Spatial and Temporal Locality of Content in BitTorrent: A Measurement Study

Taejoong Chung*, Jinyoung Han*, Hojin Lee†
Jussi Kangasharju§, Ted “Taekyoung” Kwon*, Yanghee Choi*
Seoul National University, Korea*, KAIST, Korea†, University of Helsinki, Finland§
Email: {tjchung, jyhan}@mmlab.snu.ac.kr, hojin.lee.79@gmail.com
Jussi.Kangasharju@helsinki.fi, {tkkwon, yhchoi}@snu.ac.kr

***Abstract*—**

We conduct comprehensive measurements on content locality in one of the largest BitTorrent portals: The Pirate Bay. We investigate locality phenomena from a content perspective, considering the content category, publisher, and popularity of content. In particular, we focus on (i) how content is consumed from spatial and temporal perspectives, (ii) what makes content be consumed with disparity in spatial and temporal domains, and (iii) how we can exploit the content locality. We find that content consumption in real swarms is 4.56 times and 1.46 times skewed in spatial (country) and temporal (time) domains, respectively. We observe that a cultural factor (e.g., language) mainly affects spatial locality of content. Not only the time-sensitivity of content but also the publishing purpose affects temporal locality of content. We further reveal that spatial locality of content rarely changes on a daily basis (microscopic level), but there is notably spatial spread of content consumption over the years (macroscopic level). Based on the observation, we conduct simulations to demonstrate that bundling and caching can exploit the content locality.

I. Introduction

BitTorrent is one of the most popular peer-to-peer (P2P) applications and responsible for a substantial amount of current Internet traffic [1]. However, its network-oblivious nature has posed a few challenges from the networking perspective. First, P2P connections may incur substantial transit traffic between different networks of Internet Service Providers (ISPs) [2], [3]. Second, its peering strategy may lead to sub-optimal throughput [4]. Third, its time-varying traffic patterns can be a hurdle for traffic engineering [5], [6].

To address the above issues, both ISPs and P2P application developers have considered various alternatives [7], [8], [9], [10]. ISPs have tried many traffic control techniques such as rate throttling or charging [11]. However, this appears to be inefficient because most P2P applications are dodging this control. On the other hand, some P2P applications have adopted techniques to improve efficiency by localizing P2P connections (i.e., preferring peers within the same ISP [7], [8]). These techniques, however, have limitations since network information such as topology, cost, and link status is at best inferred by application-level traffic observations. To overcome these limitations, there have been efforts to promote the cooperation between ISPs and P2P applications; ISPs can provide the above information to P2P applications [9], [10].

Recently, some studies turn their attention to *locality* among peers to fundamentally understand the above problems [12], [13], [14]. Here, the locality generally indicates how much disparity exists in content sharing patterns from the spatial and temporal perspectives. According to [13], [14], 30% more connections among peers in the same ISP compared to a random graph are observed for more than 45% of peers, which indicates that BitTorrent connections among peers are biased to local peers. They further argued that the localized nature of BitTorrent may help both ISPs to reduce inter-ISP traffic and P2P applications to improve the download speed. [12] found that substantial amount of BitTorrent traffic does not reach higher-tier ISPs, which is in line with [13], [14]. The authors of [12] also revealed that BitTorrent's temporal usage patterns are observed to vary in a diurnal fashion. Based on this observation they argued that ISPs need to devise a better price model to balance the traffic over time.

While these studies focus on how much BitTorrent traffic is localized in swarm dynamics, most of them paid little attention to investigating locality phenomena from a content perspective, which we call *content locality*. We focus on the following questions: *How are content files (spatially and temporally) consumed by human beings, and why these phenomena occur in BitTorrent? Are there any skewed patterns in the way people participating in BitTorrent swarms depending on the content properties (e.g., content types or cultural aspects)?* We argue that understanding content locality with emprically-grounded evidences is important for BitTorrent stakeholders: (i) how BitTorrent service providers deal with locality phenomena to improve the system performance and (ii) how content providers publish torrents to increase sales. For instance, ISPs may develop content caching strategies by considering locality phenomena to reduce the inter-ISP traffic.

To our knowledge, this is the first measurement study to comprehensively investigate locality phenomena from a content perspective with data from The Pirate Bay [15], one of the largest BitTorrent portals.

The main contributions of this paper are as follows.

- We show that spatial locality in BitTorrent is mainly affected by content types and publishers rather than by file dissemination mechanisms (e.g., tit-for-tat).
- We show that cultural aspects of content (e.g., the language of a torrent) is a critical factor that decides how users participate in swarms from a spatial perspective.
- We observe that the spatial locality for the most of content

(99%) are rarely changed over the day.

- We find that not only the time-sensitivity of content (e.g., periodicals or TV series) but also the publishing purpose of content affects temporal locality.
- We also find that more pronounced diurnal pattern is observed in leechers (peak-to-trough ratio: 7.29) rather than seeds (peak-to-trough ratio: 3.55).
- We demonstrate how spatial locality can be exploited to reduce inter-ISP traffic (50%) and improve availability in bundling and how temporal locality can be exploited to enhance caching performance.

This paper is structured as follows: §2 reviews the related work and §3 describes our methodology. §4 and §5 analyze the spatial and temporal locality, respectively. After discussing the implications of BitTorrent locality in §6, we conclude this paper in §7.

II. Related Work

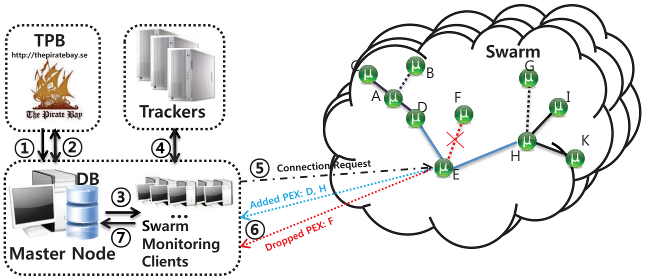
A. Peer Localization

The network-oblivious nature of P2P applications poses the above challenges to both ISPs and P2P application developers. To address these issues, many P2P applications have adopted techniques to improve networking efficiency by localizing application-level peering (i.e., preferably select peers within the same ISP). For example, [7] suggested a client-side solution without any help from ISP, which helps to select peers within the same ISP to achieve the better throughput and reduce the inter-ISP traffic. However, relying solely on peers has fundamental limitations because the network information such as topology, cost, and link status is at best inferred by application-level traffic observations. To overcome the limitation, [9] and [10] suggest explicit cooperation between ISPs and P2P applications to localize traffic within ISP.

B. Locality in BitTorrent

Recently, there have been studies of BitTorrent on locality phenomena to fundamentally understand the BitTorrent's traffic characteristics [12], [13], [14]. [14] analyzed the impact of locality-aware peer selection algorithms of BitTorrent on the inter-ISP traffic and download times of end users. The authors revealed that the localized nature of BitTorrent can help both ISPs by reducing inter-ISP traffic and P2P applications by improving download speeds. The measurement results of [13] showed that 45% of peers have more than 30% peers in the same ISP compared to a random graph, which implies that the current BitTorrent mechanism preferably selects peers in the same ISP. [12] also showed that the major BitTorrent traffic does not reach high-tier ISPs (tier 1 or 2), which signifies that locality phenomena are more prevalent in the stub ISPs. The authors further revealed that BitTorrent temporal usage exhibits the peak in the evening, which implies the presence of temporal locality.

[16], [17] showed different levels of locality phenomena are observed in BitTorrent depending on the link bandwidth of peers and the popularity of content files. The authors of [16] revealed that peers with similar upload/download speeds in a swarm tend to have more connections one another, which is mainly due to the choking algorithm. While these studies on BitTorrent locality have been focusing on traffic characteristics of BitTorrent, our focus is to empirically analyze the the locality of BitTorrent from a perspective of content considering multiple content properties such as content categories (e.g., movie or music).



① RSS notification on new torrents
② Fetch the .torrent files and store
③ Assign torrents to swarm monitoring clients
④ Request peer lists to the tracker(s) every 10 minutes
⑤ Connect each peer based on peer list and exchange PEX
⑥ Keep receiving PEX from each peer for 30 minutes
⑦ Save the snapshot of each swarm every 30 minutes

Fig. 1. We build a measurement framework to capture the torrent data and user behaviors of a real BitTorrent system.

C. Locality in Other Domains

There have been many studies to understand and exploit *"locality phenomena"* in various domains. The locality phenomena can be found in traditional libraries [18] or newspapers [19]. Interestingly, the locality phenomena also can be observed in online social networking (OSN) services in the Internet. [20] investigated the relation between popularity and geographical locality of YouTube videos and showed that sharing videos in OSN widens the geographical reach of the videos. In Twitter, [21] showed the users' geographical proximity with their followers and revealed that language and cultural characteristics determine the level of locality in Twitter. [22] observed that majority of communications in Facebook are occured within the same geographical region. By exploiting the spatial (i.e., geographical) locality phenomena, [22], [23] tried to design the system and improve performance of OSN with mininal infrastructural and operational cost. Inspired from prior work of locality phenomena in various domains, this measurement study empirically investigates content locality in BitTorrent.

III. Methodology

We have conducted a measurement study on one of the most popular BitTorrent portals, The Pirate Bay (TPB). We developed a monitoring client to keep track of swarms by modifying *Azureus* [24]. Figure 1 illustrates the overall measurement framework. To monitor each swarm from its beginning, we leverage the RSS notification of a new torrent to retrieve its publisher's username and .torrent file, from which we obtain peers from its tracker and seeds on the distributed hash table (DHT). In addition to the peer list from the tracker and the

	2010	2011
period	April 30 ~ July 23	April 6 ~ May 9
# of unique ip	13,863,126	15,884,221
# of torrents	80,173	63,793

TABLE I
DATASET DESCRIPTION.

DHT, a swarm monitoring client further leverages the peer exchange extension (PEX) by which we can discover new peers of the already known peers in a swarm.

A. Discovering Swarm Topology

Each monitoring client iteratively asks the list of peers from trackers and DHT every 10 minutes. In addition, to discover the topology of each swarm (i.e., how peers in a swarm are connected to each other), the monitoring client further exploits the peer exchange extension (PEX) [1]. By analyzing the connectivity among peers from the PEX messages, we can retrieve each peer's (peer-connectivity-level) routing table in a swarm.

When we obtain the routing table of a particular peer by analyzing her PEX messages, an entry in the routing table does not necessarily mean that they are actually exchanging the data. In current BitTorrent systems, a peer can have many connections in the routing table; however, only a small portion of connections are used for exchanging data. To identify the connections that are actually used, we consider two types of PEX messages: PEX-Added and PEX-Dropped [26]. Suppose a swarm monitoring agent monitors peer A. Whenever peer A establishes a connection with a new peer, the monitoring client receives a PEX-Added message from peer A, and thus learns when a new peer is added to A's neighbor list. Similarly, a PEX-Dropped message will be sent to A's swarm monitoring agent whenever A's connection to her peer is terminated. We notice that peer A normally drops her connection to peer B because peer B has transmitted little or no data due to B's poor network status or selfish behavior. Hence, a dropped peer within a short duration is unlikely to have exchanged (much) data. Overall, we refine the connections of each peer by removing peers who are dropped shortly during each measurement period. Note that this process is carried out iteratively (e.g., every 10 seconds). Figure 1 also illustrates how we identify the peers of peer E. Peers D and H are included, but peer F is removed since it is dropped shortly.

B. Dataset

Our datasets are composed of two different periods as shown in Table I. For the 143,966 torrents observed during the two periods, the swarm monitoring clients captured snapshots every 30 minutes. The numbers of torrents and IP addresses are described in Table I.

Based on the publisher (ID) information on TPB, we divide publishers (in the datasets) into three types, like [27], [28]: (i) *fake* publishers who publish fake content, (ii) *profit-driven* publishers who publish content for financial incentives, and (iii) *altruistic publishers* who publish content only for sharing. Also, we investigate the locality depending on the different content categories given at TPB: TV, Porn, E-book, Movie, Music, Application, and Game. We further identified the user's locale using the MaxMind database [29], which maps each IP address (of a peer) to its country or autonomous system (AS) [2]. There are 168 countries and 11,191 ASes in our datasets.

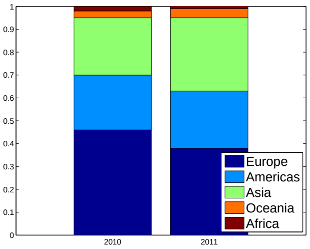


Fig. 2. Peer distribution in aspects of continental level is plotted.

Rank	2010 Country	 Portion	2011 Country	 Portion
1	United States	14%	United States	14%
2	United Kingdom	7%	India	9%
3	India	7%	United Kingdom	8%
4	Spain	6%	Canada	5%
5	Italy	5%	Korea	4%
6	Canada	5%	Italy	4%
7	France	4%	Australia	4%
8	Sweden	3%	China	3%
9	Australia	3%	Sweden	3%
10	China	3%	Brazil	3%

TABLE II
THE PORTION OF EUROPE IS DECREASED (E.G, SPAIN: RANK 4 → RANK 11 AND FRANCE: RANK 7 → RANK 15). ASIAN COUNTRIES SHOW AN INCREASE, E.G., KOREA: RANK 33 → RANK 5.

C. Representativeness

We now analyze the representativeness of the above datasets. Figure 2 shows the distribution of peers per continent from 2010 to 2011. The portion of peers in the Europe has decreased from 46% to 38%. The peers in Asia, in contrast, have increased from 25% to 32%. For reference, we also compare the observed distribution with the previous studies [12], [30] and find a similar pattern. For example, declining BitTorrent usage in Europe is also reported in [12], which is aligned with previous reports that European users are increasingly using direct download sites instead of P2P [31]. To compare and investigate the usage patterns in different countries, we also mapped peers at country level. Table II shows the top-10 countries from 2010 to 2011 sorted by peer population, and we can find the decline in Europe and growth in America and Asia (e.g, the proportions of Spain, Italy, France, and Sweden are decreased while that of India, Korea, and Australia are increased.) Overall, we find little bias of distribution in our datasets compared with prior works.

[1] Most widely used BitTorrent client software such as uTorrent and Vuze already supports PEX [25].

[2] Note that Maxmind exhibits 99.8% accuracy in country-level.

IV. Spatial Locality

A. *Locality Metrics*

In this subsection, we first introduce three metrics for spatial locality: (i) swarm locality, (ii) community locality, and (iii) neighbor locality. Note that these metrics have different searching scopes of BitTorrent peers for calculating the spatial locality; the swarm, community, and neighbor locality consider the whole peers in the given swarm, the peers in the same community, and the neighbor peers, respectively.

We first model swarm S as a graph $S = (V, E)$, where V is the set of peers (or nodes) participating in swarm S, $\{v_1, \cdots, v_n\}$, and E is the set of bidirectional edges between peers, $\{e_1, \cdots, e_m\}$. Let $L(v)$ denote the locale (e.g., AS or country) of peer v. We define *swarm locality* as the probability that randomly-selected two nodes (in the same swarm) have the same locale[3]:

$$\frac{2}{n(n-1)} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \delta\left(L\left(v_i\right), L\left(v_j\right)\right),$$

where $\delta(i, j)$ is the Kronecker's delta ($\delta(i, j) = 1$ if $i = j$, and $\delta(i, j) = 0$ otherwise). Here, the swarm locality considers not whether two peers have a connection (or edge), but whether they are interested in the same content.

We devise another metric: *community locality*. A community is a group of peers in a swarm, within which connections are denser, but between which connections are sparser. In that sense, we assume that it is more likely to be more traffic inside a community rather than outside of the community. We identify communities using the Louvain method [32], which is a well known algorithm that can quickly find the community and maximize the ratio of the number of edges within communities to that of edges between communities. Community locality is defined as the probability that randomly selected two peers *within the same community* have the same locale. Suppose we have c communities in a swarm, and community k consists of nodes $V^k = \{v_1^k, \cdots, v_{n_k}^k\}$. Then, community locality is

$$\frac{2}{\sum_{k=1}^{c} n_k\left(n_k-1\right)} \sum_{k=1}^{c} \sum_{i=1}^{n_k-1} \sum_{j=i+1}^{n_k} \delta\left(L\left(v_i^k\right), L\left(v_j^k\right)\right).$$

Since peers within the same community are likely to exchange more chunks directly or indirectly than those in different communities, community locality captures what fraction of interactions among peers (i.e., exchanging pieces) are localized.

We also consider the ratio of the actual number of peer v's neighbors with the same locale to the expected number of neighbors with the same locale assuming purely random assignment of neighbors among all the peers in its swarm [13], which we call *neighbor locality*. For instance, suppose peer v has 100 peers in the same swarm, and 40 of those are in the same locale as v. If v has currently 10 neighbors, with 5 of them in the same locale, then v's neighbor locality is 5/4. If the neighbor locality is close to unity, this means that the number of neighbors with the same locale is almost same as expected one, indicating that the peer selection mechanism exhibits marginal locality. The neighbor locality is somewhat limited in the sense that it considers only direct neighbors.

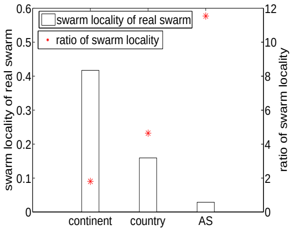


Fig. 3. The ratio of the swarm locality of real swarms to that of uniformly distributed hypothetical swarms is plotted.

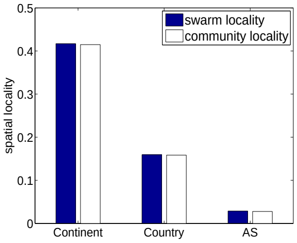


Fig. 4. There is no significant differences between swarm locality and community locality.

B. *Swarm, Community, and Neighbor*

To see if spatial locality exists, we plot the swarm locality of real swarms, compared with that of hypothetical swarms where peers are uniformly distributed among all locales. Note that the numbers of vertices and edges of the hypothetical swarms are preserved, respectively. In Figure 3, we observe that swarm locality of real swarms is significantly higher than that of hypothetical ones. As locale changes from continents to countries and to ASes, the ratio increases from 1.80 times to 4.56 times and to 11.49 times, respectively. This implies that users of the same torrent are spatially biased, which becomes stronger as locale size decreases.

We then examine the effect of BitTorrent's dissemination mechanism on locality. We calculate the neighbor locality for the real swarms, which equals 0.98, 0.98, and 0.99 when the locale is a continent, a country and an AS, respectively. It seems that the peer selection algorithm in BitTorrent contributes little to spatial locality.

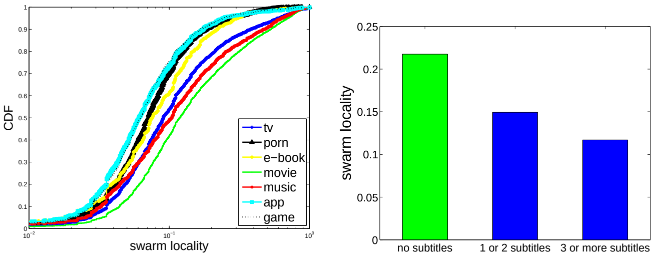
For community locality, we first check whether and how swarms make groups (i.e., communities) by calculating the modularity from the Louvain Method [32]. The modularity is defined as

$$Q = \frac{1}{2m} \sum_{i,j} \left[1 - \frac{k_i k_j}{2m}\right] \delta(c_i, c_j)$$

where k_i is the degree of node i, m is the summation of node degrees for all nodes, c_i is the community where node i belongs, and δ is the Kronecker delta. We find that the average modularity is 0.75, and 70% of swarms exhibit modularity higher than 0.7. In general, modularity above 0.3 indicates a strong presence of community structures in a swarm [33]. Figure 4 also reveals that the average community locality is similar to the average swarm locality.

In conclusion, even if there is a high locality in the wild (Figure 3), there is little difference between a measure which reflects connections (i.e., community locality) and a measure not reflecting connections (i.e., swarm locality) and, moreover,

[3] In this case, we consider all possible connections among peers regardless of traffic exchanged.



(a) Category

(b) Movie

Fig. 5. (a) Swarm locality of each content category is shown. (b) The number of subtitle files affects spatial locality of Movie torrents.

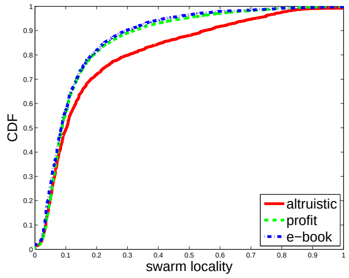


Fig. 6. Swarm locality of each publisher type is shown.

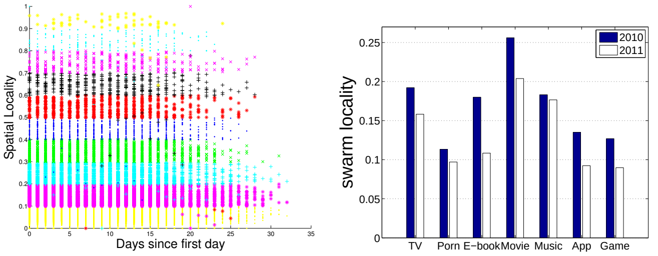
neighbor locality is close to unity. This implies that locality seems not to be much influenced by BitTorrent's sharing mechanism. Instead, we conjecture that locality is more influenced by content itself. Hence, unlike previous studies that investigate the locality only in terms of traffic, our analysis reveals the crucial role of content for better understanding of locality.

C. *Content Categories, Publishers, and Popularity*

We now investigate the spatial locality in country level depending on content categories, publishers, and consumers.

Categories: We plot the CDF of swarm locality for each content category in Figure 5(a). We see that torrents in Movie and TV categories have higher swarm locality while the ones in Porn category exhibit lower swarm locality, even though the three categories are all video-centric. We believe that the disparity across the three categories is due to the style of content consumption; movies or video content typically require understanding of content through language and culture, while porn films typically do not need such background.

To further investigate the effect of languages on locality, we examine the Movie torrents in terms of the number of subtitles. Among total 1,597 torrents, 364 torrents have one or more subtitles. We observe that (i) torrents with no subtitles show 46% higher swarm locality than the others, and (ii) as the number of subtitles increases, swarm locality becomes smaller, as shown in Figure 5(b). This is because movies with more subtitles can be consumed by users in more locations (i.e., with different languages), resulting in less spatial locality. This was conjectured in [14], but there has been no empirical study. Moreover, this result is complement to two recent work in OSN [21], [22]; not only in OSN (i.e., Twitter and Facebook) but also in BitTorrent, cultural characteristics (i.e., language) determines the level of content locality.



(a) Spatial locality on a daily basis

(b) Spatial locality over the years

Fig. 7. spatial locality of content rarely changes on a daily basis (microscopic level), but there is notably spatial spread of content consumption over the years (macroscopic level) in country level.

Figure 5(a) shows that torrents in App and Game categories have low swarm locality. For most App and Game torrents, multi-language-support packages are either included in the main program or downloadable from web sites. Hence, language is not an important factor. Also, application and game software often targets global markets, and thus their torrents are usually downloaded by users without regional inclination. For example, popular software torrents (e.g., Windows 7, Photoshop, or AutoCAD) account for 45% of App torrents.

Publisher Types: We examine swarm locality depending on publisher types: altruistic, profit-driven and fake publishers in Figure 6. We observe that torrents of profit-driven and fake publishers have lower swarm locality than the ones of altruistic ones. To examine the difference in swarm locality depending on the publisher types, we further analyze it with content categories. We find that porn torrents constitute 39% of the torrents uploaded by profit-driven publishers, aligned with [27], while porn torrents are found in 5% and 2% of the torrents of altruistic and fake publishers, respectively. Overall, the very low swarm locality of Porn torrents explains why profit-driven publishers' torrents exhibit low swarm locality.

We examine torrent titles of fake publishers and find that most torrents of fake publishers have attractive titles like those of latest popular movies. For example, the portion of Movie torrents with titles containing '2011' (i.e., torrents of latest content) but whose publishers are not fake is only 22% (384 out of 1752). On the contrary, among all the Movie torrents of fake publishers, the ratio of torrents whose titles containing '2011' is 60% (117 out of 194). As popular titles of torrents of fake publishers are attractive to users worldwide [27], [28], their naming convention results in lower swarm locality.

Popularity: We investigate the correlation between the number of downloaders (or popularity) of a swarm and its swarm locality by calculating the Pearson's coefficient, which is -0.004. Thus, the swarm locality has no or little correlation with the number of downloaders.

D. *Spatial Locality Over Time*

This subsection first analyzes whether and how spatial locality changes over time in Figure 7(a). To this end, we

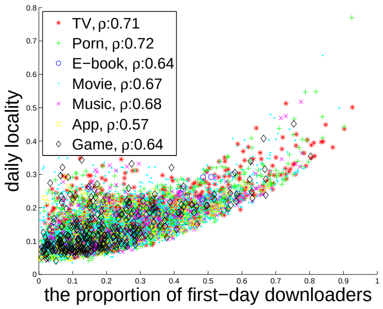


Fig. 8. Daily locality according to the proportion of the first-day downloaders is plotted.

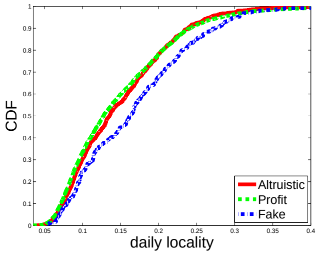


Fig. 11. Daily locality of each publisher type is shown.

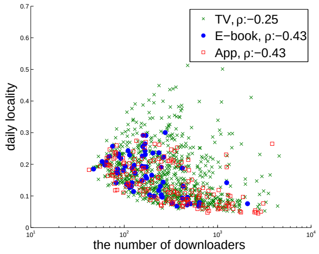


Fig. 12. App and E-book exhibit strong correlation between daily locality and popularity.

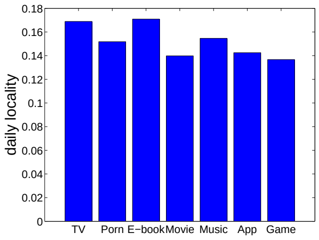


Fig. 9. Daily locality is plotted for seven content categories.

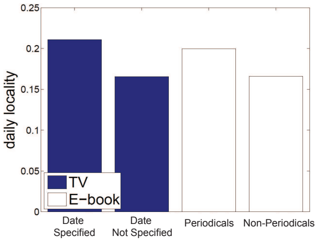


Fig. 10. Air dates and publication dates affect temporal locality.

divide the locality into ten bins with a unit of 0.1 and plot localities of each torrent during its life time from the birth (i.e., the published day of the torrent) to death (i.e., the last day at which there is no more seed in the swarm of the torrent). We assign a color to each torrent according to the locality observed in the first day. For example, if the locality (in the first day) of a torrent *"A"* is 0.55, red color is assigned to *"A"*. Even if its locality changes (e.g., 0.13 in the next day), its color is still red which has been assigned in the beginning. Interestingly, spatial locality of a content rarely changes over time as shown in Figure 7(a). Among 32,489 plots in Figure 7(a), only 74 ones have moved from the original bin to another bin. This signifies that content locality has a time-invariant property from a spatial perspective. We believe this property may be helpful for content/network providers in content caching or prefetching. For example, CDN (Content Delivery Network) providers can decide an adequate server to prefetch/cache by exploiting the time-invariant property.

We next compare the spatial localities of 2010 and 2011 to investigate how the spatial locality changes over the years. Figure 7(b) shows the average spatial localities of 2010 and 2011 across diffrent content categories. Interestingly, the spatial locality decreases as years go on; this imples that content sharing patters are increasingly globalized.

V. Temporal Locality

To see how swarm dynamics behave temporally, we define a metric: *daily locality* to indicate the probability that two peers in the same swarm download the torrent in the same day.

A. Existence of Temporal Locality

Like spatial locality, we find that daily locality of swarms is higher (1.46 times) than that of the hypothetical uniform distribution, indicating that swarm dynamics in terms of population are temporally skewed [34].

We notice that the number of users downloading a torrent tend to be highest for the first day after the torrent is published. As shown in Figure 8, positive correlation (0.53) exists between the percentage of first-day downloaders and the daily locality. This means that for all downloads in a given period, a substantial fraction of downloads happen on the first day.

Interestingly, TV and Porn contents exhibit stronger positive correlation (0.71 and 0.72) than others. We conjecture that for TV content, the air dates of TV drama episodes are fixed each week, thus many consumers already expect its publication. For Porn torrents, we conjecture that users usually download porn contents not through searching but through navigating recent contents from the meta-torrent site (e.g., TPB), resulting in older torrents being less likely to be downloaded.

B. Categories, Publishers, and Popularity

We now analyze the temporal locality of torrents depending on the content categories, publishers, and consumers.

Categories: Figure 9 shows the daily locality across the seven content categories. TV torrents exhibit higher temporal locality than torrents of other content categories except for E-book. To investigate why TV torrents show high daily locality, we first analyze their periodic nature by checking whether their titles have the form of 'S**E**', where 'S' and 'E' stand for season and episode, respectively. We find that 58% of titles (of TV torrents) follow this naming convention; for instance, a title of a torrent the drama "Game of Throne: Season 1 - Episode 6" can be "Game of Throne S01E06". The torrents with this naming convention are likely to be published weekly when a new episode is aired. However, even though a torrent has a title of 'S**E**', it does not guarantee that it is published recently. Thus, we further check the torrents whose titles include the air dates of the particular episodes (e.g., The.Daily.Show.2011.03.30). As shown in Figure 10, the TV torrents whose titles include the air dates exhibit higher temporal locality (0.211) than the others (0.165).

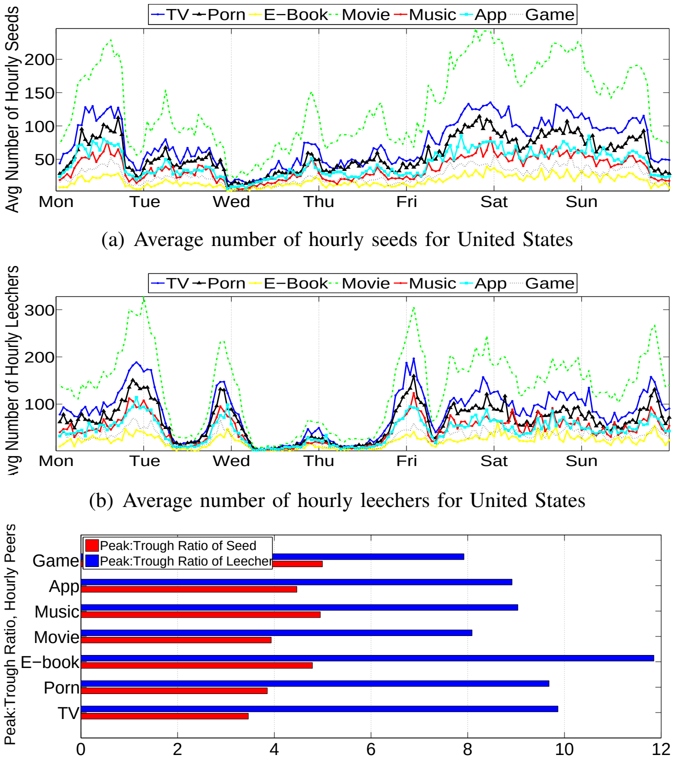
Interestingly, E-book torrents also show high temporal locality. Our investigation reveals that E-book torrents have high temporal locality since (i) periodicals are published at fixed intervals (weekly or monthly; thus, publication dates are easily expected), and (ii) the lifetime of an E-book torrent is shorter than torrents in other content categories. First, the torrents of E-book periodicals show higher temporal locality than those of non-periodicals as shown in Figure 10. Like TV, the periodic nature of periodicals leads to the higher temporal locality of E-book. Second, the average lifetime (i.e., the duration during which at least one seed is alive) of an E-book torrent is around 8∼9 days, which is shorter than that of a torrent in other categories (around 11∼12 days). The shorter lifetimes of E-book swarms are likely to result in the high temporal locality.

Publisher Types: As shown in Figure 11, torrents published by fake publishers exhibit higher temporal locality than others. This is because the administrators of TPB remove fake publishers' accounts and torrents when they are reported as fake ones, which results in shorter lifetimes of their torrents/swarms. Therefore, fake torrents can only be downloaded before they are removed, which results in higher temporal locality. The average lifetime of torrents of fake publishers (9.72 days) is shorter than those of torrents of profit-driven publishers (11.47 days) and altruistic publishers (10.63 days).

Popularity: We investigate the correlation between the total number of downloaders and daily locality by calculating the Pearson's coefficient. We found that there is a negative correlation (-0.30) between the number of downloaders and daily locality across content categories. Specifically, App (-0.43) and E-book (-0.43) show relatively strong negative correlation while TV (-0.24) shows weak negative one (TV: -0.24, Porn: -0.35, E-book: -0.43, Movie: -0.31, Music: -0.38, App: -0.43, and Game: -0.35).

To understand the disparity of correlation across the seven content categories, we analyze two contrasting content categories: TV (-0.24) and App (-0.43). As shown in Figure 12, TV shows a weaker negative correlation compared to the other content categories. This is because a significant portion of TV torrents have a large number of downloaders with high temporal locality. That is why the Pearson coefficient of TV torrents is low, so we take a deeper look at TV contents by accessing the air dates from Internet Movie Database (IMDB) [35] manually and found that 94% of torrents are periodical programs (e.g., TV shows or dramas), which 61.2% of users (average 858.9 users) have downloaded within 48 hours after it having been aired. Due to the time-sensitive nature of periodic TV torrents, a majority of people download popular TV torrents early after their air dates, which results in high popularity and daily locality. Consequently, it makes weak negative correlation for TV torrents.

Interestingly, App torrents show a strong negative correlation shown in Figure 12. We compare two torrent groups in App: (i) torrents with relatively high daily locality (> 0.2) and a small number of downloaders (< 100) and (ii) torrents with low daily locality (< 0.1) and a large number of downloaders (> 1000). Contradicting the common belief, the torrents of the first group mostly correspond to popular software (e.g., Microsoft Windows, Winzip, or Microsoft Office), and we find that lifetimes of the popular software torrents are relatively short (6 to 10 days in our datasets). To understand the phenomena, we take a look for all the app contents and we find that this is because many publishers upload these popular torrents of the same software frequently which makes users to be more likely to download a recent torrent. For example, 9 torrents of the same software 'Windows 7' are uploaded from April 6 to 10, 2011. On the contrary, the torrents of the second group correspond to software for special customer base (e.g., CAD or graphic tools). We find that the lifetimes of the special-purpose software torrents are substantially longer (> 20 days) than those of other torrents since they are not uploaded frequently, which makes users download their torrents steadily. For example, when we look at 'AutoDesk ECSCAD' from April 6 to May 9, 2011, there is only one torrent. We conclude that not only the time-sensitive nature but also the number of customers of content (i.e., general-/special-purpose) affects temporal locality.



(a) Average number of hourly seeds for United States

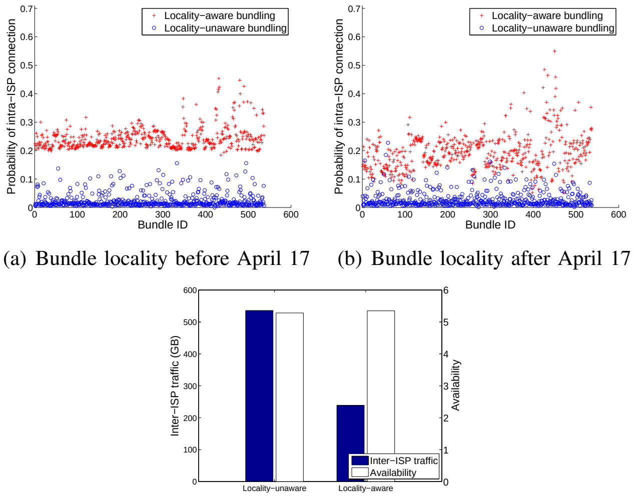
(b) Average number of hourly leechers for United States

(c) Average daily peak-to-trough ratio of hourly peers (seeds and leechers)

Fig. 13. Distributions of the number of seeds and leechers, and average daily peak-to-trough ratio of hourly peers consuming the content across the categories in United States in 2011 are plotted. Vertical grid lines in (a) and (b) correspond to midnights in its local time.

C. *Temporal Usage Trends*

In Figure 13, we plot temporal changes of the numbers of seeds and leechers over the week in United States, which ranked first in terms of the number of users in Table II. Moreover, to better illustrate changes of diurnal patterns, we plot the average daily peak-to-trough ratio (i.e., the number of peers at peak devided by that of trough.) in Figure 13(c). We observe a significant diurnal pattern with peak usage in the



(a) Bundle locality before April 17

(b) Bundle locality after April 17

(c) Inter-ISP traffic and availability

Fig. 14. Total inter-ISP traffic is significantly reduced (50%) in locality-aware bundling without degrading the availability.

late evening, which have been already reported in prior work like [12]. In addition, we can find the following interesting patterns. First, the peak-to-trough ratio is relatively higher in weekdays than weekends. We believe this is because people mostly work in a weekday, thus they often download torrents after coming from work (say, 7 pm). Second, the peak-to-trough ratio of leechers are higher than that of seeds. We believe that this because (i) BitTorrent clients tend to keep seeding after completion of downloading by default and (ii) profit-driven publishers tend to keep seeding for their financial gains [36]. Third, we find that both of the Movie and TV torrents have low peak-to-trough of seeds, but the different patterns are observed in leechers. In other words, TV torrents show the relatively higher peak-to-trough ratio of leechers, but Movie torrents exhibit the low peak-to-trough ratio of leechers. We conjecture that the low peak-to-trough ratio of seeds in Movie and TV is due to high popularity (as shown in Figure 13). However, because of time-sensitive nature of TV as shown in Section IV-C, many users tend to download soon after live broadcast, which results in higher peak-to-trough ratio than Movie torrents.

VI. How to Exploit Locality

Let us illustrate two use cases of exploiting locality: (i) bundling for both improving content availability and decreasing inter-ISP traffic, and (ii) caching for networking efficiency.

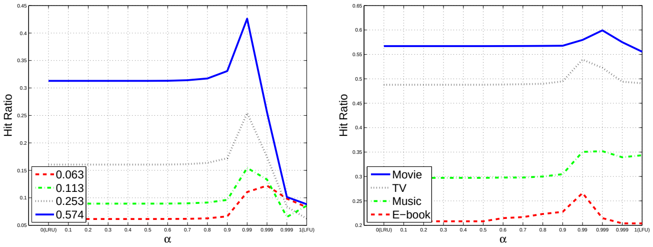
Bundling: Bundling torrents in BitTorrent has gained attention [36] since it can mitigate the unavailability problem [37] as well as reduce download times [38]. In bundling, two or more files are disseminated via a single torrent. However, prior bundling approaches may result in substantial inter-ISP traffic because a bundled torrent brings increased file size and swarm-size (i.e., users). However bundling contents mainly consumed in the same region (i.e., ISP) can reduce inter-ISP traffic.

To confirm the aforementioned conjecture, we first make bundles based on data for 17 days from April 1 to April 17 with two different bundling approaches as shown in Figure 14(a): (i) bundling two contents whose locality is high and consumption regions are mainly same (locality-aware bundling) and (ii) bundling two contents randomly but considering availability (locality-unaware bundling) [37]. Then we calculate the spatial locality of (actually) bundled torrents for the following 15-days (from April 18 to May 2, 2011). Surprisingly, we observe that torrents bundled by a locality-aware strategy still exhibit high locality as shown in Figure 14(b), which signifies that we can exploit the nature that the spatial locality changes marginally over time. We next estimate the average content availability of bundled torrents and the inter-ISP traffic after April 17. Figure 14(c) shows that total inter-ISP traffic is significantly reduced (50%) in locality-aware bundling without degrading the availability. From this, we can conclude that locality need to be seriously considered in bundling.

Caching: Recently, information-centric networking (ICN) [39] has gained momentum, and seeks to achieve efficient content distribution. One of the key components in ICN is in-network caching, where how to select content to be cached (and to be replaced) is critical. To investigate how the temporal locality affects in-network caching performance, we conduct a simulation study by generating four request patterns (each consisting of 100,000 requests) with different temporal locality (i.e., 0.063, 0.113, 0.253, and 0.574) under the same Zipf-like distribution (β=0.63 [40]). We evaluate a caching strategy, *LRFU* [41], which reflects both of the *recency* and *frequency* and its reference count of each cached element is decayed by multiplying α periodically [41]. If we set $\alpha = 0$, LRFU operates like LRU because it reflects only the current reference counts, while if we set $\alpha = 1$, LRFU is reduced to LFU because the past reference counts do not decay. We vary α from 0 to 1 and examine the relation between the temporal locality and the caching performance. As shown in Figure 15(a), the cache hit ratio with high temporal locality is higher than the one with low temporal locality since the request pattern becomes more bursty as the temporal locality is higher. Therefore, we can conclude that caching performance is affected by the temporal locality. Interestingly, the point at which the hit ratio is highest is changes depending on α, which implies that lower α fits for high temporal locality torrents (e.g., TV), while higher α fits for low temporal locality torrents (e.g., Movie).

We also conduct a simulation for the torrents of four categories from real traces[4]: (i) 'TV' with high locality, (ii) 'E-book' with high locality, (iii) 'Movie' with low locality, and (iv) 'Music' with moderate locality as shown in Figure 5(a). As shown in Figure 15(b), the cache hit ratio from the real traces is in line with that of synthetic requests. The highest hit ratio in TV/E-book category is achieved at α = 0.99, while the highest hit ratio in Movie category is α = 0.999 since its locality is relatively low. Since the locality of Music torrents is moderate, the highest point of hit ratio lies in-between. Through this evaluation, we can notice that the

[4] 147,286 content requests from 'AS0920 National Internet Backbone'



(a) Effect of different temporal locality with synthesized requests (b) Effect of different categories with requests from real trace

Fig. 15. Caching performance (Hit-Ratio) is affected by the temporal locality.

caching performance depending on content type is different resulting from different localities even in the same request distribution, which implies the importance for the adaption of temporal locality to cache replacement algorithm such as GreedyDual [42].

VII. Conclusion

We conducted comprehensive measurements on content locality of BitTorrent. From the datasets, we analyzed: (1) how content is consumed in a spatially and temporally skewed way, (2) what makes the content be consumed differently in spatial and temporal domains depending on its properties, (3) how content consuming patterns changes over the years, (4) how we can exploit the content locality. We observed that content consumption pattern is biased in both spatial and temporal domain. We also revealed how cultural factors (e.g., language) affect the spatial locality and how publishing purpose or time-sensitivity also affects the temporal locality of content. We further found that content sharing patterns are increasingly globalized and the diurnal pattern of content usage in leechers is more pronounced than that of seeds. From these observations, we also demonstrated that how spatial and temporal locality can be exploited for bundling torrents and in-network caching.

VIII. Acknowledgements

This research was supported by the KCC (Korea Communications Commission), Korea, under the R&D program supervised by the KCA (Korea Communications Agency) (KCA-2013-11-911-05-002) and International Collaborative R&D Program of Korea Institute for Advancement of Technology. The ICT at Seoul National University provides research facilities for this study.

References

[1] Sandvine: Global Internet Phenomena (Fall 2012).

[2] Karagiannis and *et al.*, "Should internet service providers fear peer-assisted content distribution?" in *ACM IMC*, 2005.

[3] S. Seetharaman and *et al.*, "Characterizing and mitigating inter-domain policy violations in overlay routes," in *ICNP*, 2006.

[4] A. R. Bharambe and *et al.*, "Analyzing and improving a bittorrent networks performance mechanisms," in *IEEE INFOCOM*, 2006.

[5] R. Keralapura and *et al.*, "Can ISPs Take the Heat from Overlay Networks?" in *ACM HotNets*, 2004.

[6] L. Qiu and *et al.*, "On selfish routing in internet-like environments," in *ACM SIGCOMM*, 2003.

[7] D. R. Choffnes and *et al.*, "Taming the torrent: a practical approach to reducing cross-isp traffic in peer-to-peer systems," in *ACM SIGCOMM*, 2008.

[8] Peterson and *et al.*, "Antfarm: efficient content distribution with managed swarms," in *USENIX NSDI*, 2009.

[9] J. Seedorf and *et al.*, "Traffic localization for p2p-applications: The alto approach." in *IEEE P2P*, 2009.

[10] Xie and *et al.*, "P4p: provider portal for applications," in *ACM SIGCOMM*, 2008.

[11] M. Dischinger and *et al.*, "Detecting bittorrent blocking," in ACM IMC, 2008.

[12] J. S. Otto and *et al.*, "On blind mice and the elephant: understanding the network impact of a large distributed system," in *ACM SIGCOMM*, 2011.

[13] Kryczka and *et al.*, "Unrevealing the structure of live bittorrent swarms: Methodology and analysis." in *IEEE P2P*, 2011.

[14] R. C. Rumín and *et al.*, "Deep diving into bittorrent locality," in *IEEE INFOCOM*, 2011.

[15] "The pirate bay," http://www.thepiratebay.com.

[16] Legout and *et al.*, "Clustering and sharing incentives in bittorrent systems," in *ACM SIGMETRICS*, 2007.

[17] D. R. Choffnes and *et al.*, "Strange bedfellows: community identification in bittorrent," in *IPTPS*, 2010.

[18] C. L. Viles and J. French, "Content locality in distributed digital libraies," *Information Processing & Management*, vol. 35, 1999.

[19] C. L. Viles, "Content locality in time-ordered document collections," 1999.

[20] A. Brodersen and e. Scellato, "Youtube around the world: geographic popularity of videos," in *ACM WWW*, 2012.

[21] e. Roberto Gonzalez, "Where are my followers? understanding the locality effect in twitter," *CoRR*, vol. abs/1105.3682, 2011.

[22] M. P. Wittie and e. Pejovic, "Exploiting locality of interest in online social networks," in *ACM CoNEXT*, 2010.

[23] J. M. Pujol and e. Erramilli, "The little engine(s) that could: scaling online social networks," in *ACM SIGCOMM*, 2010.

[24] "Open sourced bittorrent client: vuze," http://www.vuze.com.

[25] D. Wu and *et al.*, "Understanding peer exchange in bittorrent systems," in *IEEE P2P*, 2010.

[26] "Bittorrent peer exchange conventions," http://wiki.theory.org/BitTorrentPeerExchangeConventions.

[27] R. C. Rumín and *et al.*, "Is content publishing in bittorrent altruistic or profit-driven?" in *ACM CoNEXT*, 2010.

[28] S. Kim and *et al.*, "Content publishing and downloading practice in bittorrent," in *IFIP Networking*, 2012.

[29] "Maxmind," http://www.maxmind.com/.

[30] C. Zhang, P. Dhungel, and *et al.*, "Unraveling the bittorrent ecosystem," *IEEE Transactions on Parallel and Distributed Systems*, vol. 22, pp. 1164–1177, 2011.

[31] G. Maier, A. Feldmann, and P. *et al.*, "On dominant characteristics of residential broadband internet traffic," in *ACM IMC*, 2009.

[32] V. D. Blondel and *et al.*, "Fast unfolding of communities in large networks," *Journal of Statistical Mechanics: Theory and Experiment*, vol. 2008, no. 10, 2008.

[33] H. Kwak and *et al.*, "Mining communities in networks: a solution for consistency and its evaluation," in *ACM IMC*, 2009.

[34] M. Izal and e. Urvoy Keller, "Dissecting BitTorrent: five months in a torrent's lifetime," in *PAM*, 2004.

[35] "Internet movie database," http://www.imbdb.com.

[36] J. Han and *et al.*, "Bundling practice in bittorrent: What, how, and why," in *ACM SIGMETRICS*, 2012.

[37] D. S. Menasche and *et al.*, "Content availability and bundling in swarming systems," in *ACM CoNEXT*, 2009.

[38] S. Zhang and *et al.*, "Dynamic file bundling for large-scale content distribution," in *IEEE LCN*, 2012.

[39] A. Ghodsi and *et al.*, "Information-centric networking: seeing the forest for the trees," in *ACM HotNets*, 2011.

[40] A. Abhari and *et al.*, "Workload generation for youtube," *Multimedia Tools Application*, vol. 46, no. 1, 2010.

[41] D. Lee and *et al.*, "Lrfu: A spectrum of policies that subsumes the least recently used and least frequently used policies," in *ACM SIGMETRICS*, 2001.

[42] S. Jin and *et al.*, "Greedydual* web caching algorithm," *Computer Communications*, 2001.