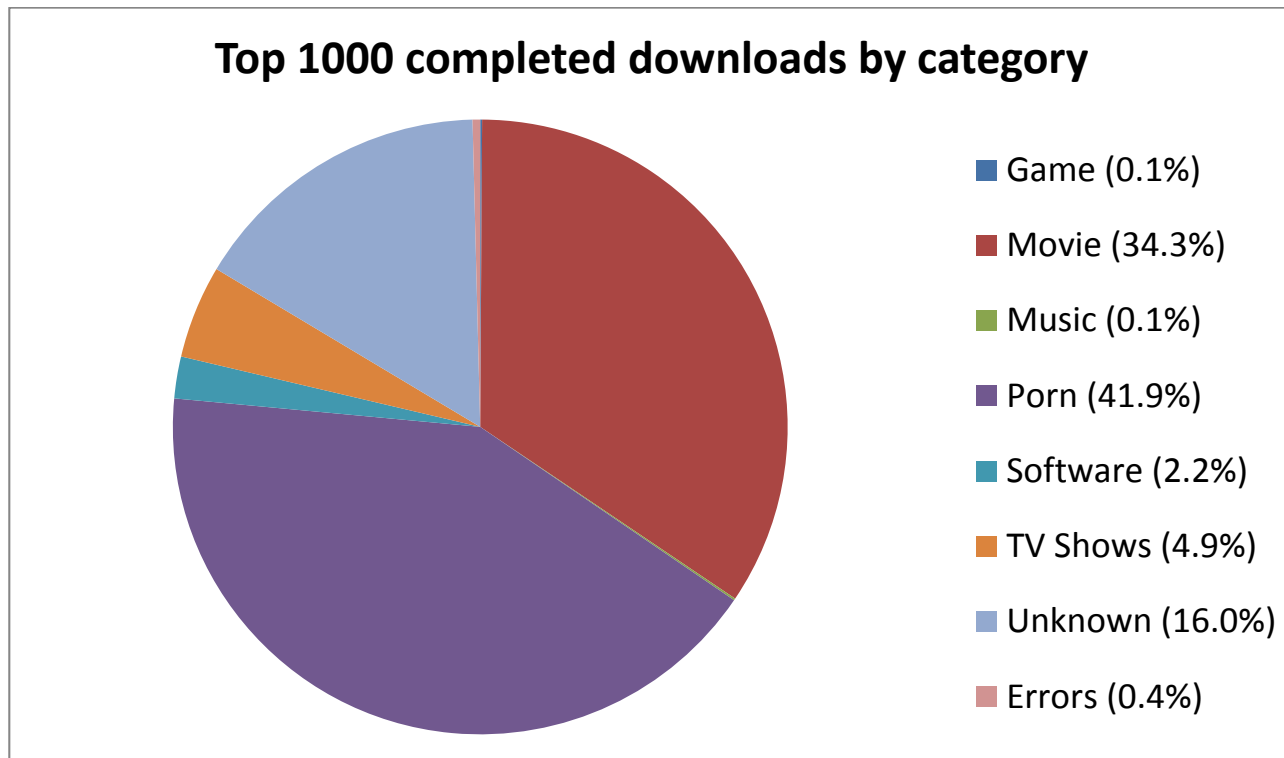
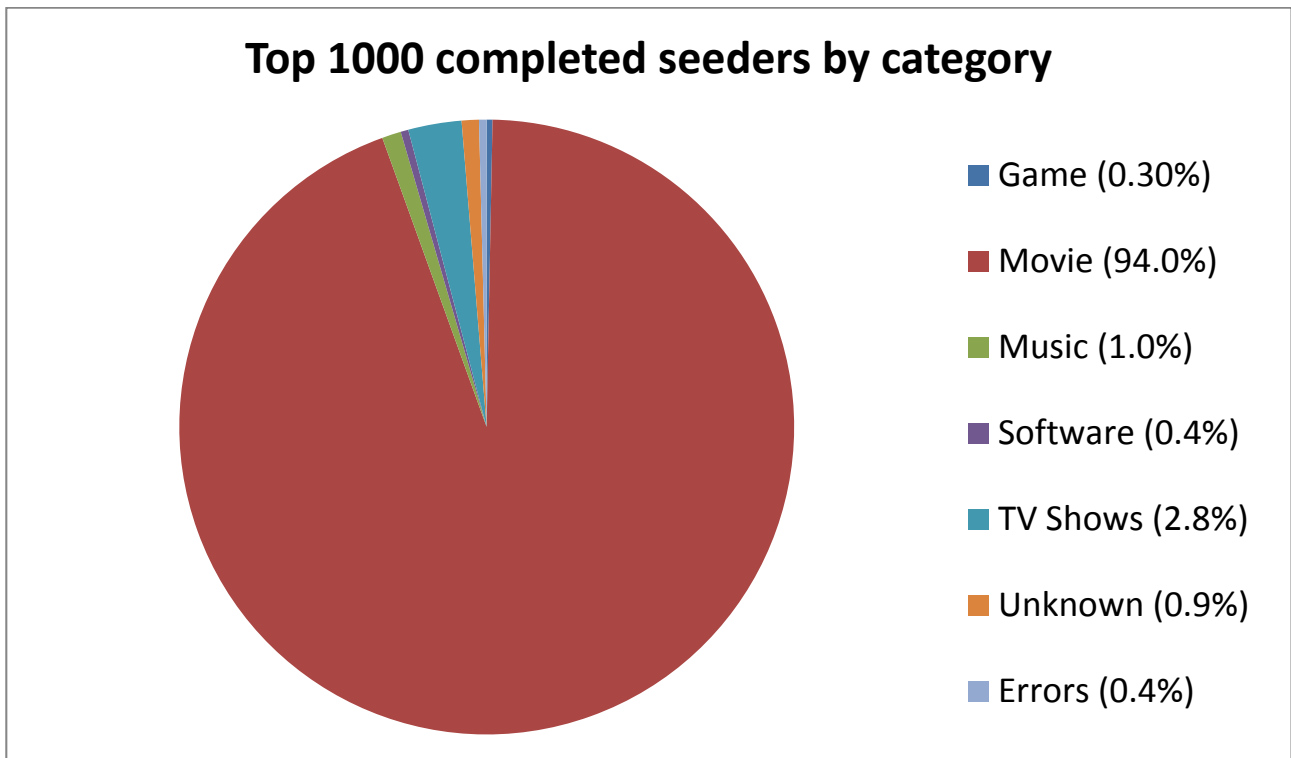


Table 4.5.2 – Category groupings of the top 1000 complete (Seeders)

Category	Number	Adjusted %
Game	3	0.30%
Movie	905	94.0%
Music	10	1.0%
Software	4	0.4%
TV Shows	27	2.8%
Unknown	9	0.9%
Errors	4	0.4%
Total	962	100.0%



4.6 Fake Detection

The results for each data set are presented below, in two ways: one, where the total for *F* and *G* are corrected for torrents not being found in the isoHunt.com database or where a network error occurred, and raw scores where no correction has been made. Thus, we can report overall levels of fake and genuine files where corrected for torrents not being found and/or not being available.

Tables 4.6.1 and 4.6.2 show the results for the Top 1,000 “downloaded” and “seeded” torrents respectively. For the “downloaded” torrents, the adjusted proportion of genuine files was 50.3% for the most torrents of the most downloaded files (adjusted percentages the same).

Table 4.6.1 – Results for the Top 1,000 “downloaded” torrents

Accuracy	Genuine	Fake	Failed	Not Found
Observed	502	498	0	0
Adjusted %	50.2%	49.8%		

For the “complete” torrents, the adjusted proportion of genuine files was 45.0% for the most torrents of the most seeded files.

Table 4.6.2 – Results for the Top 1,000 “complete” torrents

Accuracy	Genuine	Fake	Failed	Not Found
Observed	451	548	1	0
Adjusted %	45.1%	54.9%		

Figures 4.6.1 and 4.6.2 show the frequency distribution for all of the ratings, overlaid with the normal distribution “downloaded” and “seeded” torrents respectively.

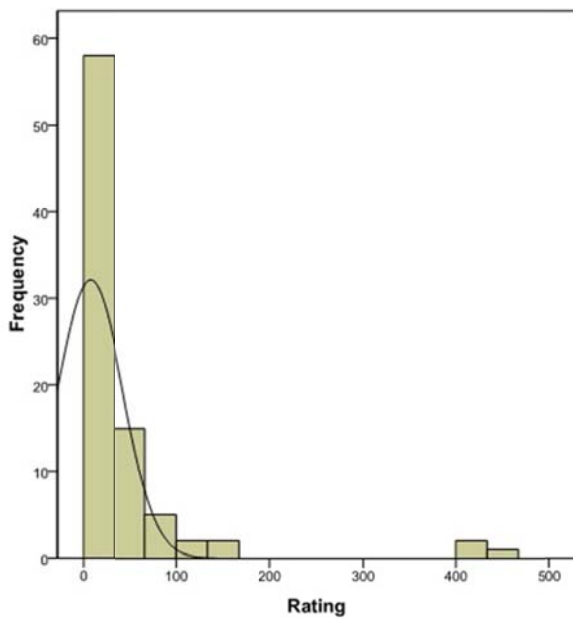


Figure 4.6.1 – Rating frequency distribution for the Top 1000 downloaded

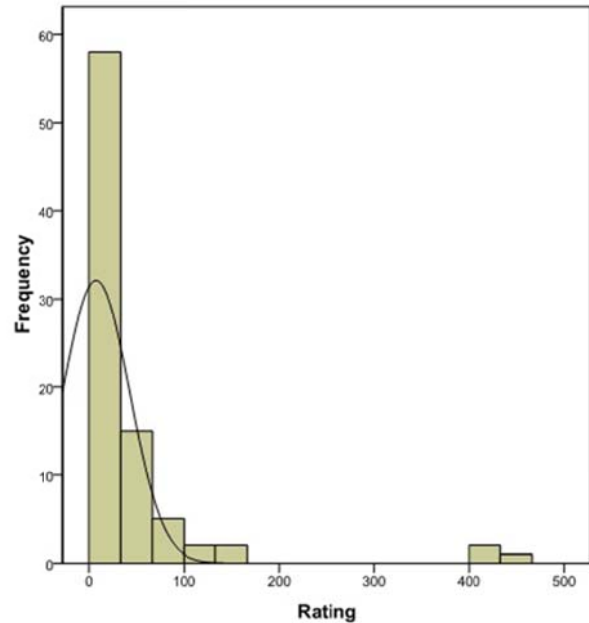


Figure 4.6.2 – Rating frequency distribution for the Top 1000 complete

4.6.1 Categories within the real files

The above figures were given for those torrents reported to have the highest median download or complete count, regardless of whether they are real or fake file. Analysis of the top 1000 downloaded/complete *real* files provides a different view on the data, with a category breakdown given in tables 4.6.3 and 4.6.4. There is a large reduction in the overall percentage of real files which were movies, with mostly moderate gains in other categories. These results confirm what was found in the previous section, where movies were highly targeted for fakes files and found a large number of downloads. TV Shows and Music gained overall, mostly in the complete count. The compare percentage in both tables is the percentage from the lists including faked files.

Table 4.6.3 – Category groupings of the top 1000 downloaded real files

Category	Number	Adjusted %	Compare %
Game	6	0.6%	0.1%
Movie	373	37.3%	34.3%
Music	26	2.6%	0.1%
Porn	512	51.2%	41.9%
Software	25	2.5%	2.2%
TV Shows	55	5.5%	4.9%
Unknown	3	0.3%	16.0%
Errors	0	0.0%	0.4%
Total	1000	100.0%	100.00%

**Top 1000 completed downloads by category
(fakes excluded)**

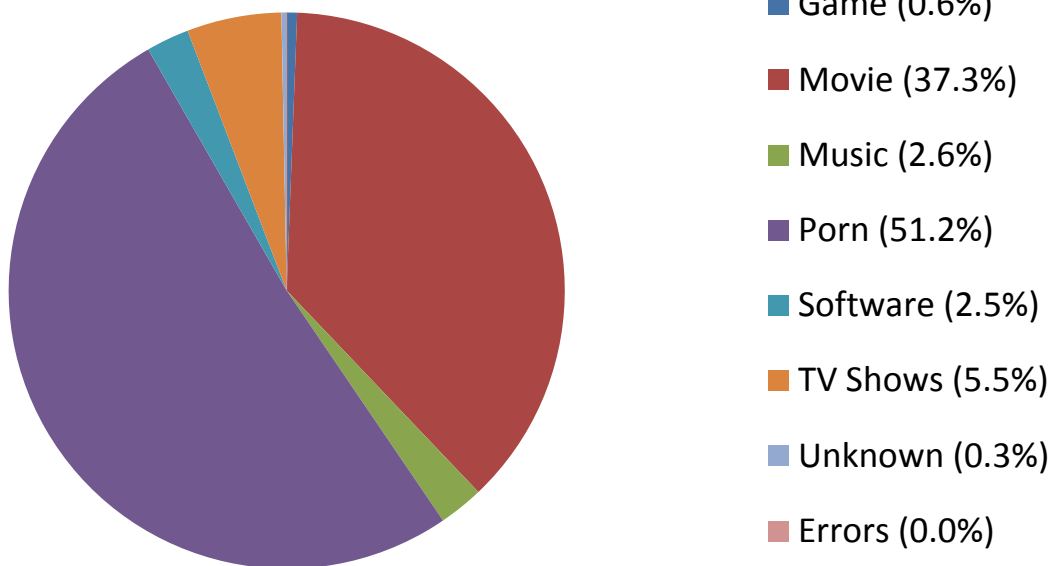


Table 4.6.4– Category groupings of the top 1000 complete real files

Category	Number	Adjusted %	Compare %
Game	34	3.4%	0.30%
Movie	694	69.4%	94.0%
Music	103	10.3%	1.0%
Software	28	2.8%	0.4%
TV Shows	109	10.9%	2.8%
Porn	16	1.6%	0.0%
Unknown	16	1.6%	0.9%
Errors	0	0.0%	0.4%
Total	1000	100.0%	100.0%

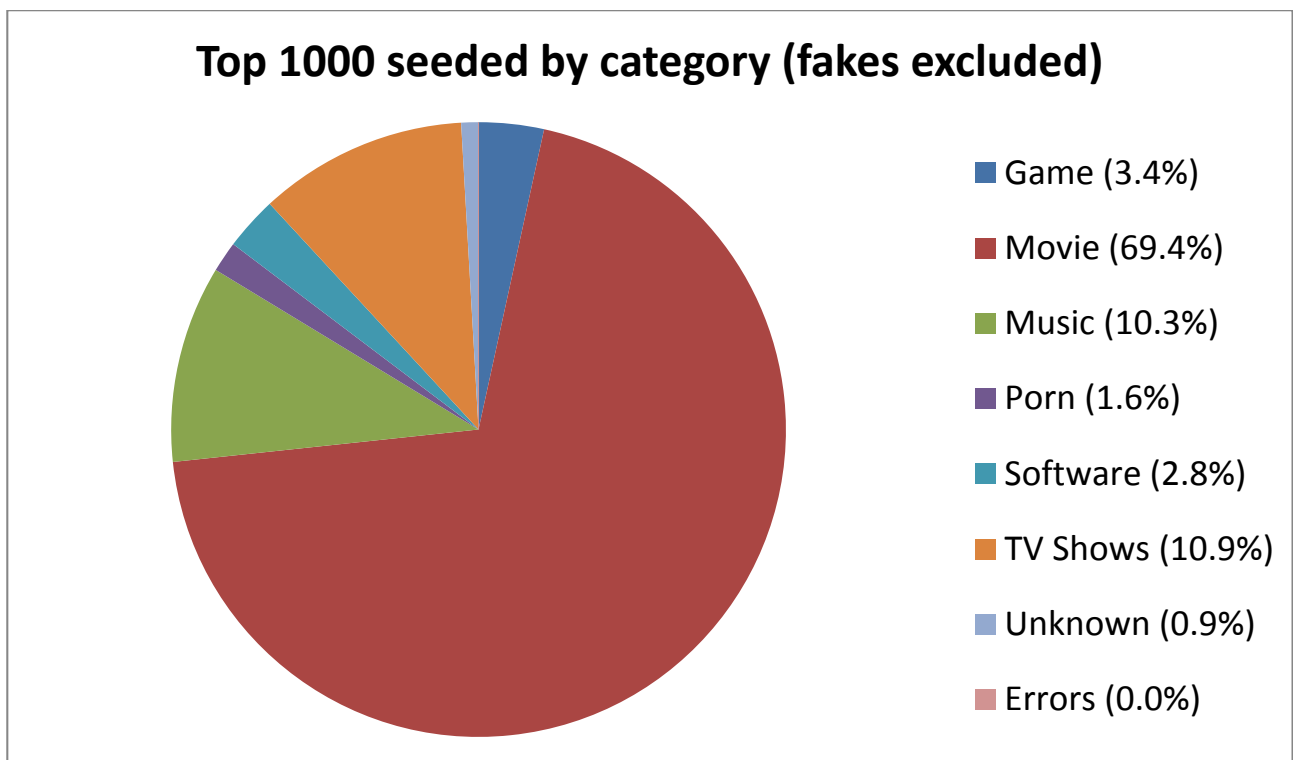


Table 4.5.3 – Category groupings of a random sample (fakes removed)

Category	Number	Adjusted %
Game	28	2.8%
Movie	233	23.3%
Books	31	3.1%
Music	109	10.9%
Hentai	2	0.2%
TV Shows	70	7.0%
Pornography	64	6.4%
Software	66	6.6%
Child Porn	1	0.1%
Pictures	2	0.2%
Unknown	394	39.4%
Errors	0	0.0%
Total	1000	100%

4.6.2 Category Separated Fake Detection

Table 4.6.1 contains the results of the fake file detection for each category. The (2000) figures are for the combined top 1000 downloaded and complete lists. The *Sample* statistics are for a completely random sample out of all torrents retrieved passing the median criteria discussed in section 3.3. Manual fake file detection was not performed for the sample. The lower fake rates are probably more a side effect of our methodology where votes on the files reliability are needed to assert whether a file is fake or not, with the default position being that files are considered real until otherwise proven.

Table 4.6.1 – Fake files, by category for the top 1000 downloaded and complete

Category	Downloads (2000)	Complete (2000)	% Fake (2000)	% Fake (Sample)	Count (Sample)
Movies	7,791,399	2,244,967	53.9%	3.0%	233
Music	404,784	35,882	11.0%	1.0%	109
TV	440,084	102,196	14.0%	1.0%	70
Porn*	4,118,739	436	0%*	0.0%	64
Game	25,590	1	0.0%	14.0%	28
Software	305,429	4,937	0.0%	3.0%	66
Book	0	0	N/A	6.0%	31
Child Porn	0	0	N/A	0.0%	1
Hentai	0	0	N/A	0.0%	2
Pictures	0	0	N/A	0.0%	2

* It is worth noting that very few of the results returned for the Porn categories had ratings, suggesting that further data sources are needed to determine the legitimacy of these ratings.

4.7 Copyright Infringement

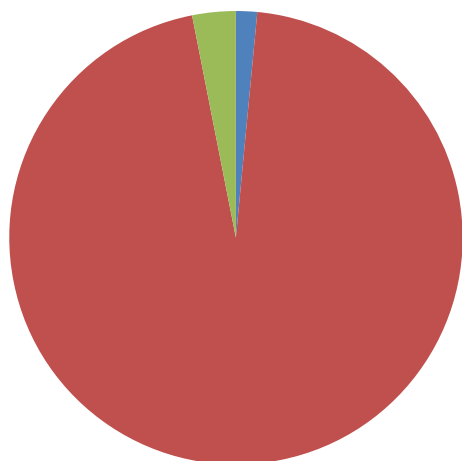
Both lists of the top 1000 downloaded and complete real files were manually checked to determine the percentage of copyright infringement. There were four categories assigned:

- **Copyright Infringing:** When a file infringes on the copyright holder of the files or material being shared.
- **Criminally Infringing:** When the files being shared included in the torrent in breaks a criminal law. The key example is child pornography.
- **Legal:** The files included in the torrent are legally being shared.
- **Unknown:** We were unable to determine the legality of the files being shared.

For the top 1,000 complete real files, we found 954 were infringing copyright, 15 were legal and 31 were of unknown legality. Most of the torrents of unknown legality did not have a filename and did not return any useful results though searching using Google. For the top 1000 downloaded real files, we found 990 were infringing copyright, 4 were legal and 6 were of unknown legality. Overall, this gives a figure of 97.2% copyright infringing material in our sample of the most popular torrents after removing fake files. Of the random 1,000 torrents, 666 were copyright infringing, 1 was criminally infringing (child pornography), 330 were of unknown legality and just 3 were legally distributed. This gives 66.7% of the sample as illegal, with a further 33% of unknown legality. Even including this sample with the top 1,000 lists, we still obtain an overall rate of 87% infringing, however many of the torrents in the random list had very low downloaded or complete figures, and many of them were of unknown legality.

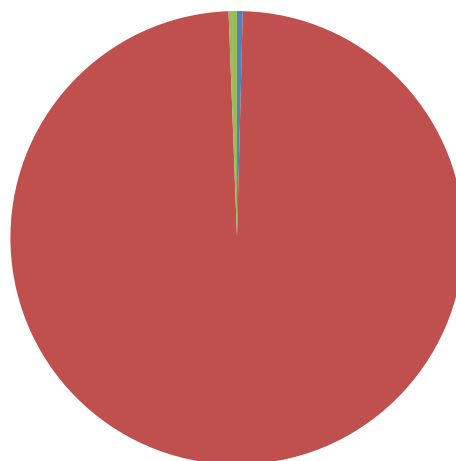
Figure 4.7.1 – Infringing and legal status of files

a) Percentage of Infringements in top 1000 completed files



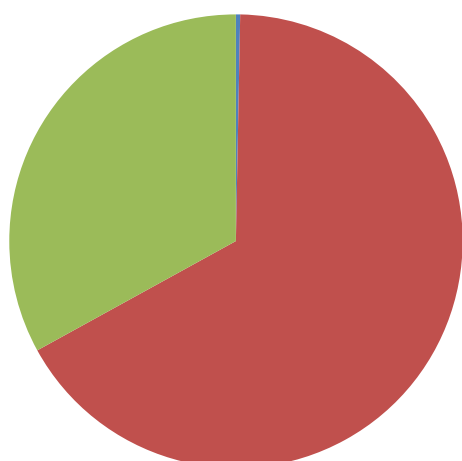
■ Legal (1.5%)
■ Infringing (95.4%)
■ Unknown (3.1%)

b) Percentage of Infringements in top 1000 real files



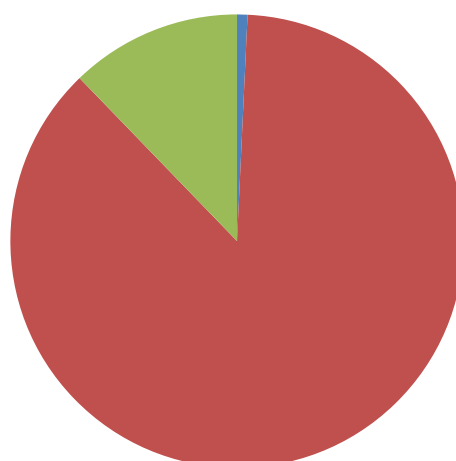
■ Legal (0.4%)
■ Infringing (99.0%)
■ Unknown (0.6%)

c) Percentage of Infringements in 1000 random files



■ Legal (0.3%)
■ Infringing (66.7%)
■ Unknown (33.0%)

d) Percentage of Infringements in all files



■ Legal (0.7%)
■ Infringing (87.0%)
■ Unknown (12.2%)

5 Criticisms of previous studies

One of the hallmarks of scientific research is inviting critical feedback through a public, peer-review process. This process is intended to ensure that scientific research is rigorously assessed for accuracy and quality, and ensures that incorrect or unfounded findings do not make their way into the scientific literature. We have published sufficient detail in our reports for other researchers to replicate our findings, which is common practice in scientific research. In broadly releasing the results of our initial studies, we have reviewed and revised our methodology to address many of the concerns, and our findings will be published in the Proceedings of the 45th Hawai'i International Conference on System Sciences (HICSS). HICSS was ranked by the Australian Research Council in 2010 as being in the top-tier (A) of information systems conferences, and we gratefully acknowledge the feedback of four anonymous, *bona fide* researchers. A further important aspect to ensuring integrity is the announcement of funding and research partners with interests in the research. We have ensured that we have disclosed this information and encourage others to do the same.

In this section, we examine some of the criticisms directed at our two previous studies, and identify those that are legitimate concerns, those that have been addressed, and those that are incorrect and/or misleading.

The main report released since the last report by the ICSL was Envisional's *An Estimate of Infringing Use of the Internet* [4]. The authors highlighted three criticisms, directed at the ICSL's first report [1]. Overall, the main concern can be pinpointed on the lack of handling of outliers in the first study.

1. **The report suggests that the choice of seeders, rather than downloads, is incorrect as there may be an inflated number of seeders.** While Envisional's report used the same information, they retrieved their data from a single source – PublicBT. This had both positive and negative points. If the sample data retrieved from PublicBT was representative of the population of download(er)s, then so would their results. However, if the sample was biased, then the results would not be representative for other trackers. The Envisional report used other methods to confirm their results. Envisional raised a valid point regarding not appropriately dealing with outliers – our initial study used the maximum value reported by trackers, which led to results higher than Envisional's, and higher than reported in other studies. Subsequently, we adopted a more conservative strategy of using the median, as a way of reducing the impact of outliers. A possible cause for this was given as the inclusion of faked files in the original results, which we have also explicitly addressed.

2. **Domain pseudonyms, where multiple entries in our tracker list point to the same tracker.** This does not cause issues with the results, as multiple entries are not summated (which would be incorrect). The figures reported in the initial ICSL report were single tracker figures.

3. **The magnitude of the download figures is much higher than other studies, and most downloaded files are inconsistent.** Again, this leads from the lack of removal of outliers in the original data. Since then, the ICSL's validation study [2] produced results much more in line with Envisional's, due to the exclusion of the tracker that reported the majority of high figures in the first study. However, the proportions (level of infringement, categorisations etc) remained very similar between the studies. Envisional's report does not mention our validation study. Note that the validation study was released in November 2010 and Envisional's report was published in January 2011. This timeframe suggests that Envisional's report was perhaps mostly written before the release of our validation report.

Further criticism has been given by online blog TorrentFreak. A large amount of criticism from TorrentFreak relies on incorrect readings of the reports, as described below.

1. **Claim that [1] states that there are only one million torrents.** This claim was not in the original report, which stated that there are 1 million torrents in our sample. While OpenBitTorrent was included in our tracker sampling, we noted that connections to full scrapes get dropped, but we include the partial scrapes. This caused us to obtain only part of OpenBitTorrent's 2.5 million torrents in our sample.

2. **Categorisation is based on most seeded, not overall figures.** The initial report [1] stated that the focus was on popular usage of BitTorrent, not overall usage. Our power law finding showed that a small proportion of torrents account for most of the usage, a claim validated by the results in this study. It would be trivial to generate an arbitrarily large number of torrents, even legal ones. However if nobody downloads them, they would not make an impact in the ICSL's findings.

3. **Outliers, particularly faked seeders and faked files, skewed the results.** As indicated above, the initial reports did not look for outliers, which did alter the results. Our validation study produced numbers more in line with other reports such as Envisional's and the enhanced sampling and fake detection in this report seeks to further enhance the accuracy of reporting.

4. **The percentage of infringing files is too high.** This argument was inaccurately backed up by citing IsoHunt as a single source for information, and that, as they point to 1.5% of legitimate files, that anything less than this figure would immediately be incorrect. The issue with this is that percentages are not additive, meaning that just because 1.5% of torrents on isoHunt are legal, it doesn't mean that this is the minimum global figure. Further, as pointed out above, our study focused on popular usage, not overall usage. Having a large number of torrents which are not actively used does not represent use of BitTorrent, therefore a study should look at usage more than occurrence. Those 1.5% of torrents come from Jamendo, but if few actually download them, they would not make a significant impact in the results of the ICSL's methodology.

5. **Categorisation is incorrect, because BitSnoop's categorisations are different.** Again, the focus of our study was on popular and not overall usage. Further to this, we noted that our categorisation had coverage of 69.9%, meaning that not all torrents were categorised. This is highlighted in both reports and the categorisation rules were given. Some categories are easier to develop rules for than other categories, leading to differences in categorisation rates.

6. **Finally, one article states that we added the figures for each torrent to produce our overall figure.** This is incorrect and is not present anywhere in any report produced by the ICSL. Such a calculation would be obviously flawed since, as stated in the ICSL's reports, torrents often have multiple trackers.

The first four criticisms are from [13] while the following 2 are from [14]. TorrentFreak also states that the "real" seed count is between 10 and 20 million at any point in time and that the highest for a single torrent was around 13,000 at the time of publication of their article on the first ICSL report. We would be keen to attempt to reproduce their results as per normal scientific practice; however we could not find details about their methodology, or sources of data, in determining this figure. TorrentFreak also publishes a weekly list of the most downloaded movies however overall figures are not given and, again, the methodology and data source is not published. Without these details, a comparison between those results and those obtained here is difficult.

Overall, the main valid criticism of the ICSL's previous studies can be directed at the lack of handling of outliers in the results, and that faked files were not accounted for. While both of these concerns have been addressed in this methodology, the overall conclusions – that a significant proportion of BitTorrent usage is copyright infringing and that movies account for a large proportion of that figure – remain the same, validated by this improved methodology.

6 Discussion

In this paper, we have provided estimates of the sharing of infringing content on BitTorrent networks, by further developing key parts of our methodology in tracker sampling and fake file detection. Unsurprisingly, our findings confirm that the majority of files shared on BitTorrent are infringing content, after fake files are removed from the results.

There are 1994 unique torrents from both top 1000 lists showing little overlap. This suggests the obvious hypothesis that torrents have a brief period of activity with many seeders active at the start, but that over time the number of seeders decreases rapidly while the number of downloads increases at a slowing rate. Specifically, there is a large difference between pornography appearing in the top downloaded torrents and not appearing on the top seeded torrents while the reverse is somewhat true for movies (movies more regularly appear in both lists). Over 50% of the most downloaded torrents were pornography, while only 1.6% of the most seeded torrents were. This suggests two things. Firstly that categories such as movies have a burst period, where files are shared a lot when they are first released but this activity reduces sharply afterwards. Secondly, categories such as pornography do not have this period of high activity followed by low rates of sharing. Instead, it suggests that for categories such as pornography, the files are shared slowly over a longer time. BitTorrent usage, specifically the flow of sharing, is therefore partially determined by the category of the torrent rather than a single "overall" model of sharing.

We found that, overall, that movies files were the most active, both with and without excluding fakes. This shows that movies are considered high value content for both the distribution of "real" copyright infringing content and faked files. Of the combined top 1000 lists including fake files, 64% (94% complete, 34% downloaded) were movies files. Of the combined top 1000 lists excluding fakes, 53.35% (69% complete, 37% downloaded) were movies. This is a large proportion of the overall usage of BitTorrent, and we found no evidence of legal usage within this category. A preliminary search of the larger database suggests that this approximate figure of 60% movie content is consistent with popular usage, although more investigation needs to be performed into larger usage of BitTorrent.

Future studies will need to enhance the rule-based approach for fake file detection. Although most torrents can be adequately classified using the isoHunt and Torrentz rules, not all torrents have been rated by the community, so any lower-ranked rules are quite brittle, although may be effective in limited domains. Categories such as pornography had a low proportion of rated files, suggesting that isoHunt and Torrentz communities do not rate these files as regularly.

Another area for future improvements would be to investigate content sharing rather than torrent sharing. The methodology presented focuses on torrents rather than content. As an example, a search for the phrase "**x-men first class**" returns over 1300 entries on isoHunt². Many of these are fakes, as noted by isoHunt's rating system. However at least 7 appear to be real versions of the recently released movie. These different versions are created by different release groups at different times in the release cycle – some versions are quickly released, while others take time to release a better version. The top rated torrent is reported (by isoHunt) to have nearly 2000 seeders at this time. Combining all of the torrents that appear to be real, there are over 5000 seeders sharing versions of this file. This fragmentation could lead to situations where the most shared content does not necessarily have the most shared torrents. Future methodologies may determine what content is the most popular rather than which torrents are.

In putting the figures above together, the results indicate that 50% of the most downloaded torrents are faked, 49.9% are copyright infringing and 0.1% are legal. For the top 1000 complete list, the results indicate that 55% are faked, 41.7% were copyright infringing, 0.2% were legal and 3.1% were of unknown legality. Overall, we find that 97.9% of popular BitTorrent use is nefarious in nature, either in the distribution of copyright infringing files or faked files. When observing the copyright infringement levels after removing faked files, we find at 97.2% are copyright infringing. A random sample found 66.7% were copyright or criminally infringing with a further 33% of unknown legality.

2 Query and figures determined in June, 2011.

Overall, regardless of the sampling method, we found little legal active use of BitTorrent on popular trackers.

6.1 Acknowledgements

The ICSL is funded by the State Government of Victoria, IBM, Westpac Banking Corporation and the Australian Federal Police. This project received financial support from Village Roadshow.

7 References

- [1] Layton, R. & Watters, P.A. (2010). Investigation into the extent of infringing content on BitTorrent networks. ICSL Technical Report ICSL-0001, April 2010, downloaded from http://www.icsl.com.au/files/bt_report_final.pdf
- [2] Layton, R., Watters, P.A., & Dazeley, R. (2010). How much material on BitTorrent networks is infringing content? A validation study. ICSL Technical Report ICSL-0001, November 2010, downloaded from http://www.icsl.com.au/files/validation_study_nov_2010.pdf
- [3] Treverton, G., Matthies, C., Cunningham, K., Goulka, J., Ridgeway, G., & Wong, A. (2009). Film Piracy, Organized Crime, and Terrorism. RAND Corporation.
- [4] Price, D. (2011). An estimate of infringing use of the Internet. Envisional Technical Report, January 2011, downloaded from http://documents.envisional.com/docs/Envisional-Internet_Usage-Jan2011.pdf
- [5] Dhungel, P. & Wu, D., Schonhorst, B. & Ross, K. (2008). A Measurement Study of Attacks on BitTorrent Leechers. *Proceedings of the International Workshop on Peer-to-Peer Systems (IPTPS)*.
- [6] Burns, A. & Jung, E. (2008). Searching for malware in BitTorrent. *University of Iowa Computer Science Technical Report UICS-08-05*.
- [7] Royal Canadian Mounted Police (2006). Future Trends in Malicious Code - 2006 Report. *Information Technology Security Report Lead Agency Publication R2-002*.
- [8] Prichard, J., Watters, P.A. & Spiranovic, C. (Submitted). On-line subcultures and onset of use of child pornography.
- [9] Tran, H., Hitchens, M., Varadharajan, V. and Watters, P.A. (2005). A trust based access control framework for P2P file-sharing systems. *Proceedings of the Hawaii International Conference on System Sciences (HICSS-38)*, Honolulu HI, USA.
- [10] Clark, D. (2004). Using consensus ensembles to identify suspect data. *Lecture Notes in Computer Science, Volume 3214/2004, 483-490*.
- [11] Bittorrent Protocol Specification v1.0, downloaded from <http://wiki.theory.org/BitTorrentSpecification>
- [12] Wikipedia (2011) Robust Statistics. Revision 426608964, downloaded from http://en.wikipedia.org/w/index.php?title=Robust_statistics&oldid=426608964
- [13] Ernesto (2010) Tech News Sites Tout Misleading BitTorrent Piracy Study, downloaded from <http://torrentfreak.com/tech-news-sites-tout-misleading-bittorrent-piracy-study-100724/>
- [14] Ernesto (2010) Incompetent BitTorrent Researchers Strike Again, downloaded from <http://torrentfreak.com/incompetent-bittorrent-researchers-strike-again-101211/>