Sampling Bias in BitTorrent Measurements

Boxun Zhang, Alexandru Iosup, Johan Pouwelse, Dick Epema, and Henk Sips

Parallel and Distributed Systems Group
Delft University of Technology, Delft, the Netherlands
{B.Zhang,A.Iosup,J.Pouwelse,D.H.J.Epema,H.J.Sips}@tudelft.nl

Abstract. Real-world measurements play an important role in understanding the characteristics and in improving the operation of BitTorrent, which is currently a popular Internet application. Much like measuring the Internet, the complexity and scale of the BitTorrent network make a single, complete measurement impractical. While a large number of measurements have already employed diverse sampling techniques to study parts of BitTorrent network, until now there exists no investigation of their sampling bias, that is, of their ability to objectively represent the characteristics of BitTorrent. In this work we present the first study of the sampling bias in BitTorrent measurements. We first introduce a novel taxonomy of sources of sampling bias in BitTorrent measurements. We then investigate the sampling among fifteen long-term BitTorrent measurements completed between 2004 and 2009, and find that different data sources and measurement techniques can lead to significantly different measurement results. Last, we formulate three recommendations to improve the design of future BitTorrent measurements, and estimate the cost of using these recommendations in practice.

1 Introduction

Peer-to-Peer file-sharing networks such as BitTorrent serve tens of millions of users daily and are responsible for a significant percentage of the total Internet traffic. Much effort [1, 2, 3, 4, 5, 6] has been put in the last decade into empirical measurements of P2P file-sharing systems including BitTorrent, with the purpose of understanding and improving their use. Similarly to early Internet measurement efforts [7, 8], due to the size of the complete network all BitTorrent measurements have employed data sampling techniques, from periodic measurements to the focus on specific BitTorrent communities. Despite this situation, there currently exists no comprehensive evaluation of the sampling biases, that is, of the ability to objectively represent the characteristics of BitTorrent, introduced by BitTorrent measurements. This work presents the first such investigation.

Understanding sampling biases in BitTorrent measurements can benefit the research on BitTorrent in the following ways. First, it can lead to a better understanding of the commonalities and of differences among different parts of the BitTorrent network by explicitly comparing the measurement results. In the Internet community, this "search for invariants" process [7] fostered many new research opportunities [8]. From the large number of empirical BitTorrent

P. D'Ambra, M. Guarracino, and D. Talia (Eds.): Euro-Par 2010, Part I, LNCS 6271, pp. 484–496, 2010.

measurements [2, 3, 9, 10], few [9, 10] consider even aspects of the sampling bias problem. Second, understanding sampling biases leads to better understanding of the usage of measurement techniques, which is key to designing and improving BitTorernt measurements. It is symptomatic for the current (lack of) understanding of BitTorrent measurement techniques that there is no agreement on the Internet traffic share due to BitTorrent–though caching companies have put forth estimates of over 50% in 2008 [11] and 30% in 2005 [12]. Towards understanding sampling biases in BitTorrent measurements, our main contribution is threefold:

1. We propose a method for exposing the sampling biases in BitTorrent measurements that focuses on both the measured BitTorrent components and the volume of the measured data (Section 3);
2. Using fifteen diverse BitTorrent datasets (Section 4) we show that the measured BitTorrent components and the volume of the measured data can both significantly bias measurement results (Section 5);
3. We formulate recommendations to improve future BitTorrent measurements, and estimate the costs of implementing these recommendations (Section 6).

This work is further motivated by the needs of two ongoing initiatives. First, we are continuing our previous work [13] on building a publicly-accessible P2P Workloads Archive, which will include in a first phase the tens of P2P measurement datasets we have acquired since 2003, and in particular the fifteen datasets we use in this work. Second, within the QLectives project[1] we are currently conducting new measurements of the BitTorrent network, and need guidelines such as the ones presented in this article.

2 Background

In this section we introduce the background needed to understand the remainder of this work. Much of the P2P-related terminology and BitTorrent description in this section is adapted from our previous work on BitTorrent [3, 9, 13].

A P2P *system* is a system that uses P2P technology to provide a set of services; this group of services forms together an application such as file sharing. We call *peers* the participants in a P2P system that contribute to or use the system's resources and services. A peer is completely disconnected until it *joins* the system, and is active until it *leaves* the system. A *real user* may run several peer *sessions*; the sessions are not overlapped in time. We call a *swarm* the group of peers, from all the peers in a P2P system, that interact with each other for a specific goal, such as transferring a file. A swarm starts being active when the first peer joins that swarm, and ends its activity when its last peer leaves. The lifetime of a swarm is the period between the start and the end of the swarm. A *community* is the group of peers who are or can easily become aware of the existence of each other's swarms.

[1] QLectives, EU FP7 project, `http://www.qlectives.eu/`

Our view on P2P systems considers three levels of operation. A P2P system includes at least a peer level, but may also include any of the community and swarm levels. The definitions of community, swarm, and peers presented here are general for peer-to-peer systems, though their implementation may differ with the P2P protocol. For example, BitTorrent and eDonkey have different interpretation and thus implementation of the swarm concept.

In this work we focus on BitTorrent, which includes all the three levels of operation defined earlier in this section. The files transferred in BitTorrent contain two parts: the raw file data and a metadata (directory and information) part. Peers interested in a file obtain the file's metadata from a web site (the community level of BitTorrent) and use the peer location services offered by a tracker (the swarm level of BitTorrent) to find other peers interested in sharing the file. The raw file is then exchanged between peers (the peer level of BitTorrent). To facilitate this exchange, the raw data are split in smaller parts, called *chunks*. Thus, to obtain a complete file a user has to obtain all the chunks by using three application levels.

In this paper we distinguish between the complete and the transient swarm population: we define the *population* of a swarm as the set of all peers ever present in the swarm at any time during the measurement, and a *snapshot* of a swarm as the set of peers present at a specific time during the measurement.

3 Exposing the Sampling Bias

In this section we introduce our method for exposing sampling bias in BitTorrent measurements.

Our method focuses on two main questions that define a measurement process: *What is the relationship between the data source and bias?* and *What is the relationship between the data volume and bias?* The first question stems from the complexity and scale of BitTorrent. For example, there currently exist tens of popular BitTorrent communities, many operating independently from the others and having specific usage characteristics. The second question expresses the trade-off between accuracy and measurement cost (see also Section 6).

3.1 Method

We say that a measurement conducted on a BitTorrent component is affected by a sampling bias if the sampled characteristics are significantly different from the real characteristics of the real BitTorrent component.To analyze the sampling bias we need an understanding of the real characteristics (the *ground truth*), a *conceptual framework* for understanding the differences between the sampled and real characteristics, and *metrics* to quantify these differences.

Ground Truth: The characteristics of BitTorrent are largely unknown: there currently exists no model of a dynamic and heterogeneous BitTorrent swarm, and scientists do not possess even a single complete measurement dataset comprising every message exchanged between the peers of a BitTorrent community of significant size. Thus, and similarly with the situation of exposing sampling biases for Internet measurements [14,8], we need to trace the presence of sampling bias without a ground truth. Instead, we make the observation that if measurements are unbiased, the measured characteristics should remain the same regardless of the sampling. Following this observation, we define for a measurement the *complete dataset*, which is the practical equivalent to the ground truth, as the dataset collected with maximal measurement capability. For example, if a real measurement has sampled data every 5 minutes, it can be used to understand the sampling bias resulting from larger sampling intervals, such as 10 minutes.

Conceptual Framework: We use the term *variability* when comparing properties, e.g., the average peer download speed or the empirical distribution of file sizes for a community, as measured across different BitTorrent components when using the same measurement technique. We also use the term *accuracy* when examining how data collected with different techniques compares with the complete dataset.

Metrics: We estimate the sampling bias using two metrics:
- The **Coverage** metric, defined as the percentage of sampled peers or events, from the peers or events comprised in the complete dataset.
- The **Error/deviation of values** metric, which mimics traditional statistical approaches for comparing probability distributions of random variables. The Kolmogorov-Smirnov test [15] uses the D characteristic to estimate the maximum distance between the cumulative distribution functions (CDFs) of two random variables. Similarly, we use the D characteristic to compare the measured and the complete dataset values. Following traditional work on computer workload modeling (see [16] and the references within), we say that measurements resulting in errors above 10% (D metric above 0.1) have *very low accuracy*, and that measurements with 5–10% error have *low accuracy*.

3.2 Data Sources

Depending on the selection of the data source, we distinguish three main sources of variability or accuracy:

1. The **Measurement level.** In Section 2 we have defined three levels for a P2P application, community, swarm, and peer. Measuring at any single level may result in measurement inaccuracy. For example, measurements taken at peer level, with an instrumented peer as the measurement tool, may fail to contact all peers in a swarm, since peers have limited uptime (presence) in the system.

2. The **Community type.** Many types of communities exist in the BitTorrent world, and this diversity leads to high variability in measurement results. We categorize BitTorrent communities based on the type of content they share, either *general* or *specific content.* The specific content may be further divided into content sub-types such as video, operating system, etc.; Garbacki et al. [17] have identified around 200 content sub-types for the SuprNova community.

3. The **Passive vs. Active Measurements.** Following the terminology introduced in our previous work [9], peer-level measurements are *active* if the instrumented peers acting as measurement probes initiate contact with other

BitTorrent peers, and *passive* if they wait for externally initiated contacts. In contrast to passive measurements, active measurements require that the other peers are accessible, for example, they are not behind a firewall. The 2007 measurement by Xie et al. [18] shows that up to 90% of the peers in a live streaming application are firewalled, and that less than 20% of them by-pass the firewalls.

3.3 Data Volume

The data volume is another major discriminant among measurements:

1. **Sampling rate and Duration.** Since peer-to-peer systems have properties that evolve over time, measurements have to observe the same property repeatedly. Then, the data volume is the product between the sampling rate and the measurement duration; reducing either leads to lower data volumes, but may also lead to inaccuracy. Rates of a sample every 2.5 [3, 9] to 30 minutes [10], and durations of a few days [9] to a few months [2] have been used in practice.

2. **Number of communities and Number of swarms.** BitTorrent communities may share properties, and within a community the most populated swarms may account for most of the traffic. Thus, including in the measurement fewer communities and swarms may reduce the volume of acquired data without reducing accuracy. Until the recent study of four communities [10], measurements have often focused on one community [3, 9], and even on only one swarm [2].

3. **Long-term dynamics.** Many BitTorrent communities have changed significantly over time or even disappeared. Thus, measurements should make efforts to catch long-term system dynamics, including seasonal and multi-year patterns. In practice, the only long-term BitTorrent measurements are the five months study of Izal et al. [2] and our own year-long measurement of SuprNova [3].

4 The Collected Traces

To understand sampling bias in BitTorrent measurements, we have acquired 15 long-term traces from 9 BitTorrent communities of hundreds of thousands of peers and responsible for transferring yearly over 13 peta-bytes of data. Overall, this paper investigates the largest number of BitTorrent datasets, to-date, as summarized in Table 1; for a complete description of the traces see our technical report [19]. Traces studied in this work are available at the Peer-to-Peer Trace Archive (`http://p2pta.ewi.tudelft.nl`); for more details and analysis results of the Archive please refer to our technical report [20].

To ensure heterogeneity among the limited number of traces, we have taken into account the following controllable factors when collecting the traces. The traces cover different community types (sharing from very specific to general content), community sizes (from small communities to the largest in the world of the time of the data collection–T1'04/SuprNova in 2004 and T2'05/PirateBay in 2005), and measurement duration (from a few days to one year). To assess the bias introduced by not following the long-term evolution of BitTorrent communities, several traces include two datasets, acquired in 2005 and 2009 respectively.

Table 1. Summary of the datasets used in this work. Only the datasets for traces T1'04 and T2'05 have been previously analyzed [3, 9].

ID	Trace Description (Content Type)	Period	Sampling (minutes)	Torrents	Sessions	Traffic (GB/day)
T1'04	BT-TUD-1, SuprNova (General)	Oct 2003 to Dec 2004	60	32,452	n/a	n/a
		06 Dec 2003 to 17 Jan 2004	2.5	120	28,423,470	n/a
T2'05	BT-TUD-2, PirateBay (General)	05-11 May 2005	2.5	2,000	35,881,338	32,000
T3'05	LegalTorrents.com	22 Mar to 19 Jul 2005	5	41	n/a	698
T3'09	(General)	24 Sep 2009 onwards	5	183	n/a	1,100
T4'05	etree.org	22 Mar to 19 Jul 2005	15	52	165,168	9
T4'09	(Recorded events)	24 Sep 2009 onwards	15	45	169,768	143
T5'05	tlm-project.org	22 Mar to 30 Apr 2005	10	264	149,071	735
T5'09	(Linux)	24 Sep 2009 onwards	10	74	21,529	15
T6'05	transamrit.net	22 Mar to 19 Jul 2005	5	14	130,253	258
T6'09	(Slackware Unix)	24 Sep 2009 onwards	5	60	61,011	840
T7'05	unix-ag.uni-kl.de	22 Mar to 19 Jul 2005	5	11	279,323	493
T7'09	(Knoppix)	24 Sep 2009 onwards	5	12	160,522	348
T8'05	idsoftware.com	22 Mar to 19 Jul 2005	5	13	48,271	19
T8'09	(Game Demos)	24 Sep 2009 onwards	5	37	14,697	12
T9'05	boegenielsen.net (Knoppix)	22 Mar to 19 Jul 2005	5	15	36,391	308

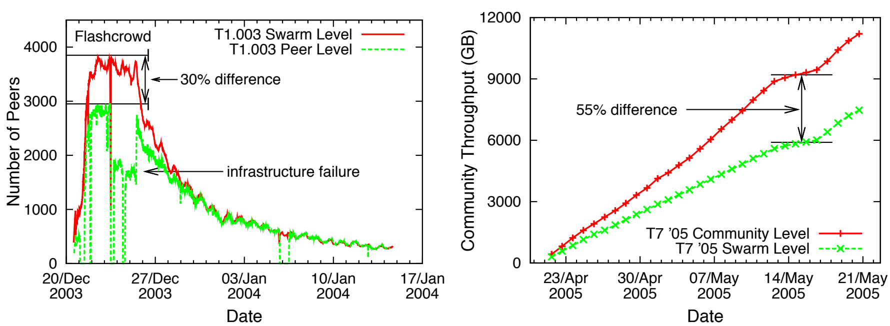
5 The Results

In this section we investigate the effects of the different ways of data source selection and data volume reduction on the variability and accuracy of BitTorrent measurements.

5.1 Data Source Selection

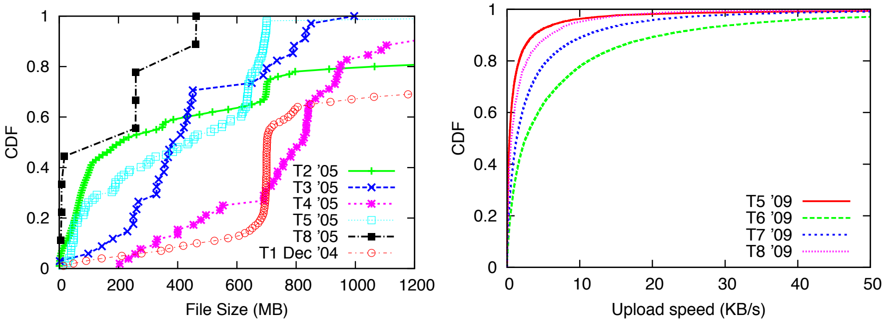
First we will assess the effects of the selection of the data source on the accuracy and variability of BitTorrent measurements.

Finding: Measurements performed at a single operational level of BitTorrent can lead to very low accuracy. When measuring swarm dynamics, we observe in T1'04 that in 4 out of 10 swarms, peer level measurements capture



(a) Swarm dynamics **(b)** Cumulative community throughput

Fig. 1. Comparison of the results obtained from measuring at different levels



(a) File Size Distributions **(b)** Upload Speed Distributions

Fig. 2. Comparison of properties in different BitTorrent communities

many fewer peers than swarm level measurements. For swarm 003 of T1'04, during the flashcrowd (see Figure 1a), the peer-level coverage drops below 70% of the swarm-level coverage. Later, due to the overloaded infrastructure, the peer-level coverage is even down to 50% for about half the duration of the flashcrowd. Similar effects can be observed from measurements at both the community and the swarm level in many communities. Taking T7'05 as an example, the community throughput as obtained from community-level measurements is more than 50% higher than the swarm-level numbers, and so the latter are very inaccurate (see Figure 1b).

Finding: Measuring different BitTorrent communities may lead to very different results. As discussed in Section 3, the diversity in community types contributes to a high variability in the measurement results of different communities. For several BitTorrent communities, we show in Figures 2a and 2b the cumulative distribution functions (CDFs) of the file size and the upload speed, which both differ significantly among communities. Furthermore, for these communities we do not see a correlation between these characteristics and the focus of the community on general versus specific content. We also observe similar differences in the distributions of the swarm sizes and the download speeds in several communities.

Finding: The results of passive and active measurements differ significantly, because the presence of firewalled peers is significant in BitTorrent, and the firewalled and non-firewalled peers have different uptime. For example, less than 60% of the peers are non-firewalled in the T1'04 (SuprNova) trace [3]. An in-depth analysis of the impact of (the fraction of) firewalled peers on upload/download performance in four communities was presented by Mol et al. [21]; their analysis also covers the data of T2'05 (The Pirate Bay), which were collected using both active and passive measurements. It turns out that over 65% of the peers discovered using the active measurements are firewalled, and that 96% of the swarms have over 50% firewalled peers. The same study found that, because BitTorrent rewards peers with connectivity, non-firewalled peers exhibit

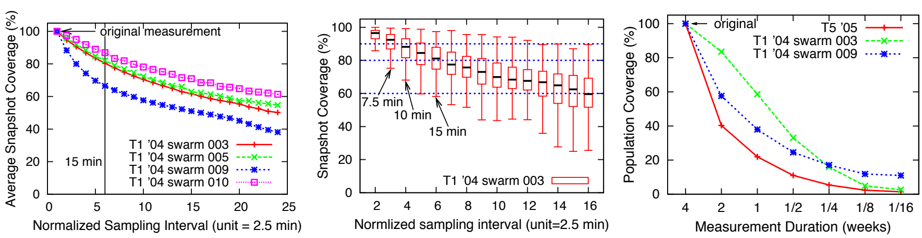
80% less uptime than firewalled peers. However, it is not possible to only perform passive measurements, since in this case it is costly to guarantee coverage and a steady sampling rate.

5.2 Sampling Rate and Duration

In order to understand the effect of the sampling rate on the measurement accuracy, we take the original datasets obtained in our measurements, with their original sampling rates and durations, as the basis. From these datasets we derive new datasets with other (lower) sampling rates by sampling the original datasets at various intervals. Similarly, we derive datasets with shorter durations by simply taking contiguous pieces of the appropriate lengths of the original datasets. We then compare the properties of the original and newly obtained datasets.

Finding: When measuring at the peer level, increasing the sampling interval leads to a higher inaccuracy and variance. Figure 3a shows how the average snapshot coverage of several swarms drops when the sampling interval is increased from 2.5 minutes (the original sampling interval) to 30 minutes. Figure 3b shows the statistics (min, max, median, and first and third quartile) of the distribution of the snapshot coverage of a single swarm obtained at different multiples of the original sampling interval. Only when the sampling interval does not exceed 7.5 (15) minutes is the median snapshot coverage at least 90% (80%).

Finding: Reducing the measurement duration quickly reduces the coverage of the measurements—a doubling of the sampling interval leads to smaller coverage loss than a halving of the measurement duration. Figure 3c depicts for several datasets the population coverage resulting from various measurement durations, including the original duration of 4 weeks. The different datasets exhibit different losses of accuracy when cutting the original measurement duration in half, but all quickly converge to less than 20% coverage with subsequent duration cuts. Swarm 003 from trace T1'04 is the least affected by halving the original measurement duration at over 80% coverage, but the coverage of the complete community in T5'05 is then already below 40%. This large difference is the result of the system evolution: swarm 003 exhibits a large



(a) Average snapshot coverage

(b) Statistics of snapshot coverage

(c) Population coverage

Fig. 3. The effects of changing sampling interval and measurement duration

flashcrowd [3] in which the peers are caught for at least a week until obtaining the content they want, while in the tlm-project.org community the peers obtain results quickly and then leave the swarm without returning.

5.3 Number of Communities and Torrents

To understand the effect of the number of measured communities on the measurement accuracy, we select one month worth of data from six communities in our datasets. We order these communities by the total amounts of traffic generated by that community, and we compute all the investigated characteristics for all the selected communities together. Then we iteratively remove from the considered datasets the data of the community with the lowest rank (total traffic), and repeat the whole process until only one community is left for analysis. We apply a similar approach to understand the effects of the number of torrents, in this case ordering the torrents by the number of peers in the corresponding swarms. This approach has been taken by many reported measurements [3, 9].

Finding: Measuring only one community or only one swarm is insufficient to obtain accurate results. Figure 4 depicts the CDF of the session length for a varying number of communities. The CDF stabilizes only after at least four communities are considered together. We have obtained similar results for other characteristics, including the download speed and the total amount of data downloaded per peer (see [19] for complete results). Similar results are found when investigating the effects of the number of measured swarms. The upload speed is the only characteristic that we have investigated that does not require multiple communities to be measured to obtain accurate results (see [19]).

5.4 Catching Long-Term Dynamics

To show evidence of the long-term evolution of BitTorrent we first extract from our long-term traces blocks of contiguous data and then compare them.

Finding: Yearly patterns exist in BitTorrent, but are diverse across different communities. Multi-year evolution is also present but is hard to characterize. Figure 5 depicts the evolution of the file sizes from Dec 2003

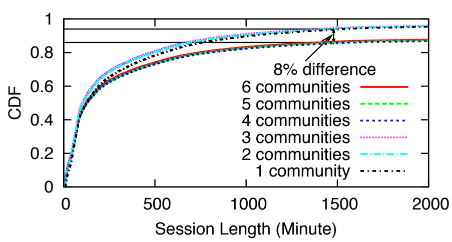


Fig. 4. Session Length Distributions from measuring different numbers of communities

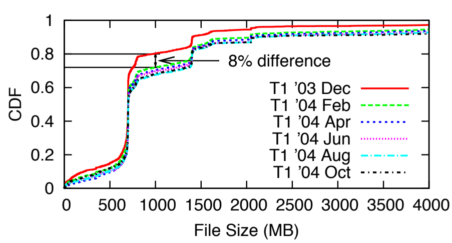


Fig. 5. Monthly file size distributions from T1 (SuprNova). Only every second month is depicted.

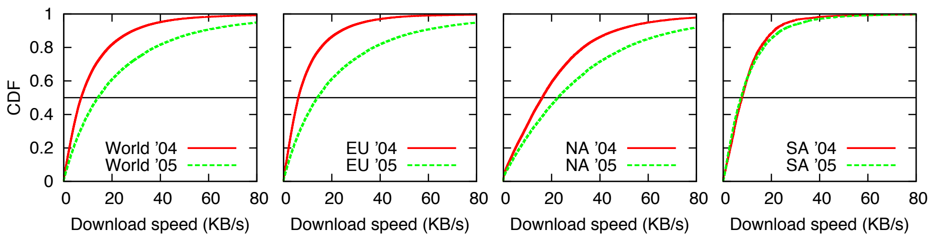


Fig. 6. Change of download speed in different continents and in different years. (Data from trace BT-TUD-1, BT-TUD-2.).

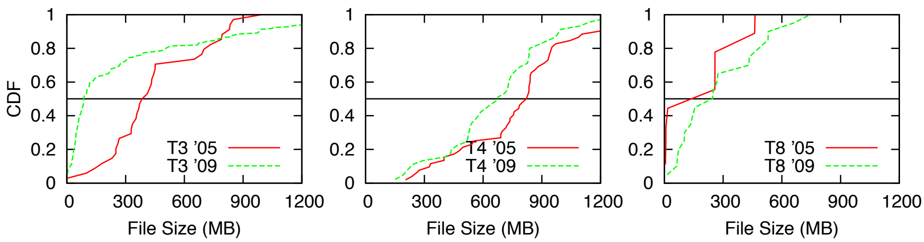


Fig. 7. Change of file size from 2005 to 2009, by community

to Nov 2004; for clarity, the figure only shows curves corresponding to every second month. The results taken in Dec 2003 reveal very different values of this characteristic versus the other measurements. Smaller differences appear between consecutive months, and overall the file sizes decrease slowly over time. We also show the evolution of swarm sizes and community in [19]. In our previous work [9] we have observed that the average download speed has doubled between T1'04 and T2'05.

We now show that the evolution is not consistent across all users. Figure 6 depicts the download speed distributions for T1'04 and T2'05 with users grouped by continent. The left sub-graph confirms our previous remark, and the other sub-graphs show that the increase of download speed varies among the different continents. The increase of download speed by community is also found for traces T5-T8 (see [19]). Similarly, we show in Figure 7 evidence that the file size distribution changed from 2005 to 2009, but the actual "direction" of the change varies greatly by community.

6 Recommendations

Based upon our findings in section 5, we formulate three recommendations for performing BitTorrent measurements with reduced sampling bias:

1. Measure at multiple operational levels of BitTorrent;

2. Use a short sampling interval (below 10 minutes) for peer level measurements;

3. Measure several communities (at least 4) and swarms in each community (at least 50) for long periods of time (months or longer).

We now estimate the cost of using these recommendations for new BitTorrent measurements, based on our past experience [9]. Following our recommendations, the new measurements should sample 200 swarms across 4 communities, use a sampling interval of 10 minutes, include both active and passive measurements, and have a duration of one year. To collect the trace T2'05, we have sampled 2,000 swarms from the PirateBay community every 2.5 minutes for 6 days, performing both passive and active measurements. We have used in the process about 300 computers and generated about 300GB of raw data. Assuming that new measurements are performed using an infrastructure with similar capabilities, they will require the use of 30 machines and 500 GB storage space for a period of one year. Excluding human and data curation costs, and using the cheapest cloud resources from the Amazon Web Services (compute resources cost $0.085 per hour, data storage $0.150 per GB, all data transfers free until June 30, 2010, $0.10 per GB afterwards) the total cost is $22,463, of which only $125 are data transfer and storage costs.

7 Related Work

Much previous work has been dedicated to measurements of P2P file-sharing networks [1, 2, 3, 4, 5, 6]. From the few studies that tried to reduce the sampling bias or even recognized its existence, none covers the full set of sampling bias sources addressed in this work.

In general, under the assumption that "more is better", these studies obtained data over long periods of time [10,22], from more peers [10,6,9], for more files [10, 9, 23] and communities [10, 23], and filtered the raw data before analysis to eliminate some of the measurement biases [10, 6].

Closest to our work, Stutzbach et al. [24] assess the bias incurred by sampling unstructured P2P file-sharing networks, and propose the MRWB technique to sample without bias from such networks. Stutzbach et al. [22] further investigate the bias introduced by the sampling rate. However, their analysis and techniques are valid only for unstructured networks, but do not apply to the tracker-based, centralized peer discovery mechanism in BitTorrent.

8 Conclusion

Despite almost a decade of BitTorrent measurements, little is known about the sampling bias of the measurement results. However, reliable measurements of BitTorrent deployments are needed to improve the quality of service for millions of BitTorrent users. In this paper we have presented an investigation of the factors that cause inaccuracy in and variability among BitTorrent measurements. Towards this end, we have first proposed a method for exposing sample bias. Our method includes a taxonomy of the sources of sampling bias comprising two axes—data source selection and data volume reduction—totaling six sources, and two metrics for quantifying the sampling bias. Then, we have evaluated the effects of the different sources of sampling bias using fifteen real traces taken from nine BitTorrent communities. Our results indicate that current measurement techniques can lead to significant sampling bias in the measurement results. Based on our findings, we have formulated three recommendations to improve future BitTorrent measurements, and analyzed the cost of implementing them in practice. For the future, we plan to extend our work towards developing a complete method for accurate yet low-volume BitTorrent measurements.

Acknowledgements

The research leading to this contribution has received funding from the European Community's Seventh Framework Programme in the P2P-Next project under grant no 216217. The Peer-to-Peer Trace Archive (`http://p2pta.ewi.tudelft.nl`) provides us with the traces studied in this work.

References

1. Sen, S., Wang, J.: Analyzing peer-to-peer traffic across large networks. In: Proc. of ACM SIGCOMM IMW, pp. 137–150 (2002)
2. Izal, et al.: Dissecting BitTorrent: Five Months in a Torrent's Lifetime. In: Proc. of PAM, Antibes Juan-les-Pins, France, pp. 1–11 (2004)
3. Pouwelse, J., Garbacki, P., Epema, D., Sips, H.: The BitTorrent P2P file-sharing system: Measurements and analysis. In: Castro, M., van Renesse, R. (eds.) IPTPS 2005. LNCS, vol. 3640, pp. 205–216. Springer, Heidelberg (2005)
4. Bhagwan, R., Savage, S., Voelker, G.M.: Understanding availability. In: IPTPS, pp. 256–267 (2003)
5. Gummadi, K., Dunn, R., Saroiu, S., Gribble, S., Levy, H., Zahorjan, J.: Measurement, modeling, and analysis of a peer-to-peer file-sharing workload. In: ACM Symp. on Operating Systems Principles, SOSP (2003)
6. Handurukande, S.B., Kermarrec, A.M., Fessant, F.L., Massoulié, L., Patarin, S.: Peer sharing behaviour in the eDonkey network, and implications for the design of server-less file sharing systems. In: EuroSys, pp. 359–371 (2006)
7. Arlitt, M.F., Williamson, C.L.: Web server workload characterization: The search for invariants. In: SIGMETRICS, pp. 126–137 (1996)
8. Floyd, S., Paxson, V.: Difficulties in simulating the Internet. IEEE/ACM Trans. Netw. 9(4), 392–403 (2001)
9. Iosup, A., Garbacki, P., Pouwelse, J., Epema, D.: Correlating topology and path characteristics of overlay networks and the Internet. In: Proc. CCGrid, p. 10 (2006)
10. Andrade, N., Santos-Neto, E., Brasileiro, F.V., Ripeanu, M.: Resource demand and supply in bittorrent content-sharing communities. Computer Networks 53(4), 515–527 (2009)
11. ipoque GmbH: Internet studies (2006-2009), `http://www.ipoque.com/resources/internet-studies/`
12. Parker, A.: The True Picture of Peer-To-Peer File-Sharing. In: IEEE Int'l. W. on Web Content Caching and Distribution Panel (2005)
13. Zhang, B., Iosup, A., Garbacki, P., Pouwelse, J.: A unified format for traces of peer-to-peer systems. In: LSAP, pp. 27–34. ACM, New York (2009)

14. Lakhina, A., Byers, J.W., Crovella, M., Xie, P.: Sampling biases in ip topology measurements. In: INFOCOM (2003)
15. Lilliefors, H.W.: On the Kolmogorov-Smirnov test for normality with mean and variance unknown. J. Am. Stat. 62, 399–402 (1967)
16. Feitelson, D.G.: Workload modeling for performance evaluation. In: Performance, pp. 114–141 (2002)
17. Garbacki, P., Epema, D., van Steen, M.: Optimizing peer relationships in a super-peer network. In: ICDCS, p. 31 (2007)
18. Xie, S., Keung, G.Y., Li, B.: A measurement of a large-scale peer-to-peer live video streaming system. In: Proc. of ICPP, p. 57 (2007)
19. Zhang, B., Iosup, A., Pouwelse, J., Epema, D., Sips, H.: On assessing measurement accuracy in BitTorrent peer-to-peer file-sharing networks. Tech.Rep. PDS-2009-005, TU Delft (2009), `http://pds.twi.tudelft.nl/reports/2009/PDS-2009-005.pdf`
20. Zhang, B., Iosup, A., Epema, D.: The peer-to-peer trace archive: Design and comparative trace analysis. Technical Report PDS-2010-003, Delft University of Technology (2010), `http://pds.twi.tudelft.nl/reports/2010/PDS-2010-003.pdf`
21. Mol, J., Pouwelse, J., Epema, D., Sips, H.: Free-riding, fairness, and firewalls in p2p file-sharing. In: P2P, pp. 301–310 (2008)
22. Stutzbach, D., Rejaie, R., Sen, S.: Characterizing unstructured overlay topologies in modern P2P file-sharing systems. IEEE/ACM Trans. Netw. 16(2), 267–280 (2008)
23. Guo, L., Chen, S., Xiao, Z., Tan, E., Ding, X., Zhang, X.: Measurements, analysis, and modeling of bittorrent-like systems. In: Internet Measurment Conference, pp. 35–48 (2005)
24. Stutzbach, D., Rejaie, R., Duffield, N.G., Sen, S., Willinger, W.: On unbiased sampling for unstructured peer-to-peer networks. IEEE/ACM Trans. Netw. 17(2), 377–390 (2009)