An Empirical Study of Seeders in BitTorrent

Justin Bieber, Michael Kenney, Nick Torre, and Landon P. Cox
Duke University, Durham, NC

Abstract

BitTorrent has attracted attention from researchers and the press for its wide deployment and explicit goal of eliminating free-riding, but many of the most important peers in BitTorrent operate outside of its incentive mechanisms. Altruistic seeders help bootstrap new peers and provide a significant fraction of the global upload bandwidth. We have taken an empirical approach to understanding seeders by studying 35 BitTorrent sites with nearly four million users at any moment over several weeks.

Our study focuses on two aspects of seeders. First, we looked at the relationship between the number of seeders and bandwidth utilization. A case study of a Linux distribution network showed that as bandwidth utilization increased, the rate of seeding decreased. Second, we looked at the relationship between site attributes and the number of seeders. A survey of 34 BitTorrent sites over two weeks found that the presence of niche-content (e.g. only anime, hip-hop, or Linux files), merchandise for sale (e.g. t-shirts with the site URL), and negative reinforcement (e.g. a posted list of the 10 least contributing peers) correlated positively with the rate of seeding.

1 Introduction

Free-riding in file-sharing networks such as Gnutella, Napster, and Kazaa is a well documented phenomenon [1, 6]. Because of this, the BitTorrent [3] content distribution system made robustness to free-riding an explicit design goal [4]. BitTorrent's attention to incentives as well as its responsibility for 35% of Internet traffic [11] has generated interest from both the research community [5] and popular press [12]. However, despite this attention, there have been no broad empirical studies of BitTorrent.

Several papers have analyzed the incentive mechanisms proposed by BitTorrent [7, 9], but we are aware of only one empirical study [8], by Izal. This is a useful start, but because it focuses on a single shared file the scope of its observations are limited. We hope to broaden the understanding of BitTorrent by presenting results from several weeks of logging 35 BitTorrent sites during February and March of 2006. These sites serve a wide range of content to nearly 4 million users at any moment.

As was the case in Izal's study, we collected data by passively monitoring BitTorrent use through the *trackers* that coordinate peers. This prevented us from directly observing how users respond to BitTorrent's incentive mechanisms, but provided us with the number of *seeders* (peers who are uploading, but not downloading) and *leechers* (peers who are downloading).

Seeders are particularly important for bootstrapping new peers and to network performance. Many sites explicitly encourage peers to continue uploading data even after their download has finished. These reminders are necessary because seeding is altruistic rather than self-interested; BitTorrent's incentive mechanisms cannot reward seeders for the upload bandwidth they contribute. This makes seeders difficult to model game theoretically and motivated our empirical study.

We have focused on two aspects of seeders. First, we examined the relationship between bandwidth utilization and the number of seeders. A case study of a BitTorrent site over 23 days showed that the ratio of seeders-to-leechers decreased dramatically when bandwidth became highly utilized. Second, we looked at the relationship between site attributes and altruism. We were interested to know which, if any, site characteristics correlate with a high seeders-to-leechers ratio. In a study of 34 different sites, we found that networks with niche-content (e.g. only anime, hip-hop, or Linux files), merchandise for sale (e.g. t-shirts with the site URL), and negative reinforcement (e.g. a posted list of the 10 least contributing peers) exhibited greater altruism than those without these features.

2 Background and Motivation

BitTorrent [3] is a popular peer-to-peer file sharing application that uses a "tit-for-tat" protocol [2] to transfer files between peers. BitTorrent files are broken into smaller fragments, which can be downloaded in parallel from multiple peers. Although not a "pure" tit-for-tat protocol [9], BitTorrent is designed to reward peers for uploading fragments to others with proportionally fast downloads.

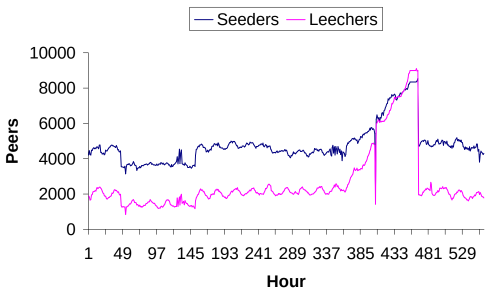

Figure 1: Hourly Behavior at elm-project.org

To download a file, peers must first have an associated `.torrent` file. `.torrent` files are usually obtained through BitTorrent web sites and contain important information about the download target, such as its length, its name, fragments' hashing information and the URL of a *tracker*. The tracker is a server that maintains information about where fragments can be downloaded from. A group of peers actively downloading and uploading fragments is a *swarm*.

Using the `.torrent` file, clients can query the tracker about the locations of their missing fragments within the swarm. Clients request fragments directly from other peers and can increase the priority of their requests by offering to upload the fragments still needed by those peers. This creates an incentive for clients to download the rarest fragments first and increases overall fragment availability.

Of course, when a client first enters the swarm, it has nothing to offer anyone else. Thus, clients must be able to download fragments "for free" from *seeders*. Seeders already have complete files and expect nothing in return for their uploading.

Seeders are also critical for BitTorrent performance. Izal's study found that over five months, seeders accounted for over two-thirds of the total upload bandwidth [8]. Their importance is reflected in the pejorative label, *leechers*, assigned to peers who are still downloading files. Peers can only reach the status of seeder by uploading *without* downloading. Importantly, leechers are not necessarily free-riders. Even peers that engage in tit-for-tat with the rest of the swarm are considered leechers until they have obtained all file fragments.

The role seeders play in bootstrapping BitTorrent and network performance makes understanding them an important data point for content-distribution network designers and administrators. Game theory has proven useful for understanding leechers' behavior [4, 9], but offers little hope of understanding seeders since altruism is not easily described by these models. Because of this, we hope that our empirical observations will provide insight into how seeders behave.

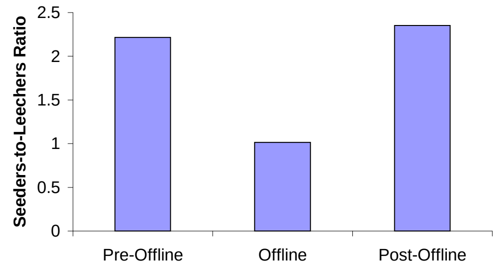

Figure 2: The Effect of LMP Server Downtime

3 Study Results

We periodically logged the number of seeders and leechers within the swarms of 35 trackers over several weeks in February and March of 2006. Each entry was an aggregation across all swarms connected to each tracker. We obtained this information by cooperating with tracker administrators as well as monitoring the statistics posted on BitTorrent web sites. Depending on the tracker, we were able to log statistics on a daily or hourly basis. The first half of our study examines the relationship among available bandwidth, download rates, and the composition of the swarm in a Linux distribution network over three weeks.

All of the BitTorrent trackers we monitored had an associated web site; tracker URLs within the `.torrent` files hosted by a web site referred to servers controlled by web site administrators. This close relationship allowed us to associate web sites' attributes with the swarm statistics we logged. Some of the attributes we recorded include the nature of files' content, whether sites provided message boards, or if registration was required to access `.torrent` files. The second half of our study looks at the relationship between these attributes and trackers' seeders-to-leechers ratios.

3.1 Case Study: The Linux Mirror Project

One of the most interesting sites in our study was the Linux Mirror Project [10] (LMP). As its name suggests, LMP uses BitTorrent to distribute Linux distributions, kernels, and other open-source software. We recorded the number of seeders and leechers tracked by LMP every hour for 23 days. These numbers are in Figure 1.

Under normal circumstances, LMP peers appear to be extremely generous, exhibiting a seeders-to-leechers ratio of over two in steady state. By comparison, Izal observed a peak ratio of approximately .6 [8]. However, between the 17th and 20th day of the trace, the swarm size crashed before nearly tripling normal levels. More interesting, during this period the ratio of seeders-to-leechers approached one and even dipped below one

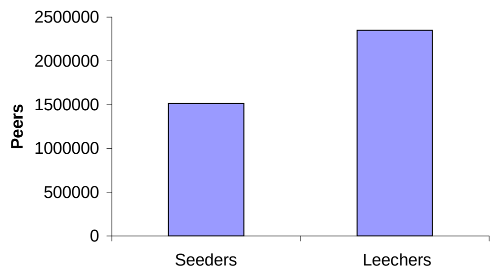

Figure 3: Aggregate Seeders and Leechers

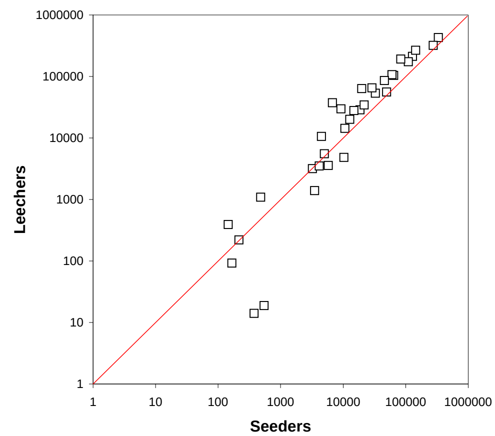

Figure 4: Average Daily Seeders and Leechers

toward the end of this growth. Around the 20th day, the network suddenly returned to normal.

Understanding this behavior requires an explanation of LMP's architecture. LMP maintains three separate servers: one that acts as a web site and tracker and two that act as seeders. Early in the morning of the 17th day, the web site and tracker went offline because of a problem with LMP's ISP. Once administrators resolved the problem, they brought the web site and tracker back online, but the two seeder servers remained offline due to a misconfiguration. It was not until mid-way through the 19th day that administrators were able to fix the misconfiguration and bring all three servers online. As the data shows, once the seeding servers came back, the number of seeders and leechers returned to normal levels.

It appears that in steady state LMP users enter and leave the system at a constant rate. During the days when the LMP seeder servers were down, the system lost significant bandwidth capacity which led to longer downloads. Peers likely continued to enter at the same rate, but were forced to stay longer. This explains why the swarm grew, but does not explain why the ratio of seeders-to-leechers declined. As Figure 2 shows, the system averaged a ratio of 2.2 before the seeder servers went offline, 1.0 while they were offline, and 2.4 when the servers came back online.

We believe that the decline was due to the different classes of seeders. *Persistent* seeders leave their clients on at all times. *Transient* seeders leave their clients running unattended while downloading and periodically check to see if the download has completed. Once the user sees that their download has completed, they close their client.

The decline in seeders-to-leechers ratio while LMP's servers were offline likely reflect a shift in the portion of the swarm composed of persistent seeders. The two-to-one ratio under normal conditions suggests that persistent seeders comprised the bulk of the swarm. However, when downloads took longer and the swarm grew, persistent seeders became outnumbered by transient seeders.

One possible explanation for this shift is that the cost of altruism increased as load shifted from LMP's servers onto the swarm. Persistent seeders may have reacted to this increased cost by exiting. However, this seems unlikely given how quickly the swarm re-established its steady state levels of seeders and leechers once LMP's servers came back online.

A more likely explanation is that transient seeders' times between completion checks are inversely proportional to the download rate. By definition, the number of persistent seeders is constant. The lower download rate makes the transition from leecher to transient seeder slower. If the rate of transition from transient seeder to out of the swarm had remained constant, the number of seeders would not have grown. Thus, the transient seeders' exit rate must have slowed.

This explanation is consistent with discussions we had with LMP administrators. When asked about the decline in seeders-to-leechers ratio they guessed that because of the longer downloads, users had checked their client's status less often. Less frequent checks would lead to a longer period between completing a download and closing a client. The administrators also mentioned that this had been their own experience when using BitTorrent. A more formal treatment of these interactions is beyond the scope of this paper, but is a target for future work.

3.2 Multi-Site Survey

The second part of our study looked at the relationship between the number of seeders and leechers of 34 trackers and the attributes of their associated web site. For each tracker, we recorded the number of seeders and leechers once per day for two weeks. Figure 3 shows the average number of seeders and leechers per data point, summed across all trackers. At any moment during our two weeks of logging, there were an average of 1.5 million seeders and 2.3 million leechers, giving a ratio of .65.

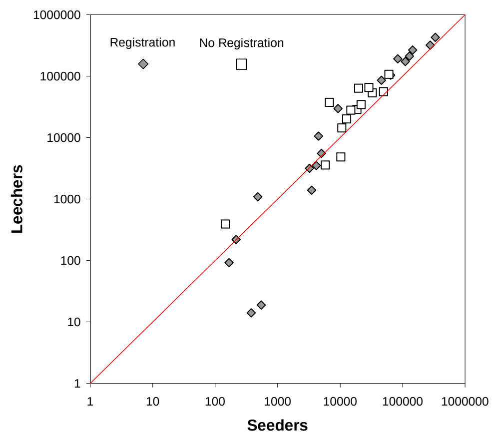

Figure 5: Registration vs. No Registration

However, the sizes and ratios of each network varied significantly. Figure 4 shows the average number of seeders and leechers at any moment for all 34 networks. Each point represents a different network and the line represents a seeders-to-leechers ratio of one. Most sites tracked fewer seeders than leechers, although smaller networks exhibited larger ratios.

To gain insight into the differences between networks, we recorded the attributes of each web site and computed the number of associated seeders and leechers with and without each attribute. We first looked at the effect of having to register to use a web site. Figure 5 shows the average number of seeders and leechers for networks with and without registration.

The effect of registering had only a minor effect on seeders-to-leechers ratio. Users of sites with registration exhibited a ratio of .67, while users of sites with out registration exhibited a ratio of .57. However, it is interesting to note that sites with registration tend to have the smallest and largest networks. We believe that both large and small sites may be using this extra hurdle to discourage network growth. Small sites may be trying to remain exclusive, while large sites may be trying to keep from growing beyond their servers' capacity to track swarms and serve web pages.

Figure 6 shows the seeders-to-leechers ratios of sites with and without five other attributes. The first attribute, "Reminder," describes whether a web site had a prominent reminder for users to leave their client open. Interestingly, sites without the reminder had higher seeders-to-leechers ratios than those with the reminder. This led us to believe that sites posted a reminder after realizing how poor their ratio was.

The second attribute we looked at was whether a network distributed niche or general content. For example, we monitored trackers that specialized in hip-hop music, anime, open source software, independent movies,

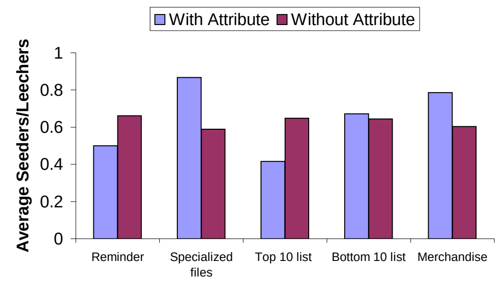

Figure 6: Site Attributes

computer games, and horror movies. Other trackers had no focus and hosted a wide range of content. The ratio of niche content trackers was much higher (.87) than for general content trackers (.59). In fact, all but one of the trackers with positive ratios hosted niche content.

There are at least two complementary explanations for this result. As Figure 7 shows, most networks with specialized content are smaller and smaller networks tend to have higher ratios. This is hardly surprising since one would expect networks with a narrow focus to be smaller. Users of specialized networks may also have a strong interest in maintaining the distribution network. For example, many peers in the independent movie network seed their own movies.

Next, we looked at web sites that posted the identities of their top 10 and bottom 10 seeders. The data shows that the negative reinforcement of posting the bottom 10 produced a much higher seeders-to-leechers ratio than the positive reinforcement of posting the top 10 seeders. In fact, sites that posted a top 10 list actually had a much lower ratio (.42) than sites without a top 10 list (.65). Conversely, sites with a bottom 10 list had a slightly higher ratio (.67) than those without a top 10 list (.65).

Figure 8 takes a closer look at this result. In total, there were two sites with only top 10 lists and one site with both a top 10 and bottom 10 list. Between the two sites with both lists, there was an average of 8,041 seeders and 11,965 leechers in their networks at any moment. The site with only a top 10 list averaged 9,231 seeders and 29,590 leechers in its network at any moment. This corresponds to a 15% increase in seeders and a 268% increase in leechers.

Finally, another attribute that correlated strongly with high seeders-to-leechers ratios was whether the site sold site-specific merchandise. These sites sold items such as T-shirts, baseball caps, and coffee mugs with the site URL and slogans. Sites with merchandise exhibited a ratio of .79 while those without merchandise exhibited a ratio of .60. Interestingly, sites with merchandise tended to be large, with an average network size of over 230,000 peers.

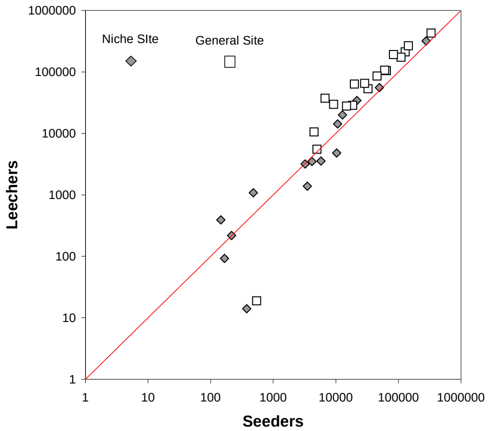

Figure 7: Niche vs. General-content Sites

There are at least two possible explanations for this. One is that sites with merchandise provoke strong comraderie among peers which encourages them to behave altruistically. Another explanation is that the money earned by sites from merchandise may be used to pay for extra bandwidth. This extra bandwidth could induce the effect on persistent and transient seeders we observed with LMP.

There are other attributes we would like to look at, but have not been able to including the distribution of file sizes, the average download rate, which clients peers use, and peers' geography. Gaining access to this information will require a closer relationship with the sites we have studied, but we are hopeful that we will be able to establish greater cooperation in the future.

4 Conclusions

BitTorrent has attracted attention for its wide deployment and explicit goal of eliminating free-riding. Many of the most important peers in BitTorrent operate outside of its incentive mechanisms. Altruistic seeders help bootstrap new peers and provide a significant fraction of the global upload bandwidth. We know of no broad empirical studies of seeders and their altruistic nature makes seeders difficult to model game theoretically. Because of this, we have studied 35 BitTorrent sites with nearly four million users at any moment over several weeks.

Our study focuses on two aspects of seeders. First, we looked at the relationship between the number of seeders and download rates. A case study of the Linux Mirror Project distribution network showed that as download rate increased, the rate of seeding decreased. We believe that this is due to a relative increase in the number of transient seeders in the swarm caused by users leaving their clients unattended for longer periods.

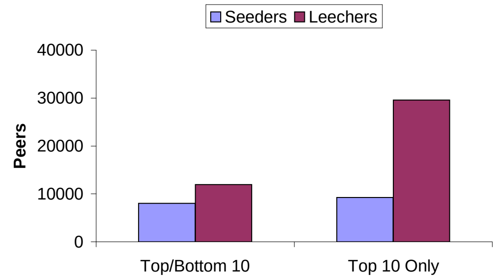

Figure 8: Posting Top and Bottom 10 Contributers

Second, we looked at the relationship between site attributes and the number of seeders. A survey of 34 BitTorrent sites over two weeks found that the presence of niche-content (e.g. only anime, hip-hop, or Linux files), merchandise for sale (e.g. t-shirts with the site URL), and negative reinforcement (e.g. a posted list of the 10 least contributing peers) correlated positively with the number of seeders. Greater altruism among niche content sites may be explained by a stronger self-interest in the distribution of content and their relatively small size. Selling merchandise may allow sites to pay for extra seeding bandwidth, which decreases utilization in the swarm.

5 Acknowledgments

We would like to thank the members of all of the sites studied in this paper. In particular, the Linux Mirror Project administrators were extremely gracious in helping us understand what happened when their seeder servers went offline.

References

[1] E. Adar and B. A. Huberman. Free riding on Gnutella. *First Monday*, 5(10), October 2000.

[2] R. Axelrod. *The Evolution of Cooperation*. Basic Books, 1984.

[3] BitTorrent. http://bittorrent.com.

[4] B. Cohen. Incentives build robustness in BitTorrent. In *Proceedings of the First Workshop on the Economics of Peer-to-peer Systems*, Berkeley, CA, June 2003.

[5] C. Gkantsidis and P. Rodriguez. Network coding for large scale content distribution. In *Proceedings of the 24th IEEE Infocom*, Miami, FL, March 2005.

[6] K. P. Gummadi, R. J. Dunn, S. Saroiu, S. D. Gribble, H. M. Levy, and J. Zahorjan. Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In *Proceedings of the 19th Symposium on Operating Systems Principles*, Bolton Landing, NY, October 2003.

[7] D. Hales and S. Patarin. How to cheat BitTorrent and why nobody does. Technical Report UBLCS-2005-12, University of Bologna, May 2005.

[8] 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 *Proceedings of the 5th Passive and Active Measurement Workshop*, Antibes Juan-les-Pins, France, April 2004.

[9] S. Jun and M. Ahamad. Incentives in BitTorrent induce free riding. In *Proceedings of the Third Workshop on the Economics of Peer-to-peer Systems*, Philadelphia, PA, August 2005.

[10] The Linux Mirror Project.
http://tlm-project.org.

[11] CacheLogic Research. The true picture of peer-to-peer filesharing, 2005.
http://cachelogic.com/research.

[12] Wired 13.01: The BitTorrent Effect, January 2005. WIRED Magazine.