Unraveling the BitTorrent Ecosystem

Chao Zhang, *Student Member, IEEE,* Prithula Dhungel, *Student Member, IEEE,* Di Wu, *Member, IEEE,* and Keith W. Ross, *Fellow, IEEE*

***Abstract*—BitTorrent is the most successful open Internet application for content distribution. Despite its importance, both in terms of its footprint in the Internet and the influence it has on emerging P2P applications, the BitTorrent Ecosystem is only partially understood. We seek to provide a nearly complete picture of the entire public BitTorrent Ecosystem. To this end, we crawl five of the most popular torrent-discovery sites over a nine-month period, identifying all of 4.6 million torrents and 38,996 trackers that the sites reference. We also develop a high-performance tracker crawler, and over a narrow window of twelve hours, crawl essentially all of the public Ecosystem's trackers, obtaining peer lists for all referenced torrents. Complementing the torrent-discovery site and tracker crawling, we further crawl Azureus and Mainline DHTs for a random sample of torrents. Our resulting measurement data is more than an order of magnitude larger (in terms of number of torrents, trackers, or peers) than any earlier study. Using this extensive data set, we study in-depth the Ecosystem's torrent-discovery, tracker, peer, user behavior, and content landscapes. For peer statistics, the analysis is based on one typical snapshot obtained over twelve hours. We further analyze the fragility of the Ecosystem upon the removal of its most important tracker service.**

I. Introduction

BitTorrent is a remarkably popular file-distribution technology, with millions of users sharing content in hundreds of thousands of torrents on a daily basis. Even in the era of YouTube, BitTorrent traffic continues to grow at impressive rates. For example, downloads of .torrent files from Mininova's site doubled in 2008, to nearly 7 million downloads in a year [1]. BitTorrent has proven to be particularly effective at distributing large files, including open-source software distributions.

Fundamental to BitTorrent's success is its openness – the BitTorrent protocol has been published, and the source code of the baseline implementation has been made widely available. This openness has enabled developers to create over 50 independent BitTorrent client implementations [2], dozens of independent tracker implementations [3], and a multitude of torrent-discovery sites. The openness of the protocol has fostered productive discussions in both the online developer and the research communities, leading to further design improvements. All of this flourishing BitTorrent technology taken together forms the *BitTorrent Ecosystem*, consisting of millions of BitTorrent peers, hundreds of active trackers, and dozens of torrent discovery sites (see Figure 1).

Chao Zhang, Prithula Dhungel and Keith W. Ross are with the Department of Computer Science & Engineering, Polytechnic Institute of NYU, Brooklyn, NY, USA 11201 (e-mail: chao@cis.poly.edu, pdhung01@students.poly.edu, ross@poly.edu).

Di Wu is with the Department of Computer Science, Sun Yat-Sen University, China (e-mail: wudi27@mail.sysu.edu.cn).

BitTorrent is not only a thriving file distribution system, but also serves as a model for many successful live and on-demand P2P video deployments. About a dozen companies in China today (including Coolstreaming, PPLive, ppstream) use the P2P paradigm to distribute live Chinese television channels, as well as live international content, to Internet users throughout the world. Most of these deployments are very similar to BitTorrent, with peers informing each other of the pieces they have, and the peers then downloading from each other their missing pieces [4]. There is also a multitude of companies today that are deploying P2P Video-on-Demand (VoD) [5]. To create these P2P VoD and live video systems, designers essentially began with the BitTorrent architecture, removed the tit-for-tat, and modified the scheduling algorithm to give priority to blocks that are to be played in the near future.

Many communities – including P2P researchers and designers, ISP researchers, copyright holders, and pedophilia and terrorist law enforcement agencies – would like to have a comprehensive and in-depth understanding of the Ecosystem in its entirety, as well as tools and methodologies for mapping the Ecosystem in the future. Despite its importance, both in terms of its footprint in the Internet and the influence it has on emerging P2P applications, the BitTorrent Ecosystem is only partially understood today. Although there are a few studies that provide limited insights, there woefully lacks an up-to-date and comprehensive picture of the Ecosystem. *However, because BitTorrent is an ecosystem involving hundreds of independently operated trackers and torrent discovery sites (public and private), as well as millions of concurrently-active peers (using many different client implementations), it is a major challenge to provide a complete snapshot that spans the entire Ecosystem. No single torrent-discovery Web site, tracker or ISP can provide the complete picture on its own.*

In this paper we aim at providing a comprehensive and up-to-date picture of the BitTorrent Ecosystem. To keep our project manageable, we limit our attention to the public English-language BitTorrent ecosystem. We do this by crawling five of the most popular discovery sites over a nine-month period, and identify all of the torrents and trackers that the sites reference. Then, over a narrow window of twelve-hours, we crawl all the trackers referenced by the discovery sites, obtaining peer lists for all referenced torrents. Our measurement data is more than an order of magnitude larger (in terms of number of torrents, trackers, or peers) than any earlier study [6], [7], [8]. The collected data has been anonymized and is publicly available to the research community[1]. Using this extensive data set, we undertake an

[1]Anonymized data available at http://cis.poly.edu/~chao/bt-ecosys.html

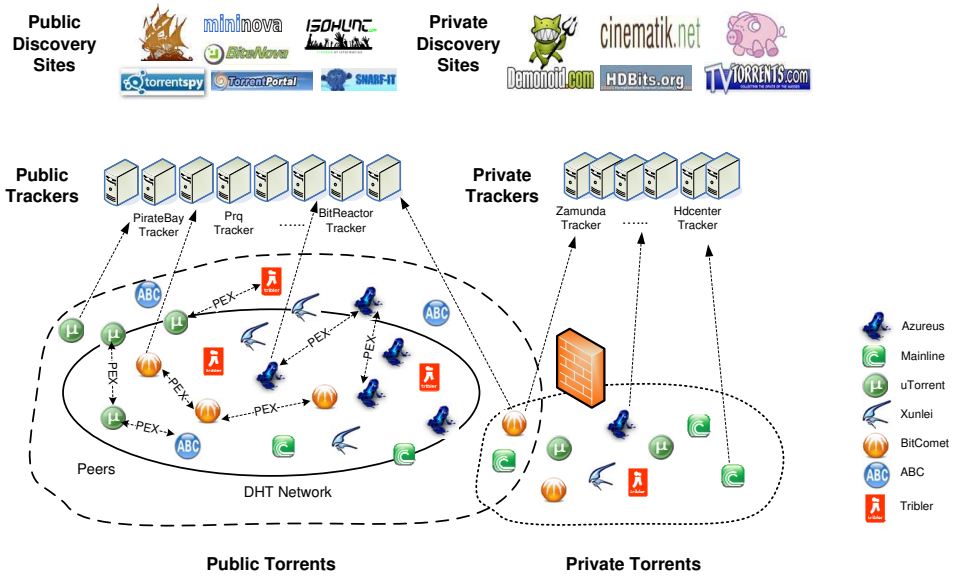

Fig. 1. The BitTorrent Ecosystem (Note that Azureus is now called Vuze)

in-depth analysis, leading to a nearly complete picture of the entire public Ecosystem.

The contributions and some of the findings of this paper are as follows:

- Based on an asynchronous I/O design, we develop a high-performance multi-tracker crawler that simultaneously crawls thousands of trackers with concurrent TCP connections. The asynchronous I/O design provides a significant performance improvement over multi-thread designs, allowing us to obtain a snapshot of the Ecosystem's millions of torrents in about twelve hours. We also develop a discovery-site crawler and adapt it to crawl five major torrent-discovery sites (Mininova, Pirate Bay, BTmonster, Torrent Reactor, and Torrent Portal).
- Using the .torrent files and meta-data Web pages from the five sites, we study the Ecosystem's torrent-discovery landscape. We find that these five sites collectively index 4.6 million unique torrents, but approximately only 1.2 million of them are active. We investigate the degree of indexing overlap among the sites, the characteristics and motivations of the users who upload .torrent files, and how the sites acquire .torrent files. We find that none of the major torrent sites on its own provides a complete picture of the Ecosystem.
- We study the Ecosystem's tracker landscape. We identify almost 39 thousand trackers, although only 728 (less than 2%) of these trackers are active. We determine the number of torrents and peers tracked by each of the active trackers. We find that although the Top 20 tracker organizations are hosted in many continents, but with a high concentration in northern Europe. We find, for example, that 40% of the trackers track no more than four active torrents, and that only 190 trackers track more than 1,000 peers. Pirate Bay, operating the largest trackers, plays a disproportionate role in the Ecosystem.
- We study the Ecosystem's peer landscape. Our analysis is based on one typical snapshot obtained within twelve hours. We find that the Ecosystem is rich in long-tail content, with many torrents being very small (82% have no more than 10 peers). Although the Ecosystem is dominated by mice, there are also elephants, with the largest torrents having more than 10,000 simultaneous peers. We investigate the number of torrents a peer joins simultaneously and the geographical distributions of the peers. We determine the countries in which BitTorrent has the highest usage per Internet user. We investigate the distribution of client types being used today, and determine that more than 50% the peers today use uTorrent. We also study which clients are being used to create .torrent files and initialize torrents.
- We study the Ecosystem's content landscape. By classifying each of the active torrents into one of ten categories (movies, music, TV shows, pornography, and so on), we determine which content types are most popular in BitTorrent today. We also perform a geographical analysis, determining in which countries movies, music and pornography are most popular per Internet user. We analyze the size of the content files being distributed in the Ecosystem for each of the categories.
- Although we find that the Ecosystem is in general highly diverse, it also contains a major pillar, namely, Pirate Bay's tracking service. We find that Pirate Bay currently tracks 90% of the torrents in the Ecosystem. We undertake an analysis of the Ecosystem's fragility to Pirate Bay, considering whether current DHT and PEX decentralized tracking services can pick up the slack.

Perhaps the most important take-away message from this paper is a vivid and complete picture of the most successful open Internet application in the current decade.

This paper is organized as follows. Section II provides an overview of the BitTorrent Ecosystem. Section III describes the measurement methodology and scope. Section IV provides

the measurement results for the torrent-discovery, tracker, and peer landscapes. Section V provides content and geography classification for the torrents. The importance of Pirate Bay to the Ecosystem is analyzed in Section VI. Section VII describes related work, and we conclude in Section VIII.

II. The BitTorrent Ecosystem

As shown in Figure 1, the BitTorrent ecosystem consists of three major components: *peers*, *peer discovery mechanisms*, and *torrent-discovery sites*. The collection of peers that participate in the distribution of a specific file at given time is called a *torrent*. Each torrent is identified with a torrent identifier called the *infohash*. At any given instant of time, each peer in a torrent is either a leecher or a seed; a seed possesses the entire file, whereas a leecher possesses only a portion of the file. Typically a torrent begins with an *initial seed*, which is the only peer to have the file. Each leecher and seed uses one of the many *BitTorrent client types*. More than 50 BitTorrent client types have been developed to date [2]. BitTorrent peers communicate with each other using the (open) BitTorrent protocol.

One common mechanism for peer discovery is to use a *tracker*. When a peer joins a torrent, it typically registers with one or more trackers. Any peer can contact a tracker at any time to obtain a random subset (IP-port pairs) of other peers in the torrent. Today there are at least a dozen tracker implementations [3]. Many BitTorrent clients also support "distributed trackers" using DHTs and Peer Exchange (PEX). For example, Azureus clients collectively form a DHT: an Azureus client can query the DHT, using an infohash for the key, to obtain a list of Azureus peers participating in the torrent. Furthermore, many BitTorrent clients (including Azureus and uTorrent) also employ PEX, which is a gossiping mechanism that allows peers in the same torrent to exchange peer lists directly with each other. Thus many client types (e.g., Azureus and uTorrent) can discover peers using three distinct mechanisms: centralized trackers, DHTs, and PEX.

When a user wants to start a new torrent, it needs to seed the content file and register the torrent with a tracker. For example, a user can start a torrent by seeding the content file locally, registering the torrent with a tracker, and uploading a .torrent file (including tracker addresses) to a torrent-discovery site.

Users learn about the existence of ongoing torrents from torrent-discovery sites such as Pirate Bay, Mininova, Isohunt, BTmonster, and Torrent Portal. There are hundreds of torrent-discovery sites. Table I shows the most-visited English-language (public) torrent discovery sites, obtained from the Web-traffic monitoring site Alexa [9]. Some torrent-discovery sites provide tracker services and others do not; in the ten sites listed in Table I, we determined (see Section III and IV) that only Pirate Bay, Mininova and SumoTorrent provide tracker services. Most torrent-discovery sites, for each of their indexed torrents, provide a *.torrrent file*, which includes the IP addresses of one or more trackers and the hashes of all the pieces in the file. Except for Torrentz, all the sites in Table I provide .torrent files for the torrents they index. (Torrentz is solely a search engine, providing links to .torrent files at other sites.) Interestingly, eight of the top-10 discovery sites are in Europe, with the remaining two in North America.

Site	Alexa rank	Location	Comment
Mininova.org	88	Netherlands	Tracker
Thepiratebay.org	109	Sweden	Tracker
IsoHunt.com	219	Canada	-
Torrentz.com	225	Netherlands	No .torrent
Torrentreactor.net	454	Germany	-
Btjunkie.org	551	Sweden	-
Sumotorrent.com	1,271	Netherland	Tracker
Btmon.com	1,410	Germany	-
TorrentPortal.com	1,525	United States	-
GamesTorrents.com	2,247	Germany	-

TABLE I
Top-10 most popular public English-language torrent-discovery sites

Site	Language	Alexa rank	Location	Theme
torrents.ru	Russian	295	Russia	General
hdcenter.ath.cx	German	1,128	Luxembourg	HD Video
Zamunda.net	Bulgarian	1,760	Netherlands	General
PureTnA.com	English	2,291	Netherlands	Porn
Empornium.us	English	2,846	Netherlands	Porn
lostfilm.tv	Russian	3,368	Finland	General
gamato.info	Greek	4,065	Netherlands	General
arenabg.com	Bulgarian/ English	4,267	Netherlands	General
bwtorrents.com	English	5,414	Denmark	General
torrentleech.org	English	5,989	United States	General

TABLE II
Top-10 most popular private torrent-discovery sites

In addition to all this flourishing "public ecosystem" activity, there is also activity within private BitTorrent sites. A BitTorrent private site restricts who can use it, typically by requiring registration of a user account. These sites commonly use invitation systems for limiting registrations. Private torrent sites usually record how much the registered users upload and download, and typically enforce a minimum upload-to-download ratio on each user. Many private trackers implement passkeys in the .torrent file, which the user's client presents to the site's private tracker for authorization. Table II lists the top-10 private sites, again based on Alexa ranking. Although the world of private sites and torrents is important and interesting, this paper focuses on the public BitTorrent ecosystem, which in itself is enormously rich and complex, and challenging to capture in a single paper.

As a cavaet, we emphasize that BitTorrent is truly an international, multi-lingual phenomenon. The content in BitTorrent is in many different languages. Naturally, there are also torrent-discovery sites specific to different languages (Chinese, Russian, Spanish, and so on). To keep the scope of this ambitious project manageable, we have limited ourselves to investigating only the torrents indexed by English-language torrent-discovery sites (as well as the trackers and peers they engender). As we shall see, mapping the BitTorrent ecosystem engendered by the English-language sites is in itself an enormously challenging problem.

III. Methodology

A. Measurement Infrastructure

As shown in Figure 2, our measurement platform includes two crawlers and one storage system. (We will discuss enhancements of this platform, including DHT support in Section 5.)

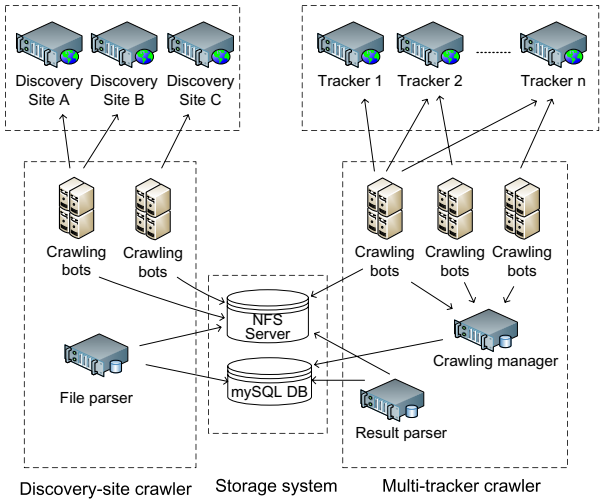

Fig. 2. Torrent-discovery and multi-tracker crawlers

The *discovery-site crawler* downloads web pages and .torrent files from torrent-discovery sites, and parses their contents to extract information of interest. We crawled five popular torrent-discovery sites (see Section 3), all of which use different formats for presenting torrent meta-data information. To improve efficiency, all the downloaded Web pages and .torrent files are first stored in the NFS file system temporarily; later, we use site-dependent parsers to extract information from the downloaded files, and store the information in the MySQL database. For different file types (e.g., .html, .torrent, .xml), we developed customized parsers to extract information in the files. For example, from the web pages, the parser extracts information about torrent category, torrent upload time, torrent uploader, the number of downloads, and so on; from the .torrent files, our parser extracts the torrent infohash, creation time, the list of trackers, data file size, and so on.

One of our goals is to obtain good estimates of the number of peers in a given torrent. Trackers support "scrape" querying, which returns, for a specific infohash, aggregate information including the total number of leechers and seeds. However, because there is significant overlap in the peer sets from different trackers, we cannot simply add the scraped results across trackers. Instead we need to obtain the IP/port pairs from each tracker, and then aggregate this lower-level data. The IP/port data can also provide a wealth of additional information, including geographical and user behavior information.

After obtaining all the infohashes indexed by the five torrent-discovery sites, and the list of trackers associated with each of the infohashes, the *multi-tracker crawler* determines the peers tracked for each (infohash, tracker) pair. This is done for each pair by repeatedly requesting peer lists. We use the term *task* for determining the peer list for a given (infohash, tracker) pair.

There are a number of challenges in designing a high-performance multi-tracker crawler. First, in order to get a snapshot of the ecosystem, the crawler must simultaneously crawl hundreds of trackers for millions of torrents. The traditional thread-pool model does not work well here, as the number of concurrent threads is limited by CPU/memory resources. Second, we need to control the crawling speed in order to avoid being banned by some trackers. Our multi-tracker crawler employs multiple tracker bots (Linux boxes) controlled by a master controller. To optimize the crawling speed, we use an asynchronous I/O model instead of the thread-pool model. The resulting tracker crawler can support more than one thousand concurrent TCP connections. To avoid being banned, we set a tunable parameter to limit the crawling speed and randomize the crawled targets to disperse the traffic evenly among multiple trackers. The multi-tracker crawler consists of one crawling manager and multiple crawling bots.

As a tracker only returns a random subset of the entire peer set for each query, multiple queries are required to get the complete peer set. Suppose that, for a given infohash, there are n peers registered with the tracker and the size of the subset returned is k; then according to [10], the expected number of queries to obtain all the peers is

$$E(n,k) \approx (\frac{n}{k} - \frac{k-1}{2k})L_n + \frac{k-1}{2k}$$

where L_n is the n-th harmonic number (i.e., the sum of the reciprocals of the first n natural numbers). When n is large, we have

$$E(n,k) \approx \frac{n}{k}(\gamma + \ln n)$$

where γ is the Euler-Mascheroni constant ($\gamma \approx 0.5772$). However, it is still difficult to determine the number of required queries to obtain the complete peer set. This is because (1) the value of n is unknown before crawling. Even when a torrent-discovery site provides the value for n, it may be inaccurate. (2) the above equations only give the expected number of queries. Therefore, we adopt a heuristic as the stopping criteria for a given task: if the bot does not see any new peers in two consecutive replies from the tracker, it assumes the peer list is almost complete and stops sending queries to that tracker for that infohash. Note that the above heuristic is not applicable if the tracker does not return a random peer list for each query.

B. Measurement Scope

Using 17 machines, we have been continuously crawling five major torrent discovery sites since 25 July 2008 until 22 April 2009. These five sites are *Mininova*, *Pirate Bay*, *Torrent Reactor*, *BTmonster*, and *Torrent Portal*. They are, respectively, the first, second, fifth, seventh, and ninth most popular English-language public torrent sites, as shown in Table I. As some torrent-discovery sites limit the rate at which an IP address can download its .torrent files, we restrict the speed of torrent-discovery crawling to avoid being banned. We first obtained all the web pages and .torrent files from those

sites; then continued to monitor these sites for new .torrent files[2].

We have collected approximately 8.8 million .torrent files from our discovery-site crawler, from which we obtained 4.6 million unique infohashes. We also discovered 38,996 trackers from the infohashes. After removing invalid tracker address, we created nearly 19 million unique crawling tasks (one task for each infohash-tracker pair).

Due to the large volume of tracker requests, we used 35 machines for the multi-tracker crawler (including 1 master controller, 34 crawling bots). We performed several tracker crawls between November 2008 and April 2009 [3]. Each crawl, performing the 19-million tasks (for each task, often requesting the peer list multiple times, since only a subset is returned), took about 12 hours. For each torrent, all the tasks for that torrent were performed over a short time span, typically less than a minute. Thus the multi-tracker crawler obtains a snapshot for each torrent, and collects one such snapshot for each of the infohashes in our database over a twelve-hour period. In this paper we only present one typical data set obtained in the early morning of 22 April 2009.

Before presenting the results of our measurements, to avoid any confusion it is useful to fix some terminology. We identify a *peer* by an < IP, port> pair. A peer can join multiple torrents at the same time. A *.torrent uploader* is a registered username on some torrent-discovery site that has uploaded at least one .torrent file. A *torrent* is the set of peers that participate in the distribution of a specific shared file identified by an infohash. A torrent is said to be an *active torrent* if the multi-tracker crawler finds at least one peer in the torrent. [4] A tracker is said to be an *active tracker* if it returns at least one peer for any of the queried infohashes.

We conclude this section with a few words about the scope of the measurement. Ideally, we would like the scope to include every active torrent indexed by the public English-language torrent-discovery sites. Also, ideally we would like to crawl all of the trackers that track these torrents. It is, however, infeasible to crawl each of the hundreds of torrent-discovery sites. In the next section we will show that by exhaustively crawling the five torrent-discovery sites, we capture the large majority of torrents and trackers.

IV. Measurements and Analysis

In this section we present the results from our torrent-discovery crawling and multi-tracker crawling. We examine the Ecosystem's discovery sites in Section 4.1, the Ecosystem's trackers in Section 4.2, and the Ecosystem's peers in Section 4.3.

[2]We were not able to obtain the Web pages at BTmonster but were able to obtain its .torrent files. For this reason, our data from BTmonster is less complete.

[3]By performing multiple crawls and comparing the results from different crawls, we are more confident about our measurement results.

[4]Although a torrent can change between active and inactive states during its lifetime, we only consider torrents that have at least one peer at the time of crawling as an active torrent.

A. Torrent-Discovery Sites

As described earlier, we found approximately 4.6 million unique infohashes. Using our multi-tracker crawler, we found that 1,192,303 of the torrents were active. Throughout the study, we focus on these 1.2 million active torrents. Table III shows the basic statistics of different torrent-discovery sites.

A torrent-discovery site learns about a torrent and obtains the associated .torrent file in one of two manners: either a registered user uploads the .torrent file directly to the site; or by crawling other torrent-discovery sites and downloading the associated .torrent file. Table III shows that all of the .torrent files are from uploaders on Pirate Bay. But for the other sites, the majority of the .torrent files come from search engines (e.g., 98.25% for Torrent Reactor). *Thus the major torrent-discovery sites are remarkably heterogeneous in their characteristics, with some sites indexing .torrent files solely from uploaders, others mostly from search engines.* Also, some major discovery sites (e.g., Torentz) do not index any .torrent files; and some discovery sites (for example, Pirate Bay and Mininova) provide complementary tracker services.

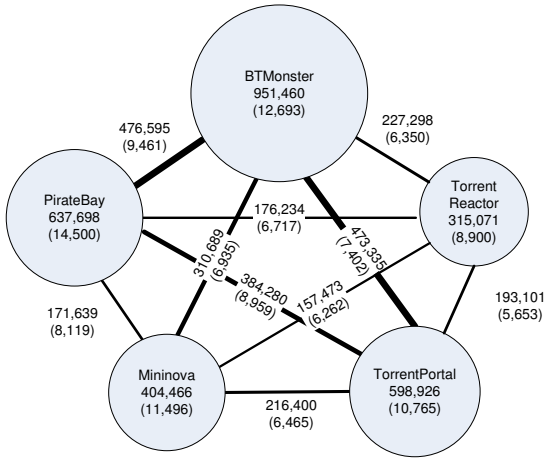

Fig. 3. Pairwise-intersection of active torrents and highly-active torrents for the five discovery sites (highly-active torrents, shown in parentheses, have more than 100 peers)

To gain deeper insight into the Ecosystem's torrent-discovery sites, we consider the overlap of their indexes. Figure 3 shows the pairwise-intersection of active torrents and highly-active torrents for the five discovery sites.The values for the highly active torrents are shown in parentheses. We define highly-active torrents as torrents with more than 100 peers (as determined by the multi-tracker crawler). We first observe that there is significant overlap among the sites for the highly-active torrents. For example, 71% of Mininova's and 75% of BTmonster's highly-active torrents are indexed by Pirate Bay. On the other hand, the overlap is much smaller for the set of all active torrents; for example, only 40% of Mininova's and 50% of BTmonster's active torrents are indexed on Pirate Bay. *We conclude that none of the major torrent site on its own provides a complete picture of the Ecosystem. Furthermore, there is significant diversity among the torrents that the sites index, particularly for less-active torrents with fewer than 100 peers.*

	Mininova	Pirate Bay	Torrent Portal	Torrent Reactor	BTMonster
# of collected .torrent files	1,454,165	902,614	2,008,285	565,270	3,832,119
Percentage of .torrent files from uploaders	36.1%	100%	34.24%	1.75%	N/A
# of active torrents	404,466	637,698	598,926	315,071	951,460
# of torrents with more than 100 peers	11,496	14,500	10,765	8,900	12,693

TABLE III
SUMMARY STATISTICS FROM TORRENT-DISCOVERY SITES(NOTE THAT ALL THE ABOVE FIVE TORRENT-DISCOVERY SITES PROVIDE .TORRENT FILES FOR THE TORRENTS THEY INDEX.)

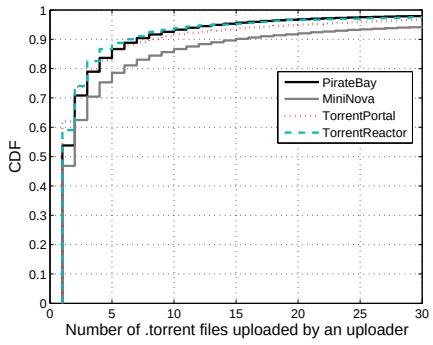

Fig. 4. CDF of number of .torrents by each uploader

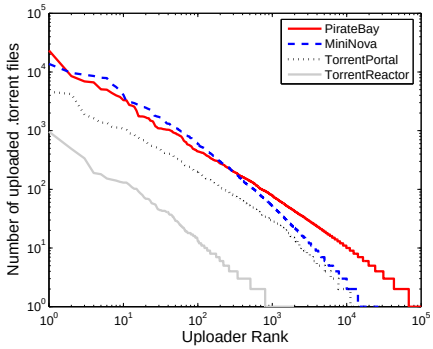

Fig. 5. Number of .torrents by each uploader

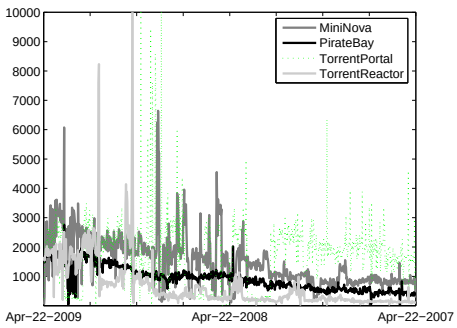

Fig. 6. .torrent indexing over past two years

We have also done an analysis of how many additional torrents each site brings beyond the union of the torrents indexed on the four other sites. For example, for highly-active torrents, the sites Torrent Portal, Mininova, and Torrent Reactor each only additionally bring 0.009%, 0.06% and 0.02% of the total number of active torrents. For small active torrents with no more than 10 peers, the corresponding percentages 0.05%, 1.4%, 0.9% are larger but remain small. Therefore, by including yet more torrent-discovery sites in our study, it is unlikely we would find significantly more active torrents. It is therefore likely that the scope of our study includes the large majority of torrents in the public (English-language) ecosystem.

Let's now take a look at BitTorrent uploaders. Figure 4 shows the CDF of the number of .torrent files uploaded by each uploader for four of the sites. We observe that, for all four sites, most of the uploaders upload only a few .torrent files. For example, 55% of the uploaders in Pirate Bay have uploaded only one .torrent file and more than 90% have uploaded less than 10 .torrent files. We conjecture that most of these infrequent uploaders of .torrent files are also the initial seeds for the corresponding content file. On the other hand, Figure 5 shows that there are also many elephants among the mice (typically less than 100), each uploading over 1,000 .torrent files. All curves in Figure 5 are straight in the log-log scale and indeed follow Zipf's law.

A natural question to consider is what is a user's incentive for initially seeding a file and uploading a corresponding .torrent file? Many of these uploaders likely act out of altruism, initiating a small number of torrents for distribution. But some users appear to be motivated by some form of self-interest. For example, the top uploader, who has uploaded over 23,000 .torrent files, is called *loder*. According to an article posted on the Slyck forum [11], loder is a founder of the private-torrent site MovieX.info [12]. In almost all of loder's .torrent files, there are two tracker addresses: a Pirate Bay tracker, and a MovieX tracker. MovieX's private tracker has been accused of only accepting connection requests from seeds and not from leechers from Pirate Bay. This has the effect of increasing the number of seeds within MovieX's private site.

Figure 6 shows the number of new .torrent files indexed per day over the past two years for four of the torrent-discovery sites [5]. Observe that the curve for Pirate Bay is relatively smooth, whereas the other sites (particularly Torrent Reactor in the last few months) have numerous sharp peaks. From this data, we can determine which sites crawl to obtain new .torrents. The sharp peaks indicate when the sites performed a crawl to obtain new .torrents. We observe that Pirate Bay has not been crawling since the very beginning. As the crawls may be executed by the administrators manually from time to time, the peaks do not exhibit a clear periodic pattern.

Figure 7 shows the CDF of the indexing times for the discovery sites relative to Pirate Bay. A positive value indicates indexing at an earlier time. It can be observed that for Pirate Bay and Torrent Portal, 80% of the torrents are indexed on the same day; for Pirate Bay and Torrent Reactor, this value is 60%. Given the knowledge that Pirate Bay has all its torrents from uploaders and that Torrent Portal and Torrent Reactor have a large fraction of torrents from their search engines, it can be concluded that these two sites are actively monitoring

[5]Although we started our crawling on July 2008, the torrent-discovery sites indexed many old torrents that were uploaded before July 2008. The upload time of .torrent files are extracted from the web pages crawled from those sites.

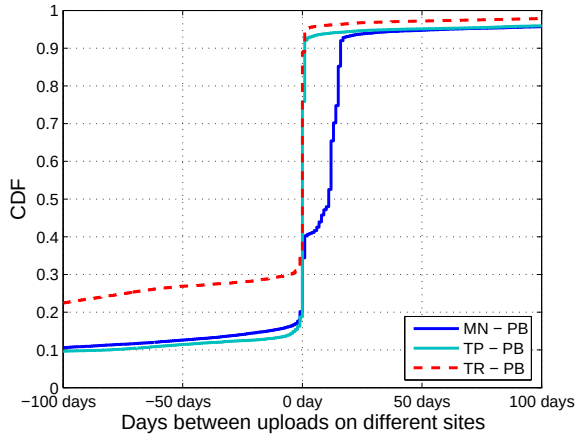

Fig. 7. Days between indexing of .torrents relative to Pirate Bay (MN: Mininova, TP: Torrent Portal, TR: Torrent Reactor, PB: Pirate Bay)

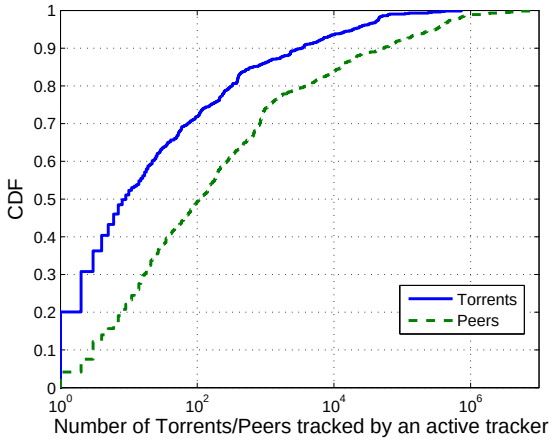

Fig. 8. CDF of the number of active torrents and peers tracked by each tracker

and indexing many of the new .torrents uploaded on Pirate Bay (but not all, as shown in Figure 3.) However, surprisingly, over 60% of .torrent files are indexed on Mininova before being indexed on Pirate Bay (and only around 18% of the .torrent files are indexed in Pirate Bay before Mininova.) From these observations it can be conjectured that many initial seeders first advertise their torrents on Mininova; in the subsequent weeks, the same or other users upload the .torrent files on Pirate Bay; then Torrent Reactor and Torrent Portal obtain the .torrent files by crawling [6].

B. Tracker Statistics

We now take a look at the BitTorrent tracker landscape. The discovery-site crawler found 38,996 trackers collectively indexed by the five discovery sites. For each of these trackers we checked to see if it tracked at least one active torrent (among the discovered torrents). Remarkably only 728, less than 2%, of these trackers are active. We explain this surprising behavior as follows. Some of the popular BT clients (e.g., Azureus and Mainline) allow users to create and locally host their own trackers. *Evidently, many of the trackers appearing in .torrent files point to ephemeral trackers created by users on their local machines with their local clients (uTorrent, Azureus, and so on).*

Table IV gives the top-20 tracker organizations, presenting both the number of active torrents and number of peers tracked by these organizations. (Note that any given tracker organization may have multiple trackers running on different hosts.) Remarkably, all of the tracker organizations track fewer than 1 million (peer, torrent) pairs, except for Pirate Bay, which tracks over 12.5 million (peer, torrent) pairs! Thus, although there are 14 organizations tracking over 100,000 (peer,torrent) pairs during the 12-hour snapshot, one organization – Pirate Bay – dominates the tracker landscape. Pirate Bay's trackers are hosted in Sweden. The other top-20 organizations are hosted in many continents, but with a high concentration in northern Europe.

Figure 8 shows the CDF of the number of active torrents and peers tracked by a tracker. We see that 40% of trackers track no more than four active torrents, and that about 32% of trackers track between 5 and 100 active torrents. Only 28% of trackers track more than 100 torrents, and only 190 trackers (26%) track more than 1,000 peers. *Thus, the Ecosystem contains tremendous tracker diversity, with a few huge trackers, a significant fraction of small trackers* [7]*, and yet another significant fraction of medium-size trackers.*

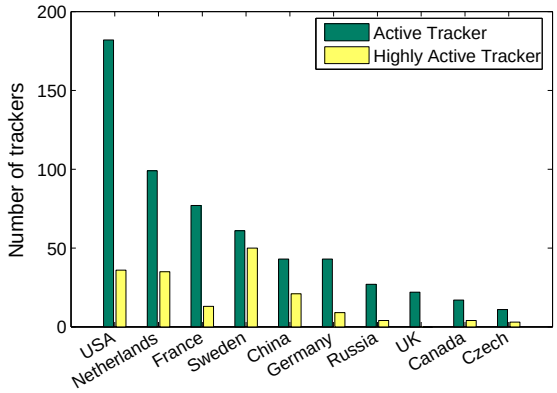

Fig. 9. Geographical distribution of active trackers and highly-active trackers

Figure 9 provides the geographical distribution of both active trackers and of highly-active trackers (with more than 1,000 peers). Throughout this paper, we use MaxMind GeoIP [13] to map IP addresses to countries. As to be expected, there are many active trackers in the USA, as it has a large population of Internet users. However, when we consider the highly-active trackers, we see that Sweden is the clear winner,

[6]Note that the uploaders of the same .torrent file may be different in different sites. As Mininova is the most popular indexing site, uploaders often prefer to first upload interesting content there. Most likely someone who downloaded that file from Mininova then uploaded it to other sites. It is also difficult to infer whether the uploader on Mininova is also the uploader of other sites simply based on uploader IDs. We have observed that the lists of Top 10 uploaders of different sites are almost completely different.

[7]The small trackers are most likely run by individuals using the tracker options in their clients

Tracker organization	Number of Tracked Peers	Number of Tracked Torrents	Location
thepiratebay.org	12,883,329	1,025,864	Sweden
rarbg.com	949,584	88,000	Sweden
torrent.to	540,476	44,084	Luxembourg
bitreactor.to	346,521	17,371	Sweden
mightynova.com	313,743	41,782	United States
torrent-downloads.to	269,104	26,776	Netherlands
paradise-tracker.com	253,851	14,827	Netherlands
torrent-download.to	245,952	14,386	Netherlands
tntvillage.org	212,319	17,943	Czech Republic
bittorrent.am	182,472	6,998	Russian Federation
9you.com	172,349	12,996	China
sharego.net	118,226	3,120	Netherlands
1337x.org	108,064	3,666	Sweden
divxfinal.com	107,979	3,520	Spain
megashara.com	94,135	3,692	Russian Federation
taquilladivx.com	93,147	1,243	Sweden
smartorrent.com	77,081	801	Canada
nyaatorrents.org	76,023	9,373	Netherlands
spanishtracker.com	62,680	5,554	Netherlands
xuntv.cn	54,013	11,002	China

TABLE IV
TOP 20 TRACKER ORGANIZATIONS RANKED BY THE NUMBER OF TRACKED PEERS

hosting more than 40% of the 150 highly-active trackers. After Sweden, the USA and Netherlands host the largest numbers of highly-active trackers.

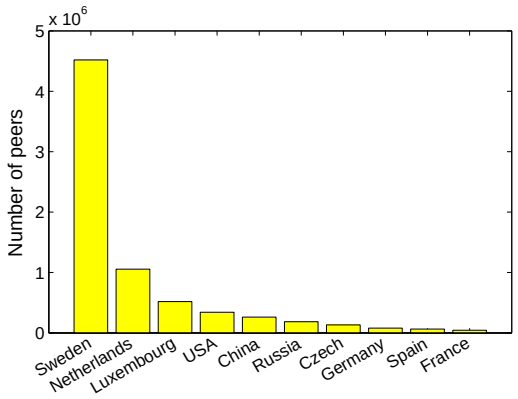

Fig. 10. Number of peers tracked from each country

Figure 10 shows the number of peers tracked by country (specifically, by the trackers residing in the country). We see that Sweden tracked *by far* the largest number of peers during our snapshot interval. Much of this tracker activity is due to Pirate Bay[14].

To analyze the scope of the study, we look at the number of active trackers brought by each of the web sites (beyond the union of the active trackers from the four other sites). Of the 728 active trackers, Mininova, Torrent Reactor and Torrent Portal only bring an additional 16 (2.4%), 10 (1.5%), and 8 (1.2%) active trackers. Among the highly-active trackers, tracking 1000 peers (of which there are 190), none of these sites individually brought any new trackers. We can therefore conclude that the scope includes the large majority of active trackers in the Ecosystem.

To complete our understanding of the tracker landscape, Figure 11 shows the CDF of the number of trackers and active trackers used by each active torrent. The curve for the active trackers is more revealing. We see that 71% of torrents are actually being tracked by two or more active trackers and that 16% of torrents are actually being tracked by 6 or more active trackers. Multiple active trackers improve availability [8].

C. Peer Statistics

Recall that the multi-tracker crawler obtained one snapshot of each active torrent, with all the snapshots for all torrents collected over a twelve-hour period on 22 April 2009. The total number of unique peers observed during the 12-hour period is 5,085,217. Figure 12 shows the CDF of the number of peers in active torrents. *Surprisingly, a large fraction of torrents are very small. For example, about 82% of the torrents have no more than 10 peers. Only about 1% of the torrents have more than 100 peers. Thus, contrary to popular belief, BitTorrent is rich in long-tail content.* Figure 12 also shows the CDF of the torrent size distributions for torrents in movie and music categories. (We discuss category classification in Section V.) It can be observed that music torrents tend to be smaller than movie torrents, with 80% of music torrents (as compared to 75% for movie torrents) having fewer than 10 peers.

Figure 13 shows the number of peers in a torrent, ordered from largest to smallest torrent. Although BitTorrent is dominated by mice (82% of torrents have not more than 10 peers), there are nevertheless many elephant torrents. There are 22 large torrents with more than 10,000 peers. Surprisingly, one torrent even has more than 150,000 peers, which is Heroes, one of the currently most popular TV series in the world. The torrent was uploaded to the Pirate Bay on 21 April 2009, one day before the tracker crawl. We see from Figure 13 that torrent size distribution approximately follows a Zipf distribution. *Thus, there is also tremendous diversity in the sizes of the BitTorrent torrents. BitTorrent torrent sizes exhibit*

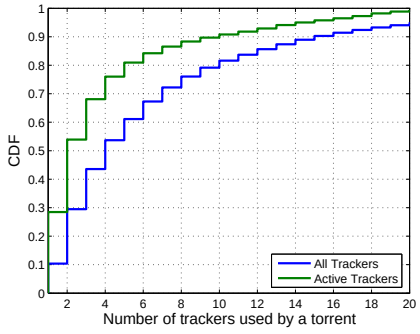

Fig. 11. CDF of the number of trackers used by each torrent.

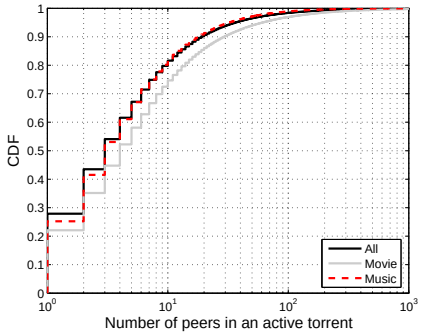

Fig. 12. CDF of the number of peers in active torrents.

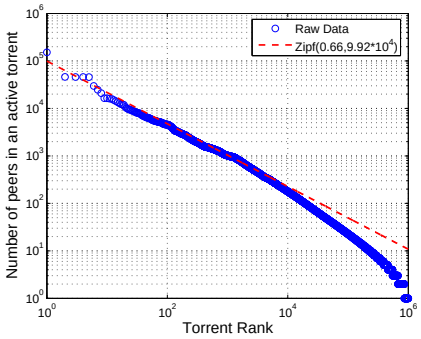

Fig. 13. Number of peers in active torrents

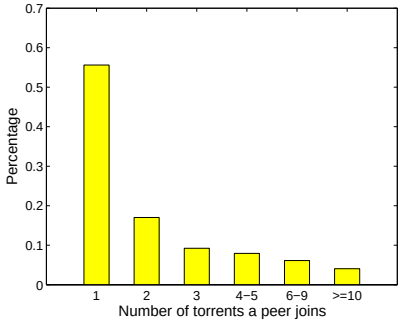

Fig. 14. The number of torrents a peer joins simultaneously

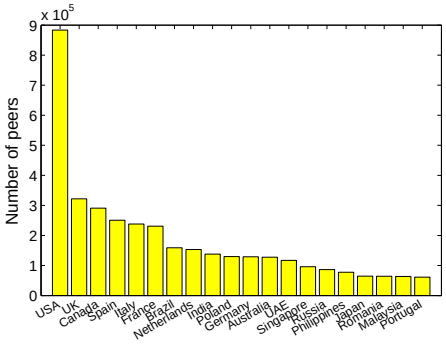

Fig. 15. Geographical distribution of all the peers

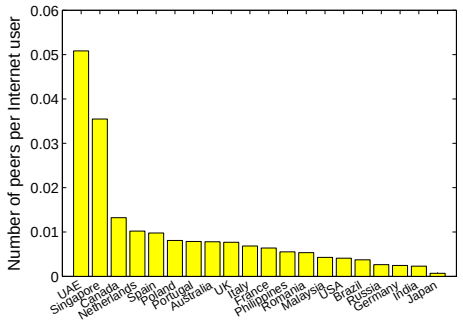

Fig. 16. Number of peers per Internet user in a country

classic Zipf behavior, both with significant long-tail content as well as the presence of very large torrents. After performing curve fitting, we observe that its Zipf exponent is around 0.66. The Zipf values (obtained from the power law parameters) reported for two categories in YouTube are 0.67 and 0.76 [15]. Assuming that those values are representative for most YouTube categories, we can conclude that BitTorrent is less biased toward popular content than YouTube.

Figure 14 shows the number of torrents a peer joins simultaneously (over the twelve-hour period). Observe that about 56% of the peers join only one torrent; about 26% of the peers join 2 or 3 torrents; about 15% of the peers join between 4 and 10 torrents; and about 4% join more than 10 torrents simultaneously. *There is significant diversity in user behavior. Although the majority of users download only one file in the 12-hour period, 44% of the users download multiple files.*

Figure 15 shows the geographical distribution of peers. Among the top-10 countries, eight countries are in Europe or North America. Neither China nor India, the two most populous countries in the world, are in the top 10. Our anecdotal investigations indicate, however, that there are indeed many BitTorrent users in China; but those users are largely participating in a disjoint torrent sets engendered by Chinese torrent-discovery sites.

To gain more insight into how popular BitTorrent is in various countries, Figure 16 shows the number of peers per Internet user in a country, with the data on Internet usage coming from [16]. Surprisingly, BitTorrent is most popular among Internet users in the UAE, Singapore and Canada, and generally popular throughout the European continent. Interestingly, the USA only ranks 15 in terms of BitTorrent usage per Internet user. We will investigate geographic behavior in more detail in Section V, where we consider content categories.

Also of interest is the popularity of different BitTorrent client types. To gain some insight here, we choose five random torrents of different sizes and content type. For each torrent, we determine all of the peers (IP,ports) in its trackers, and send handshake messages to each of those peers. From the responses, we determine the client type employed. (Not all peers respond, since some are no longer active or behind NATs). We then repeat this experiment three times for each torrent, and take averages. The results are shown in Figure 17. We see that all five torrents gave consistent results: over 50% of the peers use uTorrent in each of the torrents; Azureus is the second most popular client type, followed by Mainline and BitComet. *The Ecosystem again exhibits remarkable diversity, with four client types having significant penetration.* We also examined the distribution of client types that are creating .torrent files. Recall that the creators of .torrent files are often the initial seeds for the content files. Using the creation-client-type information in the .torrent field in the 4.6 million .torrent files, we plot the distribution of client types in Figure 18. We again see that uTorrent clients form the largest fraction of .torrent creators (and are quite likely the largest fraction of

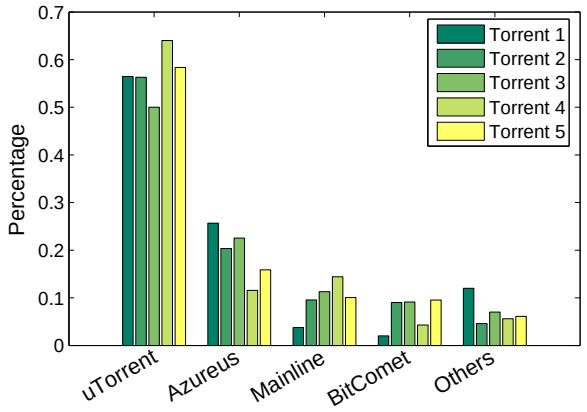

Fig. 17. Popularity of Different BitTorrent client types

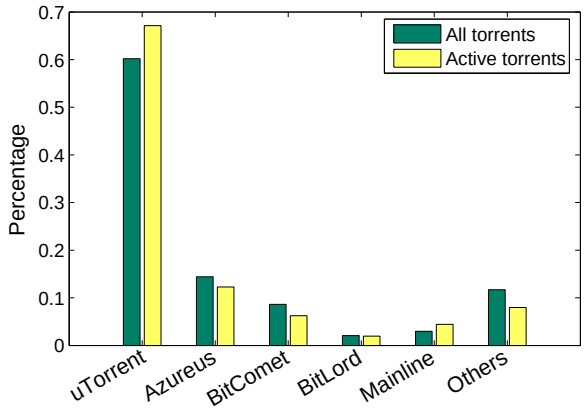

Fig. 18. Distribution of BitTorrent client types used to create .torrent files

initial seeders).

D. Torrent Popularity Versus Age

We now investigate the age distribution of active torrents. Similar to the results presented in [15] for You-Tube, we group torrents into different age groups (torrents born in the same seven-day period form a group) and record the distribution of swarm sizes in each age group.

Figure 19 shows the maximum, median, and average swarm sizes per group. It can be observed that young torrents on average have more peers compared to relatively older ones. There is, however, a clear distinction between torrents uploaded before and after 40 weeks ago: the average popularity of torrents younger than 40 weeks drops at a steep rate; and the average popularity of torrents older than 40 weeks remains nearly constant, in the 9-10 range for hundreds of weeks! For YouTube videos, as reported in [15], the nature of the curve for average number of requests is similar to this average popularity curve for BitTorrent swarms. However, the cutoff for YouTube videos is 1 month (4 weeks) as opposed to 40 weeks for BitTorrent. This means that for YouTube videos, the popularity is relatively insensitive to video ages (except for very recent ones) whereas BitTorrent file popularity is more sensitive to its age for torrents less than 10 months old. Another surprising observation is that there are torrents as old as 4 years that are still active.

Recall from the previous section, we define small active torrents as being ones having less than 10 peers, and highly active torrents with more than 100 peers. Figure 20 shows the age distribution of torrents of different activity levels: all active torrents, small active torrents, highly active torrents, and the top 100 torrents. For the top 100 torrents, about 40% of the torrents are uploaded within the past 10 days, and 95% are uploaded within the past 100 days. The top 100 torrents, the highly active torrents, and the small active torrents have a medium age of about 20 days, 100 days, and 300 days, respectively.

V. Content Analysis

In this section we provide an analysis of the content being distributed in the public (English-language) Ecosystem. To this end, we seek to classify active torrents into one of the following ten categories:

- *Movies:* DVD movies, high-resolution movies, documentaries, and so on (excluding porn).
- *Music:* Music related content, including music videos, sound tracks, songs, albums, music covers, concerts, and discographies.
- *TV/Radio Shows:* TV shows, radio shows, cartoons, anime.
- *Application:* Applications for Windows, MAC, Linux, and handheld devices; operating system installers (Windows, Linux, etc).
- *Games:* Games for PC, MAC, PS2, XBOX360, handheld, Wii, mobile phones, etc; game fixes and patches.
- *Books:* Audio books , e-Books, comics, articles, magazines, manuals, and so on.
- *Porn:* Anything related to porn, e.g., movies, movie clips, games, pictures, and so on.
- *Audio:* Content that could not be classified into any of the above 7 categories but are known to be audio files.
- *Video:* Content that could not be classified into any of the above 7 categories but are known to be video files; includes user generated content, video tutorials (e.g. for cooking).
- *Other:* Content that could not be classified into any of the above 9 categories.

Classifying the torrents is in itself a challenging problem, since the metadata available for many torrents is not always conclusive. We adopt simple heuristics for classification. Each of the five torrent discovery sites provides some kind of category information for each of the torrents it indexes. We use the category information provided by the sites to classify the torrents into one of the aforementioned categories. (For example, Pirate Bay categorizes each of its torrents into one of five major categories. In order to further narrow down the torrent into more specific category, each of its major categories have various sub-categories.) For torrents that are indexed by

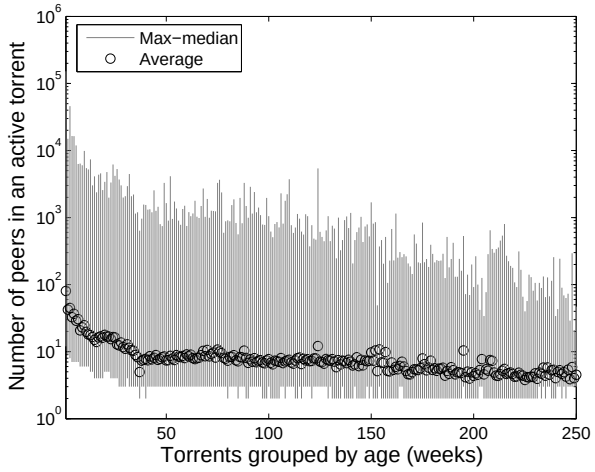

Fig. 19. Distribution of active torrent swam size over its age

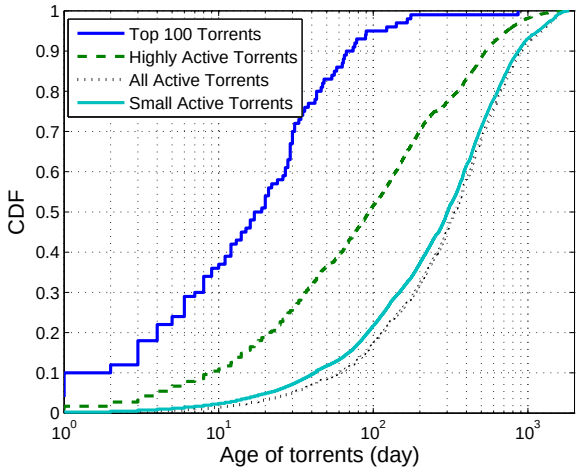

Fig. 20. Age distribution of torrents with different activity levels

more than one site, we adopt a voting system, wherein for each torrent, each site containing the torrent votes for one of the above 10 categories based on the category information it has for the torrent. The category that gets the highest number of votes for the torrent is considered the final category for the torrent. In case there is a tie in the number of votes obtained by two or more categories, we use the file extension for the torrent (available in the .torrent file) for its final classification. A torrent that cannot be classified into any of the first 9 categories (Movies - Video) using either of the above two heuristics is put into the "Other" category.

Figure 21 shows the overall classification of 1.2 million active torrents. It also shows the number of peers participating in each category. Movies, music, and TV shows are the leading category types, both in terms of the number of active torrents and in terms of the number of peers. Note that ratio of peers to torrents is larger for movies than music. This again shows that music torrents tend to be smaller than movie torrents. *Although movies, music and TV shows are the leading types, there is significant participation in books, games, applications, and porn, indicating a great diversity in the content being distributed by BitTorrent.*

To gain a deeper understanding of content popularity in BitTorrent, for each content category we examine its relative popularity in different countries. Figure 22 shows the number of peers per Internet user per country for three selected categories. We see that, at the time of crawling, UAE and Singapore led the way for the number of movies being downloaded per Internet user, and that movies are generally popular in Europe among Internet users. The popularity of movies in the UAE may be due to censorship in local theaters. The popularity in Europe may again be attributed to delayed releases. Interestingly, UAE and Singapore, although highly ranked for movies, are not in the top-10 for music. Unlike movies, music popularity does not dramatically change across countries. Porn is somewhat more popular in the UAE and Singapore.

VI. How important is Pirate Bay?

Pirate Bay plays an important role in the BitTorrent Ecosystem, not only due to its popular torrent-discovery site, but also due to its powerful tracker service. Our measurement results show that Pirate Bay has the most .torrent file uploaders, the highest ratio of active torrents to torrents, and hosts 14 of top 20 trackers (in terms of number of active torrents tracked.) Given Pirate Bay's importance, one interesting question is, if Pirate Bay is taken down by attacks or lawsuits, will the Ecosystem still be able to thrive as it does now?

To understand the importance of Pirate Bay, first consider the impact of removing the Pirate Bay torrent-discovery site. Pirate Bay indexes 637,698 (53.5%) of the 1,192,302 active torrents discovered. Among these active torrents indexed by Pirate Bay, we find that 594,654 of them are also indexed by at least one of other four sites. This means that, if Pirate Bay's torrent-discovery site were shut down, only 43,044 (3.6%) of the active torrents would no longer be indexed (at least by the 5 torrent-discovery sites considered in this paper) and therefore would be difficult to discover. As this percentage is quite small, we can conclude that the removal of Pirate Bay's torrent-discovery site would not greatly impact damage the Ecosystem.

The removal of Pirate Bay's trackers is potentially more serious, as they serve as major pillars in today's Ecosystem. Our measurements show that Pirate Bay was tracking 1,071,889 of the 1,192,302 (90%) active torrents of 22 April 2009. Many of the torrents tracked by Pirate Bay are also tracked by other trackers; in fact, only 279,627 (23%) of the active torrents are tracked exclusively by Pirate Bay. Thus one might conclude that only about a fourth of the of active torrents would be impacted by the removal the Pirate Bay's trackers. However, this analysis does not take into account the work load that would be shifted from Pirate Bay to the non-Pirate Bay trackers. The non-Pirate Bay trackers would likely have difficulty absorbing the new load.

The good news for the Ecosystem is that several of the most popular client types (including uTorrent, Azureus, Bitcomet)

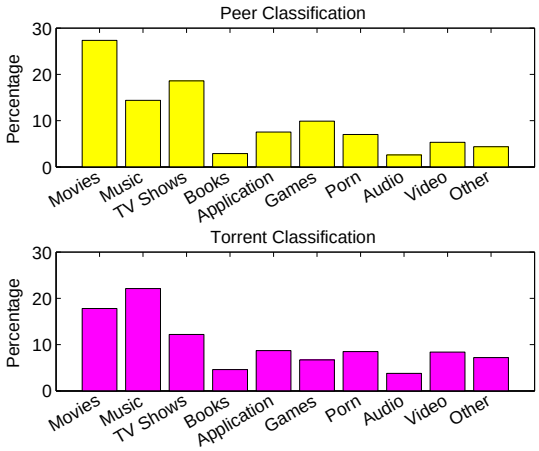

Fig. 21. Classification of peers and torrents

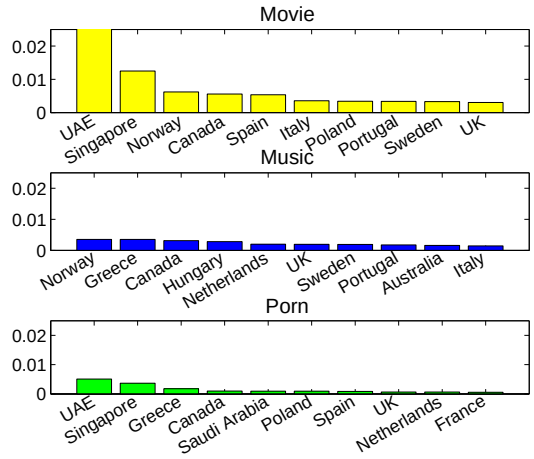

Fig. 22. Number of peers per Internet user in each category

ID	Size	AZ DHT peers	ML DHT peers	AZ/ML DHT peers
1	144	7.6%	59.7%	67.3%
2	1,231	8.7%	65.2%	73.9%
3	1,615	11.6%	52.6%	64.2%
4	3,548	4.7%	55.8%	60.5%
5	4,721	6.1%	62.9%	69.0%
6	10,656	5.3%	61.5%	66.8%

TABLE V
PERCENTAGES OF DHT PEERS AMONG THE TRACKED PEERS

support DHT and PEX, providing distributed tracker services. To test the fraction of DHT peers in the torrents, we developed a DHT crawler for the Azureus DHT and another for the Mainline DHT (employed by uTorrent and Mainline). We randomly selected six torrents of different sizes, obtained the peer sets from the trackers, and obtained the peer sets from the DHTs using our crawlers. By comparing the sets obtained from the trackers and the DHTs, we identify the percentage of Azureus (AZ) DHT peers and Mainline (ML) DHT peers among all the peers in the torrent. Table V shows the results for the six representative torrents. We observe that about 60% of peers in each of the torrents participate in one of two DHTs. Thus, even if Pirate Bay trackers are completely taken down, peers will still be able to discover many peers via the DHTs. We should note, however, that the Azureus and Mainline DHTs are two incompatible and almost completely disjoint. Many client types also support PEX, allowing them to gossip peer lists among themselves. But the client types implement incompatible versions of PEX protocol, which limits the effectiveness of peer discovery. Another problem of PEX is that a peer of a particular client type needs to first discover another peer of the same client type in the same torrent. Peers are using DHT and PEX simultaneously to improve peer discovery.

We therefore conjecture that even if Pirate Bay is taken down, large torrents will still thrive in the Ecosystem, since users will still be able to find sufficient numbers of peers using DHTs, PEX, and the non-Pirate Bay trackers. Many small torrents may suffer, however, due to splitting among the DHTs and the possibility of no remaining static trackers. Users that use non-DHT clients would have more impetus to switch to uTorrent and Azureus. Hopefully the developers of BitTorrent clients will soon employ common protocols for DHT and PEX, to avoid splitting of torrents, particularly for small torrents.

VII. RELATED WORK

In recent years, significant progress has been made in understanding the strengths and limitations of BitTorrent's protocol and its tit-for-tat mechanism [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]. However, there is little work that addresses the Ecosystem at large. Pouwelse et al. study BitTorrent usage, using statistics gathered from a single torrent-search site (the now defunct Suprnova site); their six-page workshop paper only superficially describes the measurement and analysis methodologies [6]. Bellissimo et al. collected 3-month logs from only two trackers and investigated a limited number of torrent and peer characteristics [29]. Guo et al. [7], [30] measured torrent evolution, service availability, and client performance by analyzing a limited number of tracker traces from [29] and torrent file download traces. Neglia et al. [8] investigated the availability of BitTorrent systems; they collected about 22,000 torrents from two torrent-discovery sites and mainly focused on tracker/DHT reliability issues. Izal et al. analyze the behavior of a single torrent over a five-month period [19].

The current paper goes significantly farther. The data set is more than an order of magnitude larger (in terms of torrents, trackers, and peers) than each of these earlier studies. Moreover, by addressing the torrent-discovery, peer-discovery, peer and content characteristics all at the same time, we provide a much richer picture of the BitTorrent ecosystem.

VIII. CONCLUSION

Although BitTorrent plays an important role in the Internet, there has lacked an up-to-date and comprehensive understanding of the BitTorrent Ecosystem. In this paper we provide

a nearly complete picture of the public English-language BitTorrent Ecosystem. Our large-scale measurement covers five of the most popular torrent-discovery sites. In a nine-month period, we identified over 4.6 million unique torrents and 38,996 trackers that the above five sites reference. We investigated the degree of indexing overlap among the sites, the characteristics of uploaders and how the sites acquire .torrent files. To gain further insight into the world of trackers and peers, we developed a high-performance multi-tracker crawler that simultaneously crawls thousands of trackers with concurrent TCP connections. It obtained peer lists of millions of torrents within a narrow window of twelve-hours.

We found that the Ecosystem exhibits remarkable diversity in terms of the operation of the major torrent-discovery sites, user upload behavior, numbers of torrents and peers tracked by trackers, content type, and client implementations. Nevertheless, we found that Pirate Bay currently plays a disproportionate role in the Ecosystem. We performed a detailed analysis of the Ecosystem's dependance on Pirate Bay, including an analysis of the extent the Azureus and Mainline DHTs can support the Ecosystem. We also found that the popularity of BitTorrent content is sensitive to its age.

The BitTorrent Ecosystem is by many measures the most successful open Internet application deployed in this decade. The results of this paper should benefit diverse communities including P2P researchers, ISP researchers, copyright holders, among others. The collected data has been anonymized and made publicly available to the research community.

REFERENCES

[1] "Mininova's Torrent Downloads Double to 7 Billion in a Year," http://torrentfreak.com/mininovas-torrent-downloads-doubled-in-a-year-090105.

[2] "Bittorrent clients," http://en.wikipedia.org/wiki/BitTorrent_client.

[3] "Bittorrent trackers," http://en.wikipedia.org/wiki/BitTorrent_tracker.

[4] X. Hei, C. Liang, J. Liang, Y. Liu, and K. W. Ross, "A measurement study of a large-scale P2P IPTV system," in *IEEE Transactions on Multimedia*, Dec 2007.

[5] Y. Huang, T. Z. Fu, D. Chiu, J. C. Lui, and C. Huang, "Challenges, Design and Analysis of a Large-scale P2P VoD Systems," in *ACM SIGCOMM*, 2008.

[6] J. A. Pouwelse, P. Garbacki, D. H. Epema, and H. Sips, "The BitTorrent P2P file-sharing system: Measurements and analysis," in *Proc. of IPTPS*, Ithaca, NY, Feb. 2005.

[7] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang, "Measurements, analysis, and modeling of bittorrent-like systems," in *Internet Measurement Conference*, California, USA, Oct. 2005.

[8] G. Neglia, G. Reina, H. Zhang, D. Towsley, A. Venkataramani, and J. Danaher, "Availability in BitTorrent Systems," in *IEEE INFOCOM*, 2007.

[9] "Alexa Web Site," http://www.alexa.com.

[10] G. I. Ivchenko, "How many samples does it take to see all the balls in an urn?" in *Mathematical Notes, Vol. 64, No. 1, pp. 58-63*, Jul. 1998.

[11] "Slyck Forum."

[12] "MovieX Site," http://www.moviex.info.

[13] "MaxMind GeoIP," http://www.maxmind.com/app/geoip_country.

[14] "The Pirate Bay," http://thepiratebay.org.

[15] M. Cha, H. Kwak, P. Rodriguez, Y.-Y. Ahn, and S. Moon, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System," in *ACM SIGCOMM/USENIX IMC*, 2007.

[16] "World Internet Users and Population Stats," http://www.internetworldstats.com/stats.htm.

[17] B. Cohen, "Incentives build robustness in bittorrent," in *Proc. P2PEcon*, CA, Jun. 2003.

[18] B. Fan, D.-M. Chiu, and J. C. S. Lui, "The Delicate Tradeoffs in BitTorrent-like File Sharing Protocol Design," in *ICNP*, 2006.

[19] M. Izal, G. Urvoy-Keller, E. Biersack, P. Felber, A. Hamra, and L. Garces-Erice., "Dissecting BitTorrent: five months in a torrent's lifetime," in *Passive and Active Measurements*, Apr. 2004.

[20] T. Karagiannis, A. Broido, N. Brownlee, K. C. Claffy, and M. Faloutsos, "Is P2P dying or just hiding?" in *IEEE Globecom*, 2004.

[21] A. Legout, G. Urvoy-Keller, and P. Michiardi, "Rarest First and Choke Algorithms Are Enough," in *IMC*, 2006.

[22] N. Liogkas, R. Nelson, E. Kohler, and L. Zhang, "Exploring the robustness of BitTorrent peer-to-peer content distribution systems," *Concurrency and Computation: Practice and Experience*, 2007.

[23] T. Locher, P. Moor, S. Schmid, and R. Wattenhofer1, "Free riding in BitTorrent is cheap," in *Proc. ACM HotNets*, 2006.

[24] S. Jun and M. Ahamad, "Incentives in Bittorrent induce free riding," in *ACM SIGCOMM Workshop P2PEcon*, PA, Aug. 2005.

[25] M. Piatek, T. Isdal, T. Anderson, A. Krishnamurthy, and A. Venkataramani, "Do incentives build robustness in bittorrent?" in *Proc. NDSI*, Cambridge, MA, Apr. 2007.

[26] M. Sirivianos, J. H. Park, R. Chen, and X. Yang, "Free-riding in bittorrent networks with the large view exploit," in *Proc. IPTPS*, Bellevue, WA, Feb. 2007.

[27] A. Legout, N. Liogkas, E. Kohler, and L. Zhang, "Clustering and sharing incentives in bittorrent systems," in *Proc. ACM SIGMETRICS*, San Diego, CA, Jun. 2007.

[28] N. Laoutari, D. Carra, and P. Michiardi, "Uplink allocation beyond choke/unchoke or how to divide and conquer best," in *Proc. of CoNEXT*, Spain, Dec 2008.

[29] A. Bellissimo, B. N. Levine, and P. Shenoy, "Exploring the use of bittorrent as the basis for a large trace repository." in *Tech. Rep. 04-41, UMASS Amherst*, Jun. 2004.

[30] L. Guo, S. Chen, Z. Xiao, E. Tan, X. Ding, and X. Zhang, "A performance study of bittorrent-like peer-to-peer systems," in *IEEE JSAC*, 2008.

Chao Zhang received the B.S. degree in Computer Science and Engineering from University of Electronic Science Technology of China in 2003, and M.S. degree in Computer Science and Engineering from Huazhong University of Science and Technology in 2006. He has been a PhD student in the Department of Computer Science & Engineering in Polytechnic Institute of New York University since 2007. His research interests include distributed and software systems, peer-to-peer networking, and so on.

Prithula Dhungel received her Bachelor's degree in Computer Engineering from Tribhuvan University, Nepal. She has a Master's degree in Computer Science from Asian Institute of Technology, Thailand. Currently, she is a PhD candidate at Polytechnic Institute of New York University working under the supervision of Professor Keith Ross.

Her research interests include Peer to Peer networks, Network measurement and analysis, and Network Security.

Di Wu (M' 06) received the B.S. degree from the University of Science and Technology of China in 2000, M.S. degree from the Institute of Computing Technology, Chinese Academy of Sciences in 2003, and Ph.D. degree in Computer Science & Engineering from the Chinese University of Hong Kong in 2007. From 2007 to 2009, he was a postdoctoral researcher in the Department of Computer Science & Engineering, Polytechnic Institute of NYU, advised by Prof. Keith W. Ross. He joined the Department of Computer Science, Sun Yat-Sen University, China as an Associate Professor in July 2009. He was the winner of IEEE INFOCOM 2009 Best Paper Award, and a member of IEEE, ACM and Sigma Xi.

His research interests include Peer-to-Peer Networking, Multimedia Communication, Internet Measurement, Mobile Computing, and Network Security.

Keith W. Ross (F' 08) joined Polytechnic University as the Leonard J. Shustek Chair Professor in Computer Science in January 2003. He has been Department Head since September 2008. Before joining Polytechnic, he was a professor for five years at Eurecom Institute in Sophia Antipolis, France. From 1985 through 1997, he was a professor at the University of Pennsylvania. He received a B.S.E.E from Tufts University, a M.S.E.E. from Columbia University, and a Ph.D. in Computer and Control Engineering from The University of Michigan.

Professor Ross has worked in peer-to-peer networking, Internet measurement, video streaming, multi-service loss networks, network security, queueing theory, and Markov decision processes. He is an IEEE Fellow, recipient of the Infocom 2009 Best Paper Award, and recipient of Best Paper in Multimedia Communications 2006-2007 (awarded by IEEE Communications Society). He has served on numerous journal editorial boards and conference program committees. He was PC co-chair for ACM Multimedia 2002, ACM CoNext 2008, and IPTPS 2009. He has served as an advisor to the Federal Trade Commission on P2P file sharing.

Professor Ross is co-author (with James F. Kurose) of the popular textbook, Computer Networking: A Top-Down Approach Featuring the Internet, published by Addison-Wesley (first edition in 2000, fifth edition 2009). It is the most popular textbook on computer networks, both nationally and internationally, and has been translated into twelve languages. Professor Ross is also the author of the research monograph, Multiservice Loss Models for Broadband Communication Networks, published by Springer in 1995.

From July 1999 to July 2001, Professor Ross took a leave of absence to found and lead Wimba, an Internet technology start-up. Wimba develops and markets Java-based asynchronous and synchronous voice-over-IP technologies, primarily for the on-line education and language learning markets. Wimba is now headquartered in NYC and has more than 80 employees worldwide.