

Is Content Publishing in BitTorrent Altruistic or Profit-Driven?

Ruben Cuevas
Universidad Carlos III de
Madrid
rcuevas@it.uc3m.es

Michal Kryczka
Institute IMDEA Networks and
Universidad Carlos III de
Madrid
michal.kryczka@imdea.org

Angel Cuevas
Universidad Carlos III de
Madrid
acrumin@it.uc3m.es

Sebastian Kaune
KOM Lab
TU Darmstadt
kaune@kom.tu-darmstadt.de

Carmen Guerrero
Universidad Carlos III de
Madrid
guerrero@it.uc3m.es

Reza Rejaie*
University of Oregon
reza@cs.uoregon.edu

ABSTRACT

BitTorrent is the most popular P2P content delivery application where individual users share various type of content with tens of thousands of other users. The growing popularity of BitTorrent is primarily due to the availability of valuable content without any cost for the consumers. However, apart from required resources, publishing (sharing) valuable (and often copyrighted) content has serious legal implications for users who publish the material (or publishers). This raises a question that whether (at least major) content publishers behave in an altruistic fashion or have other incentives such as financial. In this study, we identify the content publishers of more than 55K torrents in two major BitTorrent portals and examine their behavior. We demonstrate that a small fraction of publishers is responsible for 67% of the published content and 75% of the downloads. Our investigations reveal that these major publishers respond to two different profiles. On the one hand, antipiracy agencies and malicious publishers publish a large amount of fake files to protect copyrighted content and spread malware respectively. On the other hand, content publishing in BitTorrent is largely driven by companies with financial incentives. Therefore, if these companies lose their interest or are unable to publish content, BitTorrent traffic/portals may disappear or at least their associated traffic will be significantly reduced.

*Reza Rejaie was Visiting Researcher at Institute IMDEA Networks, September 2009 - August 2010, while this study was performed.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

ACM CoNEXT 2010, November 30 – December 3 2010, Philadelphia, USA.

Copyright 2010 ACM 1-4503-0448-1/10/11 ...\$10.00.

Categories and Subject Descriptors

C.2.4 [Computer Communication Networks]: Distributed Systems – *Distributed Applications*

General Terms

Measurement, Economics

Keywords

BitTorrent, Content Publishing, Business Model

1. INTRODUCTION

Peer to Peer (P2P) file-sharing applications, and more specifically BitTorrent, are a clear example of *killer* Internet applications in the last decade. BitTorrent is currently used by hundreds of millions of users and is responsible for a large portion of the Internet traffic [4]. This in turn has attracted the research community to examine various aspects of swarming mechanism in BitTorrent [10, 16, 13] and propose different techniques to improve its performance [19, 15]. Furthermore, other aspects of BitTorrent such as demography of users [22, 12, 20] along with the security [21] and privacy [6, 7] issues have also been studied. However, the socio-economic aspects of BitTorrent in particular, and other P2P file sharing systems in general, have received little attention. In particular, the availability of popular and often copyrighted content (e.g. recent TV shows and Hollywood movies) to millions of interested users at no cost is the key factor in the popularity of BitTorrent. This raises an important question about the incentive of publishers who make these content available through BitTorrent portals. Despite its importance to properly understand the popularity of BitTorrent, to our knowledge prior studies on BitTorrent have not tackled this critical question.

In this paper, we study content publishing in Bit-

Torrent from a socio-economic point of view by unravelling *who* publishes content in BitTorrent, and *why*. Toward this end, we conduct a large scale measurement over two major BitTorrent portals, namely Mininova and the Pirate Bay, to capture more than 55K published content that involve more than 35M IP addresses. Using this dataset, we first examine the contribution of the individual content publishers and illustrate that a small fraction of publishers (~ 100) are responsible of uploading 67% of the content that serve 75% of the downloads in our major dataset. Furthermore, most of these major publishers dedicate their resources for publishing content and consume few or no published content by others, *i.e.* their level of content publication and consumption is very imbalanced. In addition to allocating a significant amount of resources for publishing content, these users often publish copyrighted material, which has legal implications for them [1, 2]. These observations raise the following question: *what are the main incentives of (major) content publishers in BitTorrent?*

To answer this important question, we conduct a systematic study on the major publishers in BitTorrent. We show that these publishers can be broadly divided into two different groups: *fake publishers* who publish a large number of fake content and *top publishers* who publish a large number of often copyrighted content. We also identify the main characteristics (*i.e.* signature) of publishers in each group such as their seeding behavior and the popularity of their published content. We investigate the main incentives of major (non-fake) publishers and classify them into three categories (*i*) Private BitTorrent Portals that offer certain services and receive financial gain through ads, donations and fees, (*ii*) Promoting web sites that leverage published content at BitTorrent portals to attract users to their own web site for financial gain, and (*iii*) Altruistic Major Publishers. We characterize these three groups of publishers and present the estimated value (income) of the associated web sites to support our claims about their incentives.

The main contributions of this paper can be summarized as follows:

- We present a simple measurement methodology to monitor the content publishing activity in major BitTorrent portals. This methodology has been used to implement a system that continuously monitors and reports the content publishing activity in the Pirate Bay portal. The collected data is made publicly available through a web site.
- The distribution of the number of published content by each publisher is very skewed, *i.e.* a very small fraction of publishers (~ 100) is responsible for a significant fraction of the published content

(67%) and even more significant fraction of the downloads (75%). These major publishers can be further divided into three groups based on their incentives as follows: *fake publishers*, *altruistic top publishers* and *profit-driven publishers*.

- *Fake publishers* are either antipiracy agencies or malicious users who are responsible for 30% of the content and 25% of the downloads. These publishers sustain a continuous poisoning-like index attack [17] against BitTorrent portals that affects millions of downloaders.
- *Profit-driven top publishers* own fairly profitable web sites. They use major BitTorrent portals such as the Pirate Bay as a platform to advertise their web sites to millions of users. For this purpose they publish popular torrents where they attach the URL of their web sites in various manners. The publishers that pursue this approach are responsible for roughly 30% of the content and 40% of the downloads in BitTorrent.

The rest of the paper is organized as follows. Section 2 describes our measurement methodology. Sections 3 and 4 are dedicated to the identification of major publishers and their main characteristics (*i.e.* signature) respectively. In Section 5, we study the incentives that major publishers have to perform this activity. Section 6 presents other players that also benefit from content publishing. In Section 7 we describe our publicly available application to monitor content publishing activity in the Pirate Bay portal. Finally Section 8 discusses related work and Section 9 concludes the paper.

2. MEASUREMENT METHODOLOGY

This section describes our methodology to identify the initial publisher of a file that is distributed through a BitTorrent swarm. Towards this end, we first briefly describe the required background on how a user joins a BitTorrent swarm.

Background: A BitTorrent client takes the following steps to join the swarm associated with file X . First, the client obtains the .torrent file associated to the desired swarm. The .torrent file contains contact information for the tracker that manages the swarm and the number of pieces of file X . Second, the client connects to the tracker and obtains the following information: (*i*) the number of seeders and leechers that are currently connected to the swarm, and (*ii*) N (typically 50) random IP addresses of participating peers in the swarm. Furthermore, if the number of neighbors is eventually lower than a given threshold (typically 20), the client contacts the tracker again to learn about other peers in the swarm.

To facilitate the bootstrapping process, the .torrent files are typically indexed at BitTorrent portals. Some

| | Portal | Start | End | #Torrents | #IP addresses |
|-------------|------------|-----------|-----------|-------------|---------------|
| <i>mn08</i> | Mininova | 09-Dec-08 | 16-Jan-09 | - /20.8K | 8.2M |
| <i>pb09</i> | Pirate Bay | 28-Nov-09 | 18-Dec-09 | 23.2K/10.4K | 52.9K |
| <i>pb10</i> | Pirate Bay | 06-Apr-10 | 05-May-10 | 38.4K/14.6K | 27.3M |

Table 1: Datasets Description.

of the major portals (*e.g.* the Pirate Bay or Mininova¹) index millions of .torrent files [22], classify them into different categories and provide a web page with detailed information (content category, publisher’s username, file size, and file description) for each file. These portals also offer an RSS feed to announce a newly published file. The RSS feed provides some information such as content category, content size and publisher’s username for a new file.

Identifying Initial Publisher: The objective of our measurement study is to determine the identity of the initial publishers of a large number of torrents and to assess the popularity of each published file (*i.e.* the number and identity of peers who download the file).

Toward this end, we leverage the RSS feed to detect the availability of a new file on major BitTorrent portals and retrieve the publisher’s username. In order to obtain the publisher’s IP address, we immediately download the .torrent file and connect to the associated tracker. This implies that we often contact the tracker shortly after the birth of the associated swarm when the number of participating peers is likely to be small and the initial publisher (*i.e.* seeder) is one of them. We retrieve the IP address of all participating peers as well as the current number of seeders in the swarm. If there is only one seeder in the swarm and the number of participating peers is not too large (*i.e.* < 20), we obtain the bitfield of available pieces at individual peers to identify the seeder. Otherwise, reliably identifying the initial seeder is difficult because there are more than one seeder or the number of participating peers is large². Furthermore, we cannot directly contact the initial seeder that is behind a NAT box and thus we are unable to identify the initial publisher’s IP address in such cases. Using this technique we were able to reliably identify the publisher’s username for all the torrents and the publisher’s IP address in at least 40% of the torrents.

Once we identify a publisher, we periodically query the tracker in order to obtain the IP addresses of the participants in the associated swarm and always solicit the maximum number of IP addresses (*i.e.* 200) from

¹<http://thepiratebay.org/>, <http://www.mininova.org/>.

²Our investigations revealed two interesting scenarios for which we could not identify the initial publisher’s IP address: (*i*) swarms that have a large number of peers shortly after they are added to the portal. We discovered that these swarms have already been published in other portals. (*ii*) swarms for which the tracker did not report any seeder for a while or did not report a seeder at all.

the tracker. To avoid being blacklisted by the tracker, we issue our queries at the maximum rate that is allowed by the tracker (*i.e.* 1 query every 10 to 15 minutes depending on the tracker load). Given this constraint, we query the tracker from several geographically-distributed machines so that the aggregated information by all these machines provides an adequately high resolution view of participating peers and their evolution over time. We continue to monitor a target swarm until we receive 10 consecutive empty replies from the tracker. We use the MaxMind Database [3] to map all the IP addresses (for both publishers and downloaders) to their corresponding ISPs and geographical locations.

2.1 Dataset

Using the described methodology, we identify a large number of BitTorrent swarms at two major BitTorrent portals, namely Mininova and the Pirate Bay. Each one of these portals was the most popular BitTorrent portal at the time of the corresponding measurement according to Alexa ranking³. It is worth noting that the Pirate Bay is in particular interesting for our study since it is the only main BitTorrent portal where all the published content is contributed by users [22] (as opposed to being retrieved from other portals). Table 1 shows the main features of our three datasets (1 from Mininova and 2 from the Pirate Bay) including the start and end dates of our measurement, the number of torrents for which we identified the initial publisher (username/IP address), and the total number of discovered IP addresses associated for all the monitored swarms. We refer to these datasets as *mn08*, *pb09* and *pb10* throughout this paper. We note that dataset *mn08* does not contain the username of initial publishers, and we use a single query to identify initial publishers in dataset *pb09* after detecting a new swarm through the RSS feed. We use all three datasets for our general analysis but focused on *pb10* for our detailed analysis.

3. IDENTIFYING MAJOR PUBLISHERS

A publisher can be identified by its username and/or IP address. In our analysis, we identify individual publishers primarily by their username since the username is expected to remain consistent across different torrents. The only exceptions are publishers in *mn08* since we do not have their usernames and *fake* publishers as we describe in the next subsections.

3.1 Skewness of Contribution

First, we examine the level of contribution (*i.e.* the number of published files) by the identified content publishers in each dataset. Figure 1 depicts the percentage of files that are published by the top x% of the publishers in our three datasets. We observe that the top 3% of

³<http://www.alexa.com/topsites>.

| mn08 | | | pb09 | | | pb10 | | |
|---------------------------|------------------|-------|--------------|------------------|-------|-----------------------|------------------|-------|
| ISP | Type | % | ISP | Type | % | ISP | Type | % |
| OVH | Hosting Provider | 13.31 | OVH | Hosting Provider | 24.76 | OVH | Hosting Provider | 15.16 |
| Comcast | Commercial ISP | 4.69 | Comcast | Commercial ISP | 3.67 | SoftLayer Tech. | Hosting Provider | 4.52 |
| Keyweb | Hosting Provider | 3.18 | Road Runner | Commercial ISP | 2.3 | FDCservers | Hosting Provider | 3.64 |
| Road Runner | Commercial ISP | 3.03 | Romania DS | Commercial ISP | 2.27 | Open Computer Network | Commercial ISP | 3.59 |
| NetDirect | Hosting Provider | 2.44 | MTT Network | Commercial ISP | 1.95 | tzulo | Hosting Provider | 3.36 |
| Virgin Media | Commercial ISP | 2.42 | Verizon | Commercial ISP | 1.64 | Comcast | Commercial ISP | 2.86 |
| NetWork Operations Center | Hosting Provider | 2.39 | Virgin Media | Commercial ISP | 1.49 | Cosema | Commercial ISP | 2.25 |
| SBC | Commercial ISP | 2.38 | SBC | Commercial ISP | 1.41 | Telefonica | Commercial ISP | 2.22 |
| Comcor-TV | Commercial ISP | 2.33 | NIB | Commercial ISP | 1.26 | Jazz Telecom. | Commercial ISP | 2.07 |
| Telecom Italia | Commercial ISP | 2.02 | tzulo | Hosting Provider | 1.14 | 4RWEB | Hosting Provider | 2.06 |

Table 2: Content Publishers Distribution per ISP.

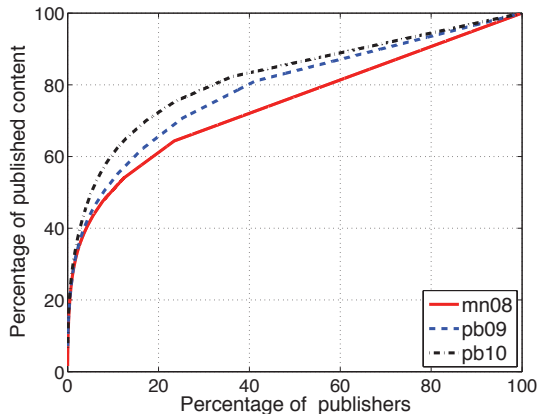


Figure 1: Percentage of content published by the top x% publishers.

the BitTorrent publishers contribute roughly 40% of the published content. Moreover, a careful examination of IP addresses for the top-100 (*i.e.* 3%) publishers in our *pb10* dataset reveals that a significant fraction of them either do not download any content (40%) or download less than 5 files (80%). This large contribution of resources (bandwidth and content) by major publishers coupled with the significant imbalance between their publishing and consumption rates seems non-altruistic and rather difficult to justify for two simple reasons:

- **Required Resources/Cost:** publishing a large number of content requires a significant amount of processing and bandwidth. For example, a major content publisher named *extv* recommends in its private BitTorrent portal web page (www.extv.it) to allocate at least 10Mbps in order to sustain the seeding of few (around 5) files in parallel.

- **Legal Implications:** As other studies have reported [6] and we confirm in our datasets, a large fraction of content published by major publishers is copyrighted material (recent movies or TV series). Thus, publishing these files is likely to have serious legal consequences for these publishers [1, 2].

This raises the question that *why this small fraction of publishers allocate a great deal of (costly) resources to contribute many files into BitTorrent swarms despite potential legal implications?* We answer this central question in Section 5.

3.2 Publishers' ISPs

To help identify content publishers in our dataset, we determine the ISP that hosts each major publisher and use that information to assess the type of service (and available resources) that a publisher is likely to have. Toward this end, we map the IP address for a publisher in each dataset to its corresponding ISP using the MaxMind database [3]. We then examine publicly available information about each ISP (*e.g.* its web page) to determine whether it is a commercial ISP or a hosting provider. We perform this analysis only for the top-100 (roughly 3%) publishers since these publishers are mostly of interest and collecting the required information for all publishers is a tedious task. Since we do not have publishers' username for *mn08*, we examine the top-100 publishers based on their IP addresses in this dataset. For these publishers, we cannot assess the aggregated contribution of a publisher through different IP addresses (*i.e.* under-estimating the contribution of each publisher).

We observe that 42% of the top-100 publishers in *pb10*, 35% of the top-100 in *pb09* and 77% of the top-100 publishers in *mn08* are located at hosting services. Moreover 22%, 20% and 45% of these top-100 publishers are located at a particular hosting services, namely OVH, in *pb10*, *pb09* and *mn08* respectively.

In short, our analysis reveals that a significant fraction of major publishers are located at a few hosting services and a large percentage of them at OVH.

We also examine the contribution of BitTorrent publishers at the ISP-level by mapping all the publishers to their ISPs and identify the top-10 ISPs based on their aggregate published content for each dataset as shown in Table 2. This table confirms that content publishers who are located at a particular hosting provider, namely OVH, have consistently contributed a significant fraction of published content at major BitTorrent portals. There are also several commercial ISPs (*e.g.* Comcast) in Table 2 with a much smaller contribution.

To assess the difference between users from hosting providers and commercial ISPs, we compare and contrast the characteristics of all publishers that are located at OVH and Comcast as representative ISPs for each class of publishers in Table 3. This table demonstrates the following two important differences: first,

| | Published torrents | # IP addr | # /16 IP Pref. | # Geo Loc. |
|----------------|-----------------------|-----------|-------------------|---------------|
| OVH (mn08) | 2766 | 164 | 5 | 2 |
| Comcast (mn08) | 976 | 675 | 269 | 400 |
| OVH (pb09) | 2577 | 78 | 5 | 2 |
| Comcast (pb09) | 382 | 198 | 143 | 129 |
| OVH (pb10) | 2213 | 92 | 7 | 4 |
| Comcast (pb10) | 408 | 185 | 139 | 147 |

Table 3: Characteristics of all OVH and Comcast publishers in *mn08*, *pb09* and *pb10*.

the aggregate contribution of each publisher at OVH is on average a few times larger than Comcast publishers. Second, Comcast publishers are sparsely scattered across many /16 IP prefixes and many geographical locations in the US whereas OVH publishers are concentrated in a few /16 IP prefixes and a handful of different locations in Europe (*i.e.* the location of OVH’s data centers). In essence, the published content by Comcast publishers comes from a large number of typical altruistic users where each one publishes a small number of files from their home or work. In contrast, OVH publishers appear to be paying for a well provisioned service to be able to publish a much larger number of files. We have also examined consumer peers in captured torrents and did not observe the presence of OVH users among the consuming peers.

In summary, the examination of ISPs that host major BitTorrent publishers suggests that these publishers are located either at a few hosting providers (with a large concentration at OVH) or at commercial ISPs. These publishers contribute a significantly larger number of files than average publishers. Furthermore, publishers who are located at hosting providers do not consume published content by other publishers.

3.3 A Closer Look at Major Publishers

We now examine the mapping between username and IP address of the top-100 content publishers in the *pb10* dataset to gain some insight about major publishers behavior. Our examination reveals the following interesting points:

First, if we focus on the top-100 IP addresses that have published the largest number of files, only 55% of them are used by a unique username. The remaining 45% of the IP addresses of major publishers are mapped to a large number of usernames. We have carefully investigated this set of IP addresses and discovered that they use either hacked or manually created accounts (with a random username) to inject “fake” content. These publishers appear to be associated with anti-piracy agencies or malicious users. The former group tries to avoid the distribution of copyrighted content whereas the latter attempts to disseminate malware. We refer to these publishers as *fake publishers*. Surprisingly, fake publishers are responsible for around 25% of the usernames, 30% of the published content and

25% of the downloads in our *pb10* dataset. This suggests that major BitTorrent portals are suffering from a systematic poisoning index attack [17] that affects 30% of the published content. The portals fight this phenomenon by removing the fake content as well as the user accounts used to publish them. However, contrary to what has been reported in previous studies [20], this technique does not seem to be sufficiently effective since millions of users initiate the download of fake content. Finally, it is worth noting that most of the fake publishers perform their activity from three specific hosting providers named *tzulo*, *FDC Servers* and *4RWEB*. Due to the relevant activity of these fake publishers, we study them as a separate group in the rest of the paper.

Second, the inspection of the top-100 usernames who publish the largest number of files shows that only 25% of them operate from a single IP. The remaining 75% of top usernames utilize multiple IPs and can be classified into the following common cases: (i) 34% of the usernames with multiple IP addresses (5.7 IP addresses on average) at a hosting provider in order to obtain the required resources for seeding a large number of files. (ii) 24% of the usernames with multiple IP addresses (13.8 IP addresses on average) located at a single commercial ISP. Their mapping to multiple IP addresses must be due to the periodical change of their assigned IP addresses by their ISPs. (iii) The other 17% of these usernames are mapped to multiple IP addresses (7.7 IP addresses on average) at different commercial ISPs. These are users who inject content from various locations (*e.g.* home and work computer). To properly characterize different types of publishers, we exclude the 16 usernames who publish fake content from the top-100 usernames. We refer to the remaining top-100 usernames (non-fake publishers) as *Top publishers* who are responsible of 37% of the published content and 50% of the total downloads in our *pb10* dataset.

In summary, the major portion of the content comes from two reduced group of publishers: Top publishers and Fake publishers that collectively are responsible of 67% of the published content and 75% of the downloads. In the rest of the paper we devote our effort to characterize these two groups.

4. SIGNATURE OF MAJOR PUBLISHERS

Before we investigate the incentives of major BitTorrent publishers, we examine whether they exhibit any other distinguishing features, *i.e.* whether major publishers have a distinguishing signature. Any such distinguishing features could shed some light on the underlying incentives of these publishers. Toward this end, in the next few subsections, we examine the following characteristics of major publishers in our datasets: (i) the type of published content, (ii) the popularity of published content, and (iii) the availability and seeding

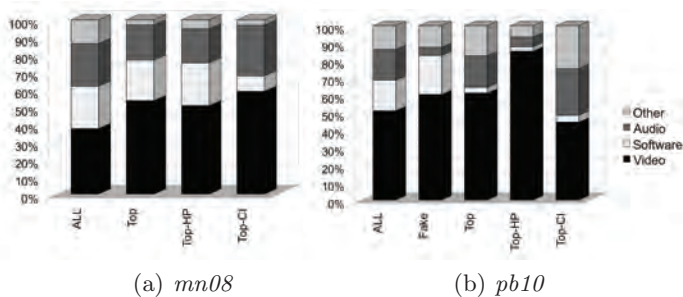


Figure 2: Type of published content distribution for the different classes of publishers: *All*, *Fake*, *Top*, *Top-HP* and *Top-CI*.

behavior of a publisher.

To identify distinguishing features, we examine the above characteristics across the following three *target groups* in each dataset: all publishers (labeled as “All”), all fake publishers (labeled as “Fake”) and all top-100 (non-fake) publishers regardless of their ISPs (labeled as “Top”). We also examine the break down of Top publishers based on their ISPs into hosting providers and commercial ISPs, labeled as “Top-HP” and “Top-CI”, respectively.

4.1 Content Type

We leverage the reported content type by each publisher to classify the published content across different target groups. Figure 2 depicts the break down of published content across different content type for all publishers in each target group for our Mininova and our major Pirate Bay datasets. We recall that without username information for each publisher in *mn08* dataset, we cannot identify fake publishers. Figure 2 reveals a few interesting trends as follows:

First, Video files (which mainly include movies, TV-shows and porn content) constitute a significant fraction of published files across most groups with some important differences. The percentage of published video across all publishers is around 37%-51% but it is slightly larger among top publishers. However, video is clearly a larger fraction of published content by the top publishers located at hosting providers in our *pb10* dataset. Fake publishers primarily focus on Videos (recent movies and TV shows) and Software content. This supports our earlier observation that these publishers consist of anti-piracy agencies and malicious users because the former group publishes a fake version of recent movies while the latter provides software that contains malware.

4.2 Content Popularity

The number of published files by a publisher shows only one dimension of its contribution to BitTorrent. The other equally important issue is the popularity of each published content (*i.e.* the number of download-

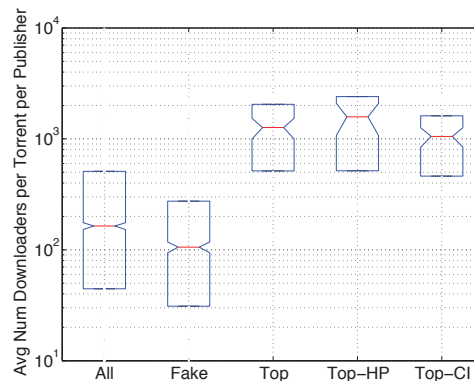


Figure 3: Avg Num of Downloaders per torrent per publisher for the different classes of publishers: *All*, *Fake*, *Top*, *Top-HP*, *Top-CI*.

ers regardless of their download progress) by individual publishers. Figure 3 shows the box plot of the distribution of average downloaders per torrent per publisher across all publishers in each target group where each box presents the 25th, 50th and 75th percentiles.

On the one hand, the median popularity of top publishers’ torrents is 7 times higher than a typical user (represented by *All*). A closer examination of the Top publishers shows that the content published by users at hosting providers is on average 1.5 times more popular than those published by users at commercial ISPs. On the other hand, fake publishers’ content is the most unpopular among the target groups. This is because the portals actively monitor the torrents and immediately remove the content identified as fake to avoid users from downloading it. Furthermore, users quickly realize the fake nature of these content and report this info on forums that inform others and limit their popularity.

In summary, top publishers are responsible for a larger fraction of popular torrents. This in turn magnifies the contribution of the 37% of the injected files by the top publishers to be responsible for 50% of all the downloads. The low popularity of fake publishers’ content has the opposite effect and limits their contribution to the number of downloads to 25%.

4.3 Seeding Behavior

We characterize the seeding behavior of individual publishers in our target groups using the following metrics: (*i*) average seeding time of a publisher for its published content, (*ii*) average number of parallel seeded torrents, and (*iii*) aggregated session time of a publisher across all its torrents. Since calculating these properties requires detailed analysis of our dataset that are computationally expensive, we are unable to derive these values for all publishers. We use 400 randomly selected publishers to represent the normal behaviour of all publishers and refer to this group as “All” in our analysis.

In order to compute these metrics we need to estimate

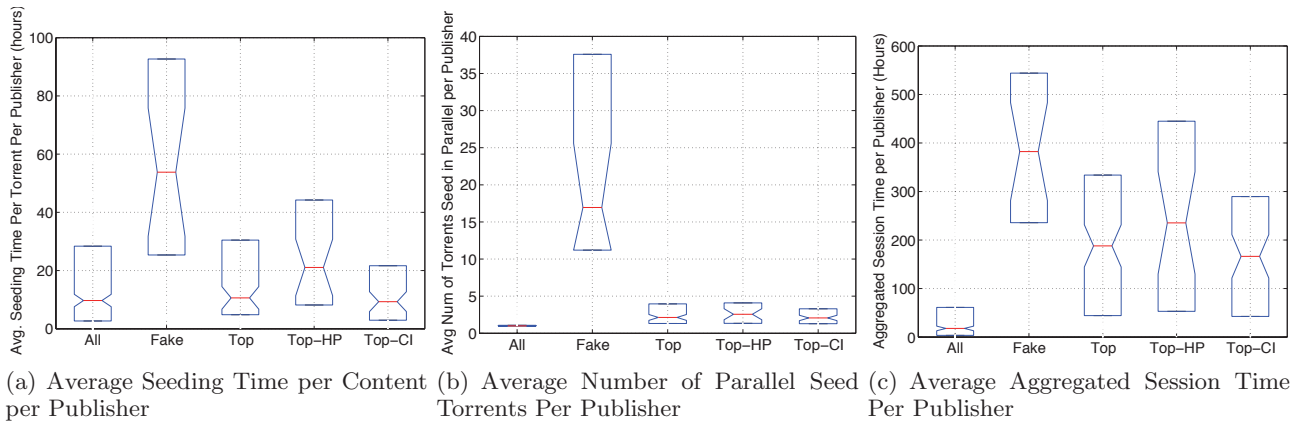


Figure 4: Seeding Behaviour for the different classes of publishers: *All*, *Fake*, *Top*, *Top-HP* and *Top-CI*.

the time that a specific publisher has been connected to a torrent (in one or multiple sessions). Since each query to the tracker just reports (at most) a random subset of 200 IPs, in big torrents (>200 peers), we need to perform multiple queries in order to assess the presence of the publisher in the torrent. In our associated Technical Report [8], we detail the technique used to estimate the session time of a specific user in each torrent.

Average Seeding Time: We measure the duration of time that a publisher stays in a torrent since its birth to seed the content. In general, a publisher can leave the torrent once there is an adequate fraction of other seeds. Figure 4(a) depicts the summary distribution of average seeding time across all publishers in each target group. This figure demonstrates the following points: first, the seeding time for fake publishers is significantly longer than publishers in other groups. Since these publishers do not provide the actual content, the initial fake publisher remains the only seed in the session (*i.e.* other users do not help in seeding fake content) to keep the torrent alive. Second, Figure 4(a) shows that top publishers typically seed a content for a few hours. However, the seeding time for top publishers from hosting providers is clearly longer than top publishers from commercial ISPs. This suggests that publishers at hosting providers are more concerned about the availability of their published content.

Average number of Parallel Torrents: Figure 4(b) depicts the summary distribution of the average number of torrents that a publisher seeds in parallel across publishers in each target group. This figure indicates that fake publishers seed many torrents in parallel. We have seen that fake publishers typically publish a large number of torrents and other users do not help them for seeding. Therefore, fake publishers need to seed all of their seeded torrents in parallel in order to keep them alive. The results for top publishers show that their typical number of seeded torrents in parallel is the same

(around 3 torrents) regardless of their location. However, we expect that a regular publisher seed only 1 file at a time.

Aggregated Session Time: We have also quantified the availability of individual publishers by estimating the aggregated session time that each publisher is present in the system across all published torrents. Figure 4(c) shows the distribution of this availability measure across publishers in each target group. As expected fake publishers present the longest aggregated session time due to their obligation to continuously seed their content to keep them alive. If we focus on top publishers, they exhibit a typical aggregated session 10 times longer than standard users. Furthermore, top publishers at hosting services are clearly more available than those from commercial ISPs.

4.4 Summary

BitTorrent content publishers can be broadly divided into three groups as follows: *(i)* Altruistic users who publish content while consuming content that is published by other users. *(ii)* Fake publishers publish a significant number of files that are often Video and Software content from a few hosting providers. Due to the fake nature of their content, their torrents are unpopular and they need to seed all torrents to keep them alive. These publishers appear to be associated with antipiracy agencies or malicious users. We validate this hypothesis in the next section. *(iii)* Top publishers publish a large number of popular files (often copyrighted video) and remain for a long time in the associated torrents to ensure proper seeding of their published content. These publishers are located at hosting facilities or commercial ISPs. Their behavior suggests that these publishers are interested in the visibility of the published content possibly to attract a large number of users. The (cost of) allocated resources by these publishers along with legal implications of publishing copyrighted material cannot be considered as altruistic.

Therefore, the most conceivable incentive for these publishers appears to be financial profit. We examine this hypothesis in the rest of the paper.

5. INCENTIVES OF MAJOR PUBLISHERS

In this section, we examine the incentive of different groups of publishers in more detail. First we focus on fake publishers by examining the name and content of the files they published. We noticed that these publishers often publish files with catchy titles (*e.g.* recently released Hollywood movies) and in most cases the actual content has already been removed at the moment of downloading it⁴. The few files we were able to download were indeed fake content. Some of them included an antipiracy message whereas some others led to malware software. In the latter case, the content was a video that had a pointer to a specific software (*e.g.* <http://flvdirect.com/>) to be downloaded in order to reproduce the video. This software was indeed a malware⁵. These observations validate our hypothesis that fake publishers are either antipiracy agents who publish fake versions of copyrighted content or malicious users who led users to download a malware. Therefore, we have clearly identified the incentives of fake publishers.

Second, another group of major publishers allocate significant amount of resources to publish non-fake an often copyrighted content. We believe that the behavior of these users is not altruistic. More specifically, our hypothesis is that these publishers leverage major BitTorrent portals as a venue to freely attract downloading users to their web sites. To verify this hypothesis, we conduct an investigation to gather the following information about each one of the *top* (*i.e.* top-100 non-fake) publishers:

- *Promoting URL*: the URL that downloaders of a published content may encounter,
- *Publisher's Username*: any publicly available information about the username that a major publisher uses in the Pirate Bay portal, and
- *Business Profile*: offered services (and choices) at the promoting URL.

Next, we describe our approach for collecting this information.

Promoting URL: We emulate the experience of a user by downloading a few randomly-selected files published by each top publisher to determine whether and where they may encounter a promoting URL. We identified

⁴We tried to download these files a few weeks after the correspondent measurement study was performed.

⁵<http://www.prevx.com/filenames/X2669713580830956212-X1/FLVDIRECT.EXE.html>

three places where publishers may embed a promoting URL: (*i*) name of the downloaded file (*e.g.* user mois20 names his files as *filename-divxatope.com*, thus advertising the URL www.divxatope.com), (*ii*) the textbox in the web page associated with each published content, (*iii*) name of a text file that is distributed with the actual content and is displayed by the BitTorrent software when opening the .torrent file. Our investigation indicates that the second approach (using the textbox) is the most common technique among the publishers.

Publisher's Username: We browsed the Internet to learn more information about the username associated with each top publisher. First, the username is in some cases directly related to the URL (*e.g.* user UltraTorrants whose URL is www.ultratorrants.com). This exercise also reveals whether this username publishes on other major BitTorrent portals in addition to the Pirate Bay. Finally, posted information in various forums could reveal (among other things) the promoting web site.

Business Services: We characterize the type of services offered at the promoting URL and ways that the web site may generate incomes (*e.g.* posting ads). We also capture the exchanged HTTP headers between a web browser and the promoting URL to identify any established connection to third-party web sites (*e.g.* redirection to ads web sites or some third party aggregator) using the technique described in [14].

5.1 Classifying Publishers

Using the above methodology, we examined a few published torrents for each one of the top publishers as well as sample torrents for 100 randomly selected publishers that are not in the top-100, called *regular publishers*. On the one hand, we did not discover any interesting or unusual behavior in torrents published by regular publishers and thus conclude that they behave in an altruistic manner. On the other hand, a large fraction of seeded torrents by the top publishers systematically promote one or more web sites with financial incentives. Our examination revealed that these publishers often include a promotional URL in the textbox of the content web page. We classify these top publishers into the following three groups based on their type of business (using the content of their promoting web sites) and describe how they leverage BitTorrent portals to intercept and redirect users to their web sites.

Private BitTorrent Trackers: A subset of major publishers (25% of top) own their BitTorrent portals that are in some cases associated with private trackers [11]. These private trackers offer a better user experience in terms of download rate (compared to major open BitTorrent portals) but require clients to maintain certain seeding ratio. More specifically, each participating BitTorrent client is required to seed content propor-

tionally to the amount of data it downloads across multiple torrents. To achieve this goal, users are required to register in the web site and login before downloading the torrent files. The publishers in this class publish 18% of all the content while they are responsible for 29% of the downloads. 2/3 of these publishers advertise the URL in the textbox at the content web page. Furthermore, they appear to gain financial profit in three different ways: (i) posting advertisement in their web sites, (ii) seeking donations from visitors to continue their basic service, and (iii) collecting a fee for VIP access that allows the client to download any content without requiring any kind of seeding ratio. These publishers typically inject video, audio and application content into BitTorrent portals. Interestingly, a significant fraction of publishers in this class (40%) publish content in a specific language (Italian, Dutch, Spanish or Swedish) and specifically a 66% of this group are dedicated to publishing Spanish content. This finding is consistent with prior reports on the high level of copyright infringement in Spain [5].

Promoting Web Sites: Another class of top publishers (23% of top) promote some URLs that are associated with hosting images web sites (*e.g.* `www.pixsor.com`), forums or even religious groups (*e.g.* `lightmiddleway.com`). These publishers inject 8% of the content and are responsible of 11% of the downloads. Most publishers in this class advertise their URL using the textbox in the content web page. Furthermore, most of these publishers (70%), specifically those that are running a hosting image web site, publish only porn content. Inspection of the associated hosting image web sites revealed that they store adult pictures. Therefore, by publishing porn content in major BitTorrent portals, they are targeting a particular demography of users who are likely to be interested in their web sites. The incomes of the portals within this class is based on advertisement.

Altruistic Publisher: The remaining top publishers (52% of top) appear to be altruistic users since they do not seem to directly promote any URL. These publishers are responsible of 11.5% of the content and the same fraction of downloads. Many of these users publish small music and e-book files that require smaller amount of seeding resources. Furthermore, they typically include a very extensive description of the content and often ask other users to help with seeding the content. These evidences suggest that these publishers may have limited resources and thus they need the help of others to sustain the distribution of their content.

In summary, roughly half of the top publishers advertise a web portal in their published torrents. It seems that their intention is to attract a large number of users to their web sites. The incomes of these portals come from ads and in the specific case of private BitTorrent portals also from donations and VIP fees. Overall, these

| | Lifetime (days) | Avg. Publishing Rate (torrents per day) |
|---------------------|--------------------|--|
| Private Portals | 63/466/1816 | 0.57/11.43/79.91 |
| Promoting Web sites | 50/459/1989 | 0.38/4.31/18.98 |
| Altruistic | 10/376/1899 | 0.10/3.80/23.67 |

Table 4: Lifetime and Avg. Publishing Rate for the different classes of content publishers: BitTorrent Portals, Promoting Web Sites and Altruistic Publishers. The represented values are min/avg/max per class.

profit-driven publishers generate 26% of the content and 40% of the downloads. Therefore, the removal of this small fraction of publishers is likely to have a dramatic impact on the BitTorrent open ecosystem. Finally, a fraction of publishers appear to be altruistic and responsible for a notable fraction of published content and downloads (11.5%). This suggests that there are some seemingly ordinary users who dedicate their resources to share content with a large number of peers in spite of the potential legal implications of such activity.

5.2 Longitudinal View of Major Publishers

So far we focused on the contribution of major publishers only during our measurement intervals. Having identified the top publishers in our *pb10* dataset, we examine the longitudinal view of the contribution by major publishers since they appeared on the Pirate Bay portal. Toward this end, for each top publisher, we obtain the username page on the Pirate Bay portal that maintains the information about all the published content and its published time by the corresponding user till our measurement date (June 4, 2010)⁶. Using this information for all top publishers, we capture their publishing pattern over time with the following parameters: (i) *Publisher Lifetime* which represents the number of days between the first and the last appearance of the publisher in the Pirate Bay portal, (ii) *Average Publishing Rate* that indicates the average number of published content per day during their lifetime.

Table 4 shows the min/avg/max value of these metrics for the different classes of publishers: Private Portals, Promoting Web Sites and Altruistic publishers. The profit-driven publishers (*i.e.* private portals and promoting web sites) have been publishing content for 15 months on average (at the time of the measurement) while the most longed-lived ones have been feeding content for more than 5 years ago. Furthermore, some of these publishers exhibit a surprisingly high average rate of publishing content (80 files per day). The altruistic publishers present a shorter lifetime and a lower publishing rate that seems to be due to their weaker incen-

⁶Note that we cannot collect information about fake publishers since the web pages of their associated usernames are removed by the Pirate Bay just after identifying they are publishing fake content.

tives and their lower availability of resources.

In summary, the lifetime of major publishers suggests that content publishing in BitTorrent seems have been a profitable business for (at least) a couple of years. Furthermore, the high seeding activity by profit-driven publishers (e.g. private portals) over a long period of time implies a high and continuous investment for required resources that should be compensated by different types of incomes (e.g. ads) for these portals. We examine the incomes of the profit-driven publishers in the next subsection.

5.3 Estimating Publishers’ Income

The evidences we presented in previous subsections suggest that the goal of half of the top publishers is to attract users to their own web sites. We also showed that most of these publishers seem to generate income at least by posting ads in their web sites. In essence, these publishers have a clear financial incentive to publish content. In order to validate this key point, we assess their ability to generate income by estimating three important but related properties of their promoting web sites: (i) average value of the web site, (ii) average daily income of the web site, and (iii) average daily visits to the web site. We obtain this information from several web sites that monitor and report these statistics for other sites on the Web. To reduce any potential error in the provided statistics by individual monitoring web sites, for each publisher’s web site we collect this information from six independent monitoring sites and use the average value of these statistics across these webs⁷.

Table 5 presents the min/median/avg/max value of the described metrics for each class of profit-driven publisher classes (i.e. private portals and promoting web sites). The median values suggest that the promoted web sites are fairly profitable since they value tens of thousand dollars with daily incomes of a few hundred dollars and tens of thousand visits per day. Furthermore, few publishers (<10) are associated to very profitable web sites valued in hundreds of thousand to millions of dollars, that receive daily incomes of thousands of dollars and hundreds of thousand visits per day. In summary, these statistics confirm that these web sites are indeed valuable and visible, and generate a substantial level of incomes.

6. OTHER BENEFICIARIES IN THE BITTORRENT MARKETPLACE

In previous sections we analyzed the main characteristics of major content publishers in BitTorrent, demonstrating that content publishing is a profitable *business* for an important fraction of the top publishers;

⁷www.sitelogr.com, www.cwire.com, www.websiteoutlook.com, www.sitevaluecalculator.com, www.mywebsiteworth.com, www.yourwebsitevalue.com

| | Web site Value (\$) | Web site Daily Income (\$) | Web site Daily visits |
|---------------------|---------------------|----------------------------|-----------------------|
| Private Portals | 1K/33K/313K/2.8M | 1/55/440/3.7K | 74/21K/174K/1.4M |
| Promoting Web Sites | 24/22K/142K/1.8M | 1/51/205/1.9K | 7/22K/73.5K/772K |

Table 5: Publisher’s web site value (\$), daily income (\$) and num of daily visits for the different classes of profit-driven content publishers: BitTorrent Portals and Promoting Web Sites. The represented values are min/median/avg/max per class.

a business that is responsible of 40% of the downloads. However, although content publishers are the key players, there are other players who help sustain the business and obtain financial benefits including: *Major BitTorrent Portals*, *Hosting Providers* and *ad companies*. Figure 5 depicts the interactions between different players in BitTorrent content publishing where the arrows indicate the flow of money between them. Next, we briefly describe the role of each player.

Major Public BitTorrent Portals such as the Pirate Bay are dedicated to index torrent files. They are rendezvous points where content publishers and clients publish and retrieve torrent files respectively. The main advantage of these major portals is that they offer a reliable service (e.g. they rapidly react to remove fake or infected content). All this makes that millions of BitTorrent users utilize these portals every day. These portals are the perfect target for profit-driven publishers in order to publish their torrents and advertise their web sites (potentially) to millions of users. Therefore, these major portals are one of the key players of the BitTorrent Ecosystem [22] that brings substantial financial profit. For instance, the Pirate Bay is one of the most popular sites in the whole Internet (ranked the 94th in the Alexa Ranking) as well as one of the most valued ones (around \$10M).

Hosting Providers are companies dedicated to renting servers. Heavy seeding activity performed by some publishers requires significant resources (e.g. bandwidth and storage). Thus a large fraction of major publishers obtain these resources from rented servers in hosting providers who receive an income in return for the offered service. Let’s focus on OVH, the ISP responsible of a major portion of the content published in BitTorrent. Our measurement study shows that OVH contributes between 78 and 164 different servers (i.e. IP addresses) across the different datasets. Considering the cost of the average server offered by OVH in its web page (around 300 €/month) we estimate that the average income obtained by OVH due to BitTorrent content publishing ranges between roughly 23.4K and 42.9K €/month. It is worth noting that some hosting providers have defined strict policies against sharing copyrighted material through P2P applications using their servers due

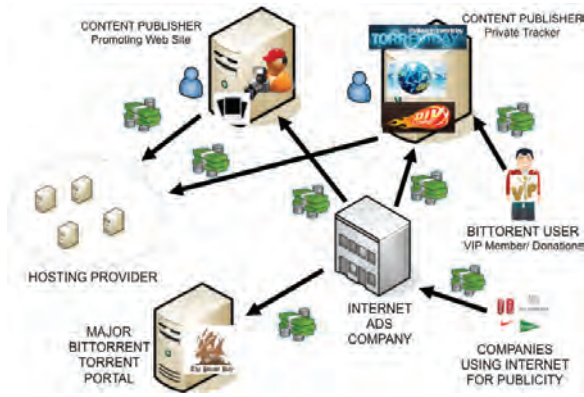


Figure 5: Business Model of Content Publishing in Bittorrent.

to legal issues⁸. However the income obtained by some hosting providers such as OVH seems to justify the risk of potential legal actions taken against them.

Ad Companies are responsible for advertisements in the Internet. They have a set of customers who wish to be advertised in the Internet and a set of web sites where they put their customers ads. They apply complex algorithms to dynamically select in which web site and when to post each ad. They charge their customers for this service and part of this income is forwarded to the web sites where the ads have been posted. Therefore, ad companies look for popular web sites where to put ads for their costumers. We have demonstrated in this paper that profit-driven BitTorrent content publishers' web sites are popular, thus most of them post ads from ad companies⁹. Hence, part of the income of ad companies is directly linked to the BitTorrent content publishing. Unfortunately, there is no practical way to estimate the value of this income.

7. SOFTWARE FOR CONTENT PUBLISHING MONITORING

Using the methodology described in Section 2, we have implemented a system that continuously monitors the Pirate Bay portal, and leverages RSS feed to quickly detect a newly published content. Once a new torrent is detected, our system retrieves the following information for that torrent: filename, content category and subcategory (based on the Pirate Bay categories), publisher's username, and (in those cases we can) the publisher's IP address as well as the ISP, City and Country associated to this IP address. Furthermore, for profit-driven publishers described in this paper, we have created an individual publisher's web page that provides specific information such as the publisher's promoted URL or

⁸<http://www.serverintellect.com/terms/aup.aspx>.

⁹We have validated this by looking at the header exchange between the browser and the publishers' web site servers.

business type. The system stores all this information in a database. Finally, we have built a simple web-based interface to query the resulting database. This interface is publicly available¹⁰ and access is granted to interested parties by contacting the authors.

Our application has two goals. On the one hand, we want to share this data with the research community to permit further analysis of different aspects of the BitTorrent content publishing activity. On the other hand, we believe that this application can be useful for regular BitTorrent clients. First, a BitTorrent client can easily identify those publishers that publish content aligned with her interest (*e.g.* an e-books consumer could find publishers responsible for publishing large numbers of e-books). Furthermore, we are working on implementing a feature to filter out fake publishers, allowing BitTorrent users of our application (in the future) to avoid downloading fake content.

8. RELATED WORK

Significant research effort has focused on understanding different aspects of BitTorrent by gathering data from live swarms. Most of these studies have primarily examined demographics of users [12, 20, 22] and technical aspects of swarming mechanism [18, 13, 9]. However, to our knowledge the socio-economics aspects of BitTorrent that we addressed in this paper have received little attention. The most relevant work to this paper is a recent study that examined the weakness of BitTorrent privacy [6]. The authors analyzed the demography of BitTorrent content publishers and presented a highly skewed distribution of published content among them as well as the presence of a significant fraction of publishers located at hosting providers. This indeed validates some of our initial observations. In another study, Zhang et al. [22] presented the most extensive characterization of the BitTorrent ecosystem. This study briefly examined the demography of content publishers and showed a skewed distribution of the contributed content among them. The authors identify the publishers by their usernames. We have shown that this assumption may miss an important group of publishers who post fake content, *i.e.* fake publishers. Our work goes beyond the simple examination of demographics of content publishers. We identify, characterize and classify the major publishers and more interestingly reveal their incentives and their motivating business model.

9. CONCLUSION

In this paper we studied the content publishing activity in BitTorrent from a socio-economic perspective. The results reveal that a small fraction of publishers are responsible for 67% of the published content and 75% of

¹⁰<http://bittorrentcontentpublishers.netcom.it.uc3m.es/>

the downloads. We have carefully examined the incentives of major publishers and identified the following key characteristics: first, antipiracy agencies and malicious users perform a systematic poisoning index attack over major BitTorrent portals in order to obstruct download of copyrighted content and to spread malware, respectively. Overall, this attack contributes 30% of the content and attracts 25% (several millions) of downloads. Second, 37% of the (non-fake) published content is published by a small fraction of users that serve 54% of the (non-fake files) downloads. Our evidence suggests that these publishers have financial incentives for posting these contents on BitTorrent portals. The removal of these financial-driven publishers (e.g. by antipiracy actions) may significantly affect the popularity of these portals as well as the whole BitTorrent ecosystem. If this happens, *will BitTorrent survive as the most popular file-sharing application without these publishers?*

10. ACKNOWLEDGEMENTS

The authors would like to thank anonymous CoNEXT reviewers for their valuable feedback as well as Jon Crowcroft for his insightful comments. We are also grateful for David Rubio's help on developing the web-based application. This work has been partially supported by the European Union through the FP7 TREND Project (257740), the Spanish Government through the T2C2 (TIN2008-06739-C04-01) and CONPARTE (TEC 2007-67966-C03-03) projects, the Regional Government of Madrid through the MEDIANET project (S-2009/TIC-1468) and the National Science Foundation under Grant No. 0917381.

11. REFERENCES

- [1] http://www.theregister.co.uk/2008/06/02/onk_further_arrests/.
- [2] <http://yro.slashdot.org/yro/05/01/13/2248252.shtml?id=123&tid=97&tid=95&tid=1>.
- [3] MaxMind. <http://www.maxmind.com>.
- [4] The Impact of P2P File Sharing, Voice over IP, Skype, Joost, Instant Messaging, One-Click Hosting and Media Streaming such as YouTube on the Internet, 2007. <http://www.ipoque.com/resources/internet-studies/internet-study-2007>.
- [5] Bay TSP Annual Report, 2008. <http://tech.mit.edu/V129/N28/piracy/BayTSP2008report.pdf>.
- [6] S. Le Blond, A. Legout, F. Lefessant, W. Dabbous, and M. Ali Kaafar. Spying the world from your laptop. *LEET'10*, 2010.
- [7] D. R. Choffnes, J. Duch, D. Malmgren, R. Guimera, F. E. Bustamante, and L. Amaral. Strange bedfellows: Communities in bittorrent. *IPTPS'10*, 2010.
- [8] R. Cuevas, M. Kryczka, A. Cuevas, S. Kaune, C. Guerrero, and R. Rejaie. Is content publishing in bittorrent altruistic or profit-driven?. Tech report: <http://arxiv.org/abs/1007.2327>, 2010.
- [9] R. Cuevas, N. Laoutaris, X. Yang, G. Siganos, and P. Rodriguez. Deep Diving into BitTorrent Locality. In *ACM SIGMETRICS'10 (Poster Session)*.
- [10] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang. Measurements, Analysis, and Modeling of BitTorrent-like Systems. In *ACM IMC'05*.
- [11] David Hales, Rameez Rahman, Boxun Zhang, Michel Meulpolder, and Johan Pouwelse. Bittorrent or bitcrunch: Evidence of a credit squeeze in bittorrent? In *WETICE '09*, 2009.
- [12] M. Izal, G. Urvoy-Keller, E.W. Biersack, P.A. Felber, A. Al Hamra, and L. Garces-Erice. Dissecting bittorrent: Five months in a torrent's lifetime. In *PAM '04*.
- [13] S. Kaune, R. Cuevas, G. Tyson, A. Mauthe, C. Guerrero, and R. Steinmetz. Unraveling BitTorrent's File Unavailability: Measurements, Analysis and Solution Exploration. In *IEEE P2P'10*.
- [14] B. Krishnamurthy and C. E. Wills. On the leakage of personally identifiable information via online social networks. In *ACM WOSN '09*, 2009.
- [15] N. Laoutaris, D. Carra, and P. Michiardi. Uplink allocation beyond choke/unchoke or how to divide and conquer best. In *ACM CoNEXT'08*.
- [16] A. Legout, N. Liogkas, E. Kohler, and L. Zhang. Clustering and sharing incentives in bittorrent systems. In *ACM SIGMETRICS '07*.
- [17] J. Liang, N. Naoumov, and K.W. Ross. The index poisoning attack in p2p file sharing systems. In *IEEE INFOCOM'06*, 2006.
- [18] D. Menasche, A. Rocha, B. Li, D. Towsley, and A. Venkataramani. Content availability in swarming systems: Models, measurements and bundling implications. In *ACM CoNEXT'09*.
- [19] M. Piatek, T. Isdal, T. Anderson, A. Krishnamurthy, and A. Venkataramani. Do incentives build robustness in BitTorrent? In *NSDI'07*, 2007.
- [20] J.A. Pouwelse, P. Garbacki, D.H.J. Epema, and H.J. Sips. The BitTorrent P2P file-sharing system: Measurements and analysis. In *IPTPS'05*.
- [21] M. Sirivianos, J. Han, P. Rex, and C.X. Yang. Free-riding in bittorrent networks with the large view exploit. In *IPTPS '07*, 2007.
- [22] C. Zhang, P. Dhungel, D. Wu, and K.W. Ross. Unraveling the bittorrent ecosystem. *IEEE Transactions on Parallel and Distributed Systems*, 2010.