Order-of-Magnitude Popularity Estimation of Pirated Content

Charalampos Chelmis and Daphney-Stavroula Zois

Abstract Understanding the spread of information in complex networks is a key problem. Content sharing in popular online social networks such as Facebook and Twitter has been well studied, however, the future trajectory of a cascade has been shown to be inherently unpredictable. Nonetheless, cascade virality has recently been studied as a classification problem, resulting in good prediction accuracy. Herein, we address the important problem of pirated media popularity estimation in torrent applications, such as Project Free TV, Popcorn-Time, and The Pirate Bay. Although pirating software and media is illegal, the practice of pirating is actually growing in popularity. On a large sample of data acquired from The Pirate Bay, we demonstrate high accuracy in the task of identifying whether the popularity of a torrent will continue to grow in the future. Specifically, we achieve close to perfect accuracy in estimating the order-of-magnitude popularity of torrents.

1 Introduction

Video popularity estimation is an important problem [2, 6], particularly for the movies industry due to the massive economic losses of copyrighted content infringement through digital piracy. Even though earlier studies had reported that peer-to-peer traffic was declining due to copyright laws, more recent work has argued that peer-to-peer traffic is in fact increasing and still constitutes a significant fraction of the total workload transmitted in the Web [17, 24]. In fact, according to Columbia University's American Assembly's "Copy Culture" study on copyright infringement in the United States and Germany [16], it was revealed that 45% of US citizens and 46% of German citizens actively pirate media.

C. Chelmis (✉) · D.-S. Zois
University at Albany, State University of New York, Albany, NY, USA
e-mail: cchelmis@albany.edu; dzois@albany.edu

T. Özyer, R. Alhajj (eds.), *Machine Learning Techniques for Online Social Networks*,
Lecture Notes in Social Networks, https://doi.org/10.1007/978-3-319-89932-9_5

Because of its economic importance, the problem of accurate estimation of the potential popularity of a video has been extensively studied for years, with emphasis on the design and management of recommendation systems and targeted advertising services [22]. However, most of the related work is not applicable to the problem of pirated content popularity. The reason is that, as we show in this article, popularity of pirated content is related to a variety of features through a heavy-tailed distribution, which leads to a severe imbalance in this problem. In addition, when a media delivery system, such as Popcorn Time, is used that can hide clients' consumption even from the content distributor by means of cryptographic primitives, the actual network topology of the peer network or the cascade of the copyright infringed content is usually unavailable.

We propose an approach to estimate the order of magnitude of pirated content popularity based on a small set of publicly available metadata associated with the actual content; the estimation does not utilize actual content information. In contrast to other work which focuses only on one type of content, i.e., video or microblogs in online social media (e.g., [3, 6]), we obtained estimation of order-of-magnitude increases in the size of numerous types of pirated content. In our experimental evaluation, we maintained the imbalances of the real-world dataset to better mimic reality. This differs from some previous studies on cascade prediction, which balance the data before conducting prediction or classification [12]. Finally, using torrent features, we estimate torrent "quality."

The overarching goal of our work is to understand fundamental properties of popularity of torrents used to share pirated content. Defining popularity itself in this context is not straightforward; proxies of popularity can be defined as the number of seeders or leechers of a torrent, or based on the number of votes from users that have already downloaded the pirated content. We believe that an in-depth study of torrent popularity as a function of a set of features is necessary to understand the reasons people take the risk of being prosecuted while illegally downloading and sharing pirated media. Our work solves an important problem and paves the way towards a more comprehensive study of pirating media with applications to digital forensics for the movies and audio industry.

The main contributions of this work are as follows:

1. Through an extensive empirical study of a large-scale real-world torrent dataset acquired from The Pirate Bay, we provide novel insights regarding torrent characteristics and we identify correlations between torrent features.
2. We construct a vector space model of pirated media, which we use to estimate the exact number of seeders.
3. We demonstrate that our approach can achieve near optimal estimation of torrent popularity despite the severe data imbalance and the skewed distribution of torrent features on a dataset of 679,515 unique torrents from The Pirate Bay.

2 Anatomy of Pirated Content

2.1 Background

First, we introduce some necessary terminology that we use in later sections to describe, measure, and understand piracy data. Peer-to-Peer networks have become extremely popular in content sharing by facilitating the exchange of file transfers between users in a decentralized manner. In order to download a file, users must connect to others who are providing the file for download. Trackers keep track of users (e.g., IP address and port) that are downloading and uploading files, so that they can connect to each other using torrents (small file with metadata describing contained files). The Pirate Bay (hereafter referred to as TPB) used to be one of the most popular torrent hosting sites, where anyone could download torrent files. In 2009, the website was brought down on account for copyright infringement [11, 21]. Subsequently, proxies have been providing access to it and its content by multiple servers, collectively nicknamed the "TPB hydra" [19].

In order for peers to experience fast downloads, they have to discover high quality torrents, which further have many seeders uploading desired files. Inherently, torrent users have to rely on a dynamically changing number of seeders, i.e., peers that own a complete version of the desired file, and user feedback in order to decide whether to download a torrent or not. The unique characteristic of torrents is that files are broken into pieces, which constantly shift between users. The advantage of this is that when one person leaves the network, the data is transferred to another person so that it is always available. Conversely, leechers are users who are downloading data, and as such only own parts of a file. "Leecher" and "peer" are commonly used interchangeably.

2.2 TPB Torrent Description

A typical torrent description page is shown in Fig. 1. It provides a variety of information including the torrent title, its type, who uploaded it and when, and its quality based on user feedback. Clicking on the torrent type brings one to a page listing torrents of the same category. Clicking on the username of the person who uploaded the torrent results in a listing of all the torrents uploaded by that user. Finally, clicking on a tag brings one to a list of torrents associated with this tag.

The number of seeders and leechers is also available. The number of seeders is extremely important, since without seeders, only parts of the file are available for download, resulting in an incomplete, non-usable file. Registered users are further able to upload torrent files, write comments and leave personal messages. Inherently, TPB (as well as other BitTorent sites) does not provide a mechanism for promoting the k-most "interesting" torrents nor does it provide a ranking of torrents based on their "quality."

Fig. 1 A typical torrent page on The Pirate Bay

2.3 Data Set

We obtained a dataset D_I, containing metadata (ID, category, size, number of leechers, number of seeders) of 679,515 unique torrents available at TPB on December 5, 2008 [7]. We enriched D_I by crawling TPB hyperlinks (of the form http://thepiratebay.org/torrent/torrentId) during October 2011 to harvest detailed information about each torrent. 200,950 torrents out of 679,515 in D_I did not exist at that time. Our enriched dataset, D_{TPB}, contains 478,565 unique TPB torrents, represented as tuples of the following form $< t_{id}, tp, sz, n_f, n_t, n_c, Q, n_{os}, n_{ol}, n_{cs}, n_{cl} >$, where $t_{id} =$ torrent id, $tp =$ category, $sz =$ torrent size in MBs, $n_f =$ number of files, $n_t =$ number of associated tags, $c =$ number of comments, $Q =< n_p, n_n, n_a >$ is a quality vector, where $n_p =$ number of positive votes, $n_n =$ number of negative votes, $n_a =$ average number of votes, $n_{os} =$ number of seeders in D_I, $n_{ol} =$ number of leechers in D_I, $n_{cs} =$ number of seeders in D_{TPB}, and $n_{cl} =$ number of leechers in D_{TPB}. Unless otherwise specified, our analysis refers to D_{TPB}.

2.4 Data Characteristics

Figures 2, 3, and 4 present the distribution of torrents with respect to their characteristics, whereas Figs. 5, 6, and 7 show various features associated with each torrent. All features follow a heavy-tailed distribution. All distributions are broad,

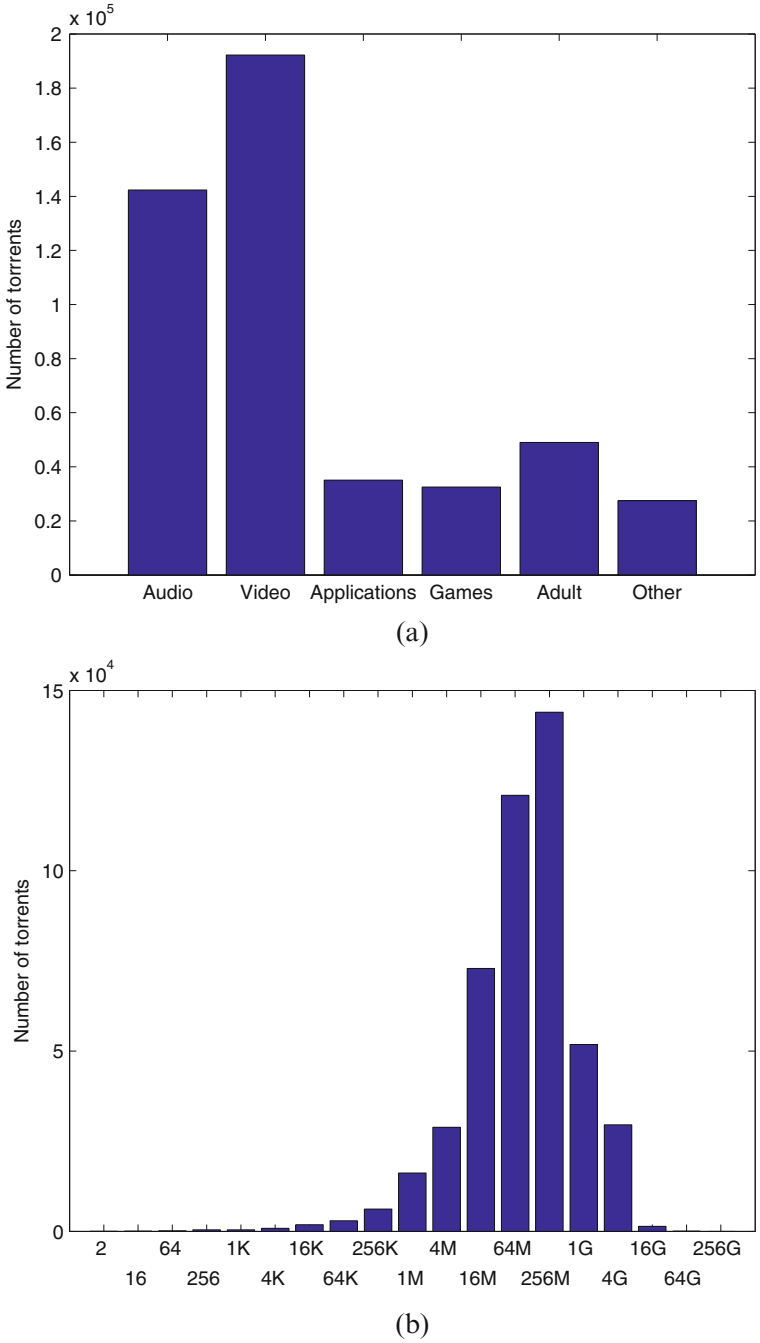


Fig. 2 Torrent distribution per (**a**) category and (**b**) size

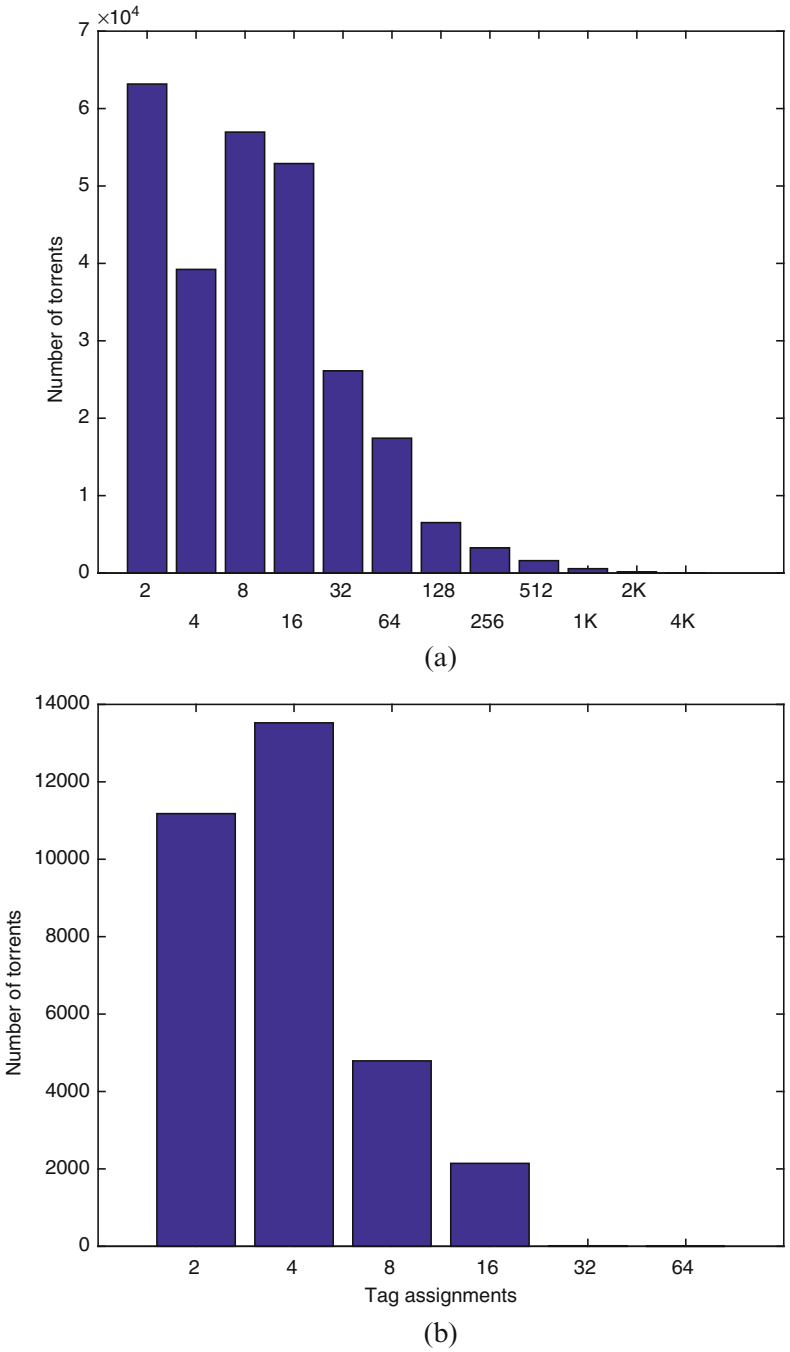
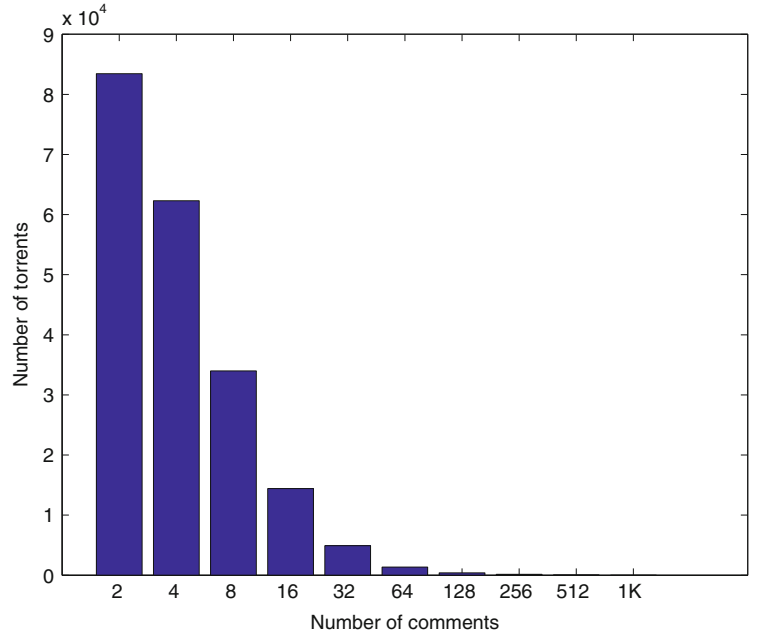
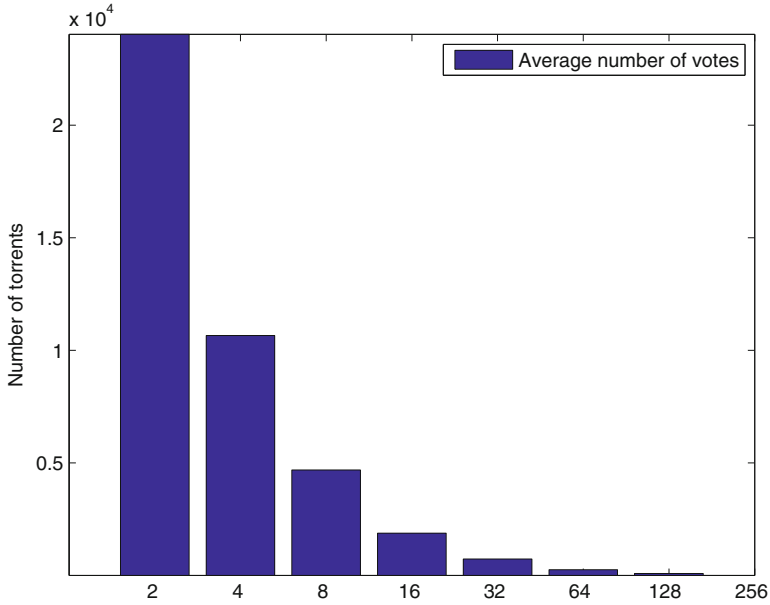


Fig. 3 Torrent distribution per (**a**) number of files, and (**b**) tag assignments



(a)



(b)

Fig. 4 Torrent distribution per (**a**) number of comments, and (**b**) average number of votes

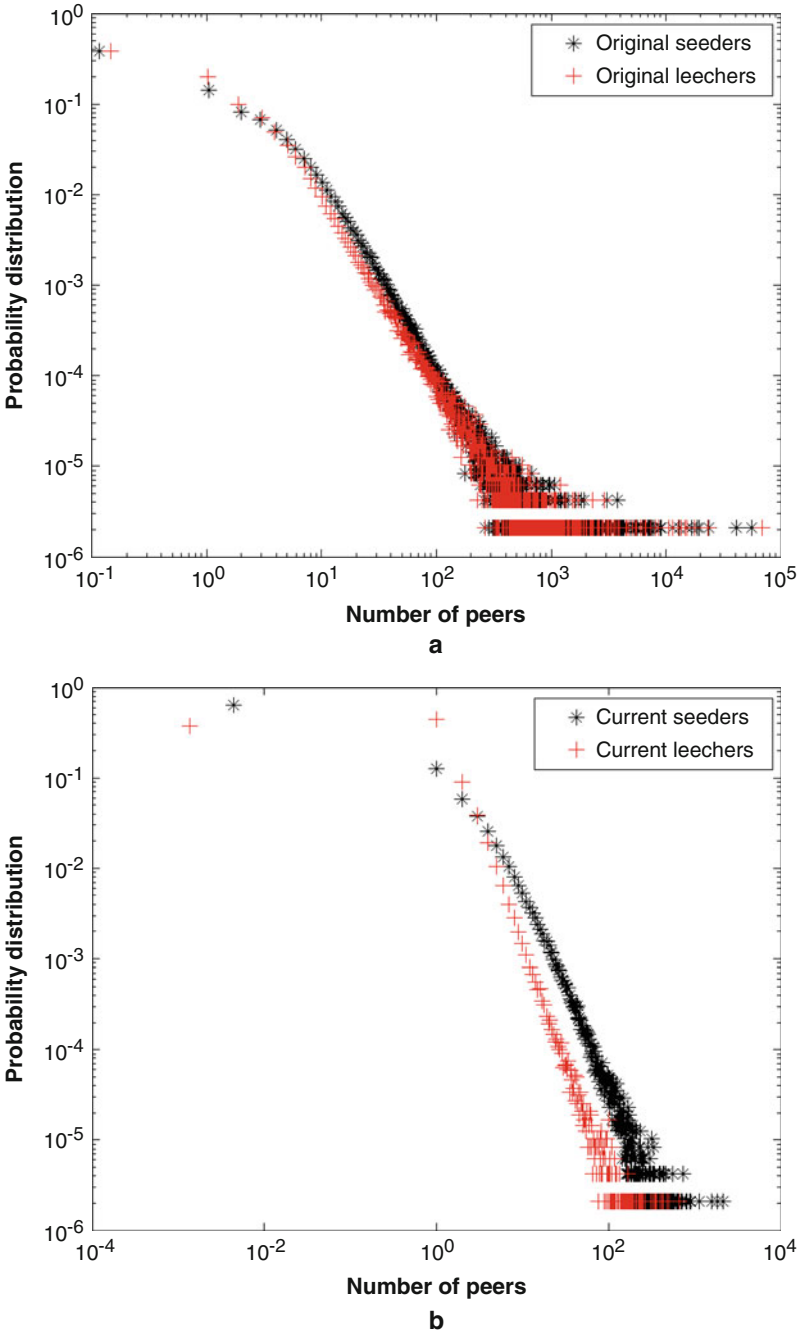


Fig. 5 Probability distributions of the number (**a**) n_{os} and n_{ol} of original seeders and leechers per torrent, and (**b**) n_{cs} and n_{cl} of current seeders and leechers per torrent

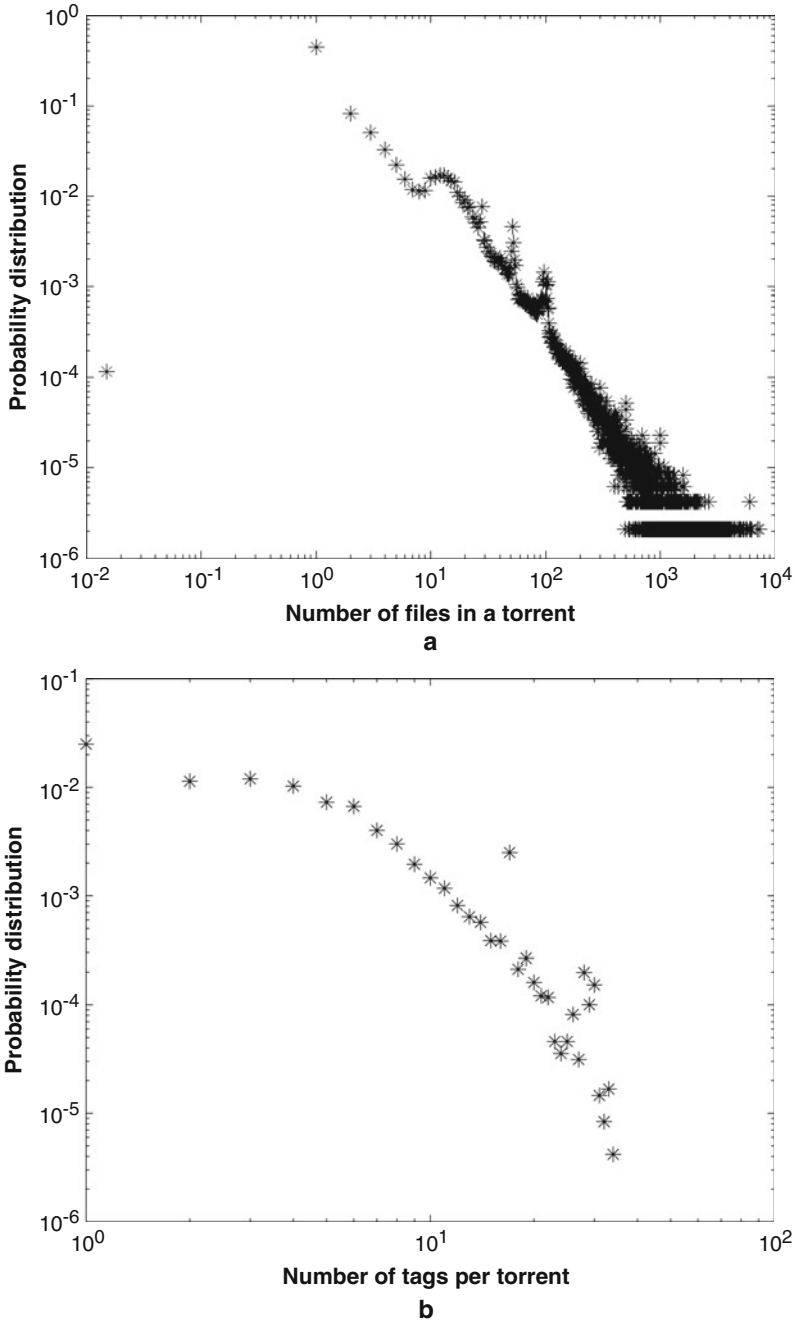


Fig. 6 Probability distributions of the number (**a**) n_f of files per torrent, and (**b**) n_t tag assignments per torrent

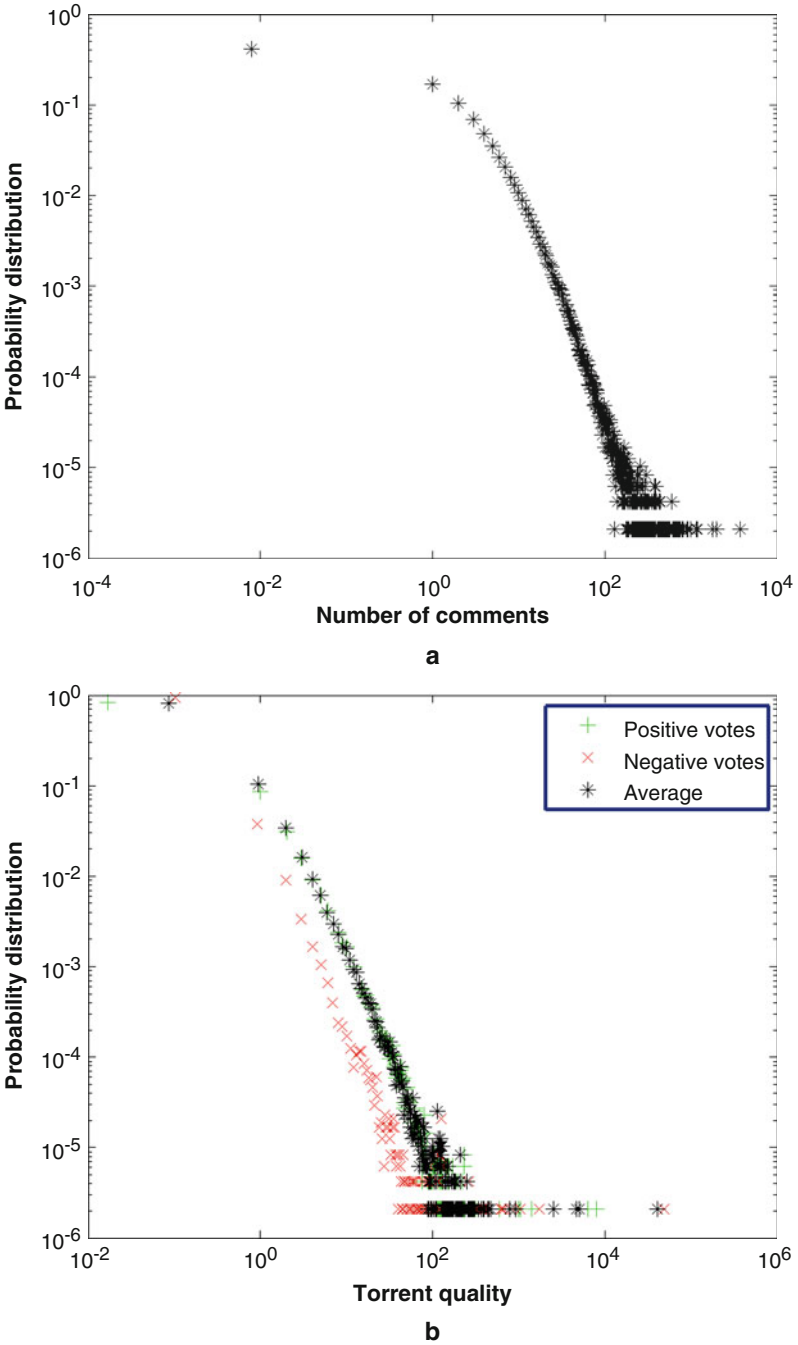


Fig. 7 Probability distributions of the number (**a**) n_c comments per torrent, and (**b**) n_p positive votes, n_n negative votes, and n_a average votes per torrent

Table 1 Averages and fluctuations of torrent features

Measure x	Average (x)	Variance (x)
n_{os}	8.32	19,767
n_{ol}	6.59	27,708
n_{cs}	2.36	164.34
n_{cl}	1.19	15.51
n_{sz}	863.16	588.41
n_f	20.74	9515.9
n_t	0.42	3.59
n_c	3.66	217.17
n_p	0.7	302.23
n_n	0.27	5128.5
n_a	0.8	3702

showing that these features are highly heterogeneous. For reference, Table 1 reports the averages and variances of these quantities.

A few comments are in order. First, the number of seeders/leechers follow similar distributions. Figure 8a shows the scatter plot (loglog) of the number of seeders versus the number of leechers. The points are close to the diagonal, indicating approximately linear relation between the number of seeders and leechers. The result is similar for the number of seeders versus the number of leechers in D_I. We further examined the correlation between positive and average votes, the scatterplot of which (Fig. 8b) demonstrates that a linear relationship between the two measures does exist. Secondly, the average number of tags per torrent is small, limiting tag-based search of torrents. Similarly, the average number of comments per torrent is small, which indicates a tendency of people not to comment on torrents, either to support their quality or indicate bad quality torrents (e.g., fakes) to other users. This conclusion is further supported by the average number of votes (n_p, n_n, n_a). The variability of negative votes is quite large however, indicating that many users vote negatively for bad torrents but do not necessarily support good torrents.

2.5 Correlations

2.5.1 Correlations w.r.t. Torrent Popularity

Do torrents with more seeders or leechers also have more files and tags, are they more commented, and do they attract more votes? Figures 9 and 10 show that this is indeed the case. Figure 9 displays the average features values of torrents with k current seeders. Figure 10 shows average features values as a function of torrent leechers. The results are similar for the number of seeders/leechers in D_I. Next, we characterize average attribute value n_p of torrents with k seeders

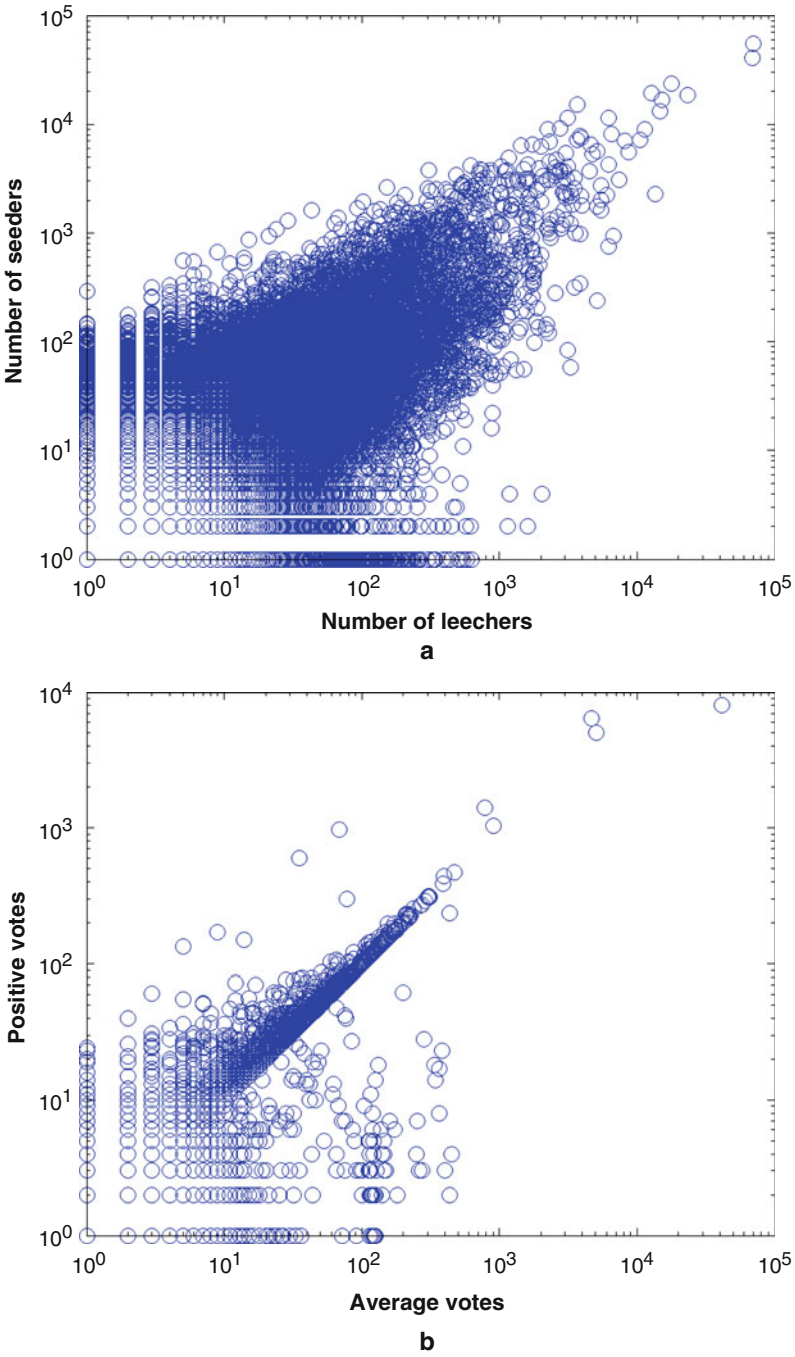


Fig. 8 Loglog scale scatter plot of (**a**) the number of seeders and the number of leechers, and (**b**) the number of positive and average votes

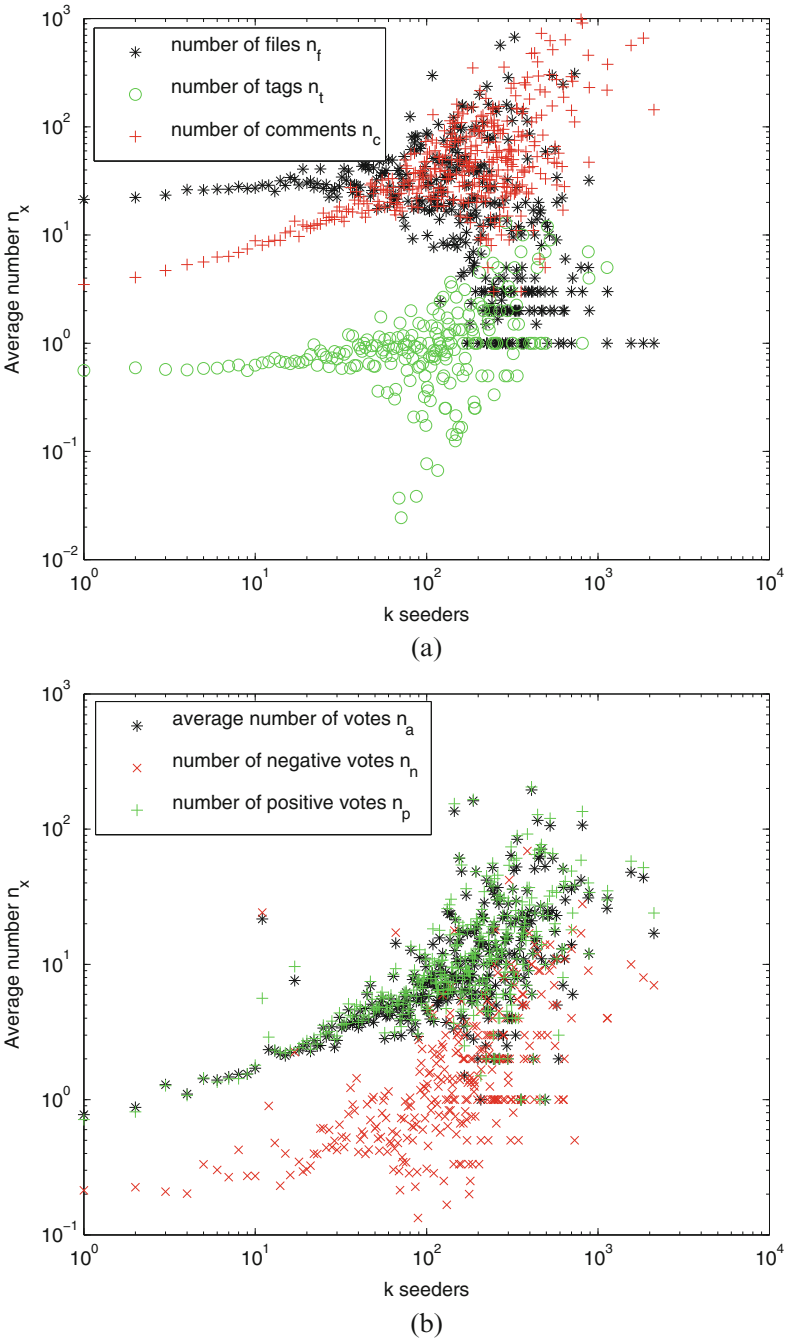


Fig. 9 Average number of (**a**) files (n_f), tags (n_t) and comments (n_c), and (**b**) quality (n_p, n_n, n_a) as a function of seeders n_{cs}

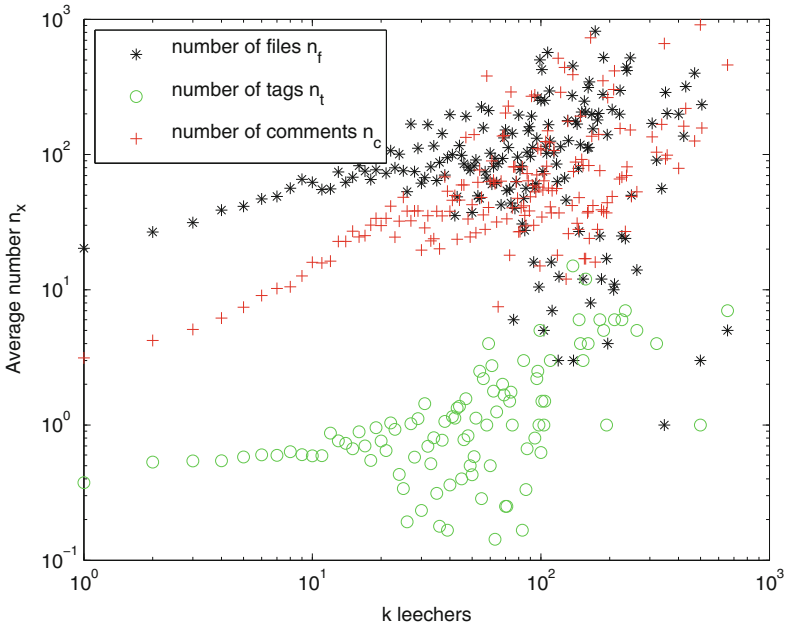


Fig. 10 Average number of (**a**) files (n_f), tags (n_t) and comments (n_c), and (**b**) quality (n_p, n_n, n_a) as a function of leechers n_{cl}

Table 2 Pearson correlation coefficients w.r.t. torrent popularity

Feature i	Popularity k	Pearson correlation
Size	Seeders	0.182
	Leechers	0.417
Number of files	Seeders	−0.055
	Leechers	0.246
Number of tags	Seeders	0.049
	Leechers	0.133
Number of comments	Seeders	0.569
	Leechers	0.469
Positive votes	Seeders	0.378
	Leechers	0.456
Negative votes	Seeders	0.279
	Leechers	0.289
Average votes	Seeders	0.343
	Leechers	0.394

(similarly for leechers) as a weighted average of respective values. For instance, $n_p(k) = \frac{1}{|t:t_s=k|} \sum_{t:t_s=k} n_p(t)$.

All features have an increasing trend for increasing values of k, and of course all features exhibit strong fluctuations for all values of k. The strong fluctuations visible for large k values are due to the fewer seeded torrents over which the averages are performed. Notably, torrents with a large number of seeders/leechers but having very few files, tags and receiving few comments and votes can be observed. Despite these important heterogeneities in the features of torrents with the same seeders/leechers k, the data clearly indicate a strong correlation between the different types of features up to about $\frac{10^3}{2}$ seeders/leechers. The disparity of measurements after this point however clearly decreases the value of the Pearson correlation coefficients overall, for almost all pairs of features. For reference, Table 2 reports these quantities.

Discriminative features exhibiting increasing trends for the number of seeders also show such trends for the number of leechers. This is not the case however when considering torrent size. Figure 11 shows average torrent size as a function of torrent popularity. The trend is non-increasing for the number of seeders but is definitely increasing as a function of leechers both in D_I and D_{TPB}.

2.5.2 Correlations w.r.t. Torrent Quality

Similarly, it seems natural to ask whether different types of features are correlated with one another, with respect to torrent quality: do torrents with more positive (also negative or average) number of votes also have more files and tags, are they more commented, and do they have more seeders (also leechers)? As shown in Fig. 12, this does not seem to be the case for the number of files contained in torrents,

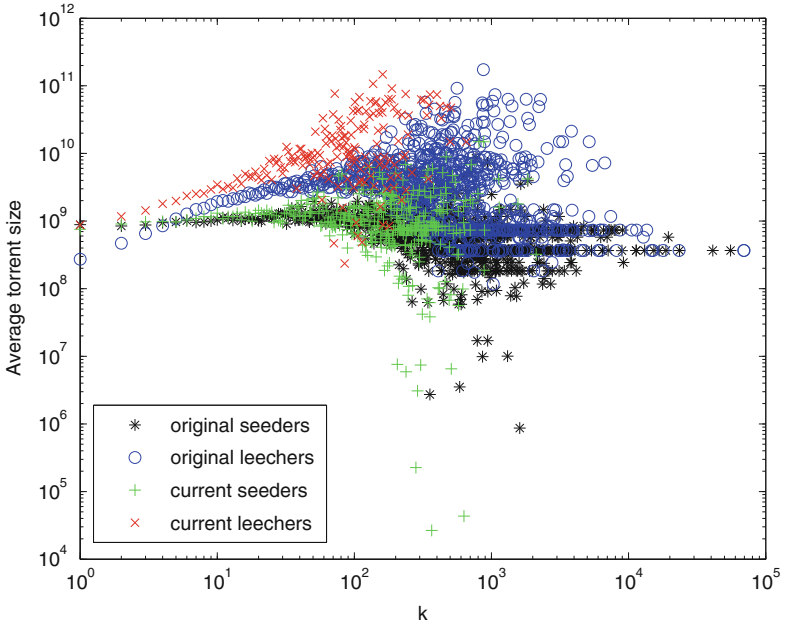


Fig. 11 Average torrent size as a function of torrent popularity

neither does this fact hold true for torrent size. On the contrary, Fig. 13 shows that there is indeed a positive relation between torrent quality and the number of tags and the number of comments. Table 3 reports the value of the Pearson correlation coefficients with respect to torrent quality. The correlation coefficients in this case are much smaller, and often the overall effect of disperse values results in negative averages. However, non-decreasing trends are apparent for all features up to $\approx 10^2$ number of votes.

2.6 Feature Alignment

We now focus on the similarity between torrents in relation to their popularity, as measured by the similarity of their respective feature vectors. This approach is similar to the exploration of topical locality in the Web, where the question is whether pages that are closer to each other in the link graph are more likely to be related to one another [20]. Similar results can be obtained for torrent quality.

We define a robust measure of feature similarity between two torrents u and v; we regard torrents as feature vectors whose elements correspond to different features and whose entries are the values for that specific features. To compare the feature

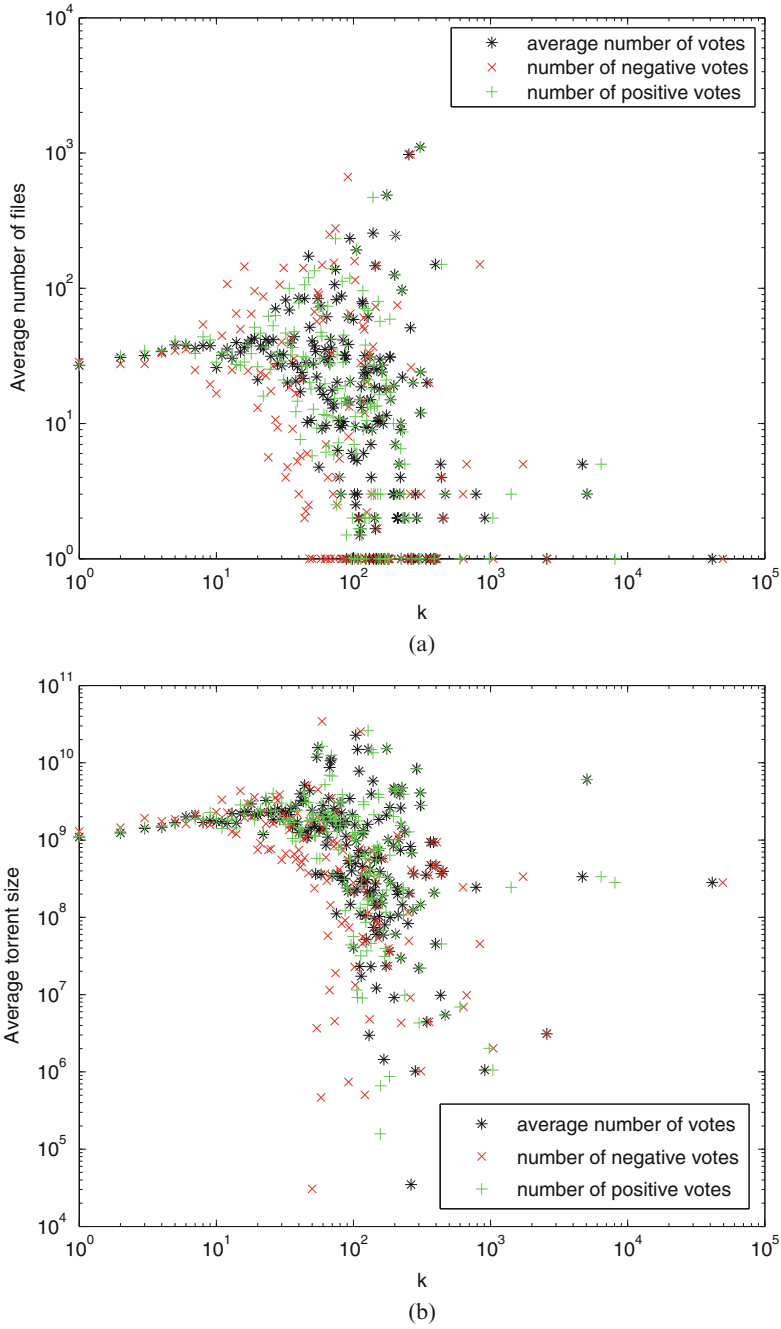
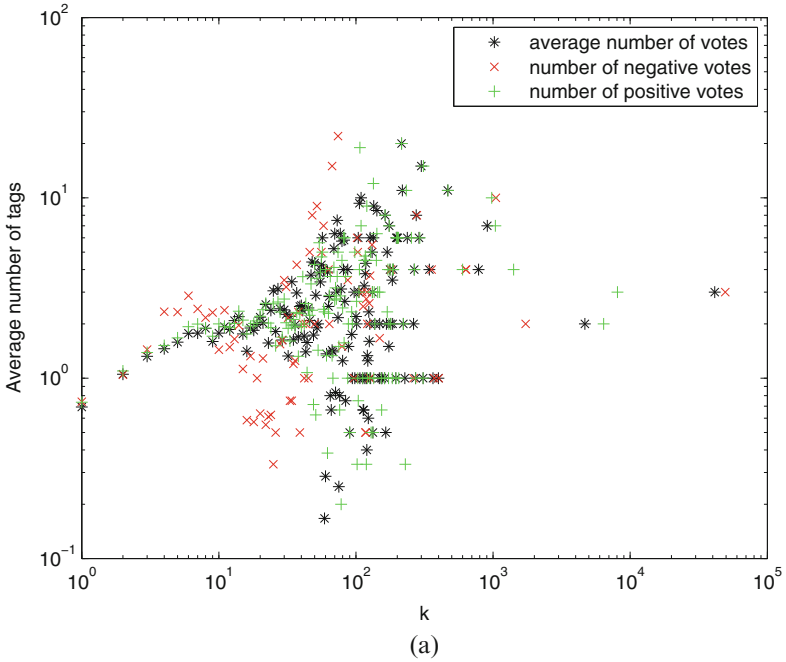
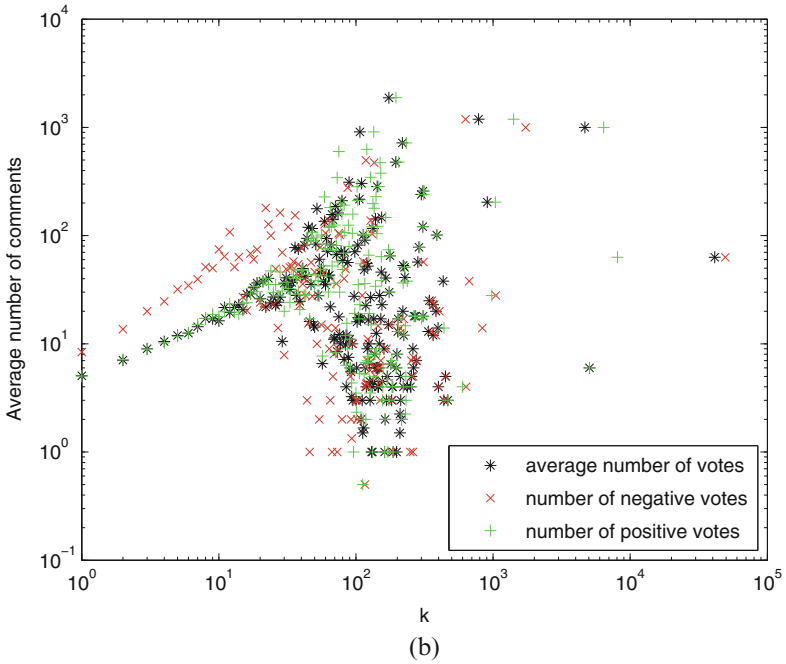


Fig. 12 Average number of (**a**) files (n_f), (**b**) size (n_{sz}) of torrents having k positive/negative/average number of votes



(a)



(b)

Fig. 13 Average number of (**a**) tags (n_t), and (**b**) comments (n_c) of torrents having k positive/negative/average number of votes

Table 3 Pearson correlation coefficients w.r.t. torrent quality

Feature i	Quality k	Correlation
Size	Positive votes	-0.032
	Negative votes	-0.033
	Average votes	-0.037
Number of files	Positive votes	-0.043
	Negative votes	-0.038
	Average votes	-0.029
Number of tags	Positive votes	0.016
	Negative votes	0.038
	Average votes	0.010
Number of comments	Positive votes	0.207
	Negative votes	0.023
	Average votes	0.040

vectors of two torrents, we use the standard cosine similarity. Denoting by $f_t(m)$ the value of feature m for torrent t, the cosine similarity $\sigma(u, v)$ is defined as

$$\sigma(u, v) = \frac{\sum_m f_u(m) f_v(m)}{\sqrt{\sum_m f_u(m)^2} \sqrt{\sum_m f_v(m)^2}}. \quad (1)$$

This quantity is 0 if u and v have no shared features, and 1 if they have exactly the same values for the same features, in the same relative proportions. To compute averages of $\sigma(u, v)$, we calculated the cosine similarity between torrents with k seeders, over all torrents in D_{TPB}. In Fig. 14 we show the density of similarity scores between pairs of torrents for different feature sets. We observe that the first two feature sets, namely (a) $[n_f, n_t, n_c]$ and (b) $[n_f, n_t, n_c, n_p, n_n, n_a]$ do not perform as well as the last two feature sets (c) $[n_c, n_p, n_n]$ and (d) $[n_p, n_n, n_a]$. These results confirm our intuition that torrents with similar numbers of positive/negative/average votes tend to have similar numbers of seeders. Similar results can be obtained for the number of leechers.

3 Torrent Popularity Estimation

The analysis in the previous section strongly suggests that torrents of high similarity, as captured by their corresponding feature vectors, are more likely to have the same number of seeders. Therefore a natural question to ask is "*whether similarity among torrents based solely on their respective features can be employed as accurate estimator of the number of seeders of a torrent*." We test this hypothesis on our D_{TPB} dataset as well as on D_{AVG}, a dataset derived from D_{TPB} by calculating average feature values for k number of seeders and constructing the corresponding

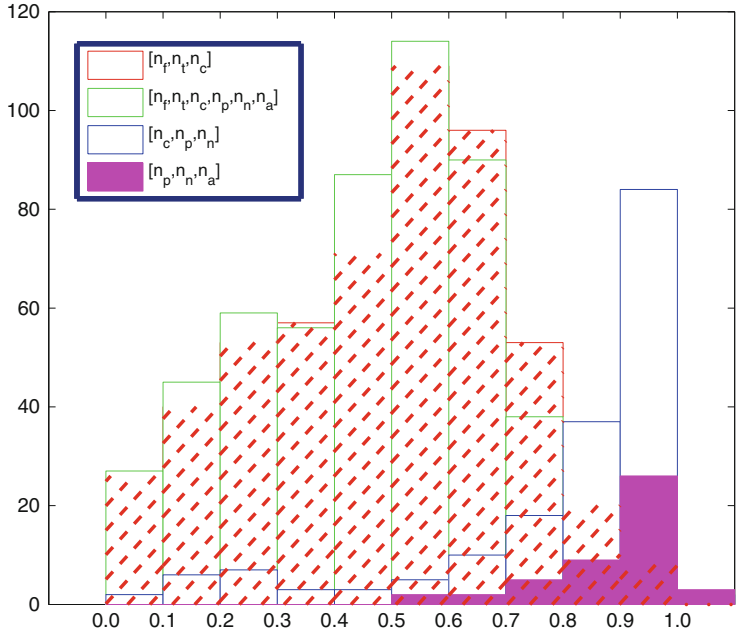


Fig. 14 Distribution of cosine similarity between torrents with same number of seeders. x-axis: similarity score; y-axis: estimated probability density

average feature vectors. The purpose of constructing D_{AVG} is to examine if we can achieve better estimation power by averaging feature values of torrents with the same number of seeders. For brevity, we report results only for the number of seeders, however the same approach can be used to estimate the quality of torrents as well.

Given a query torrent q_t, we compute the similarity between q_t and torrents in the training dataset using Eq. (1), and then select the k-most similar torrents $\{t_1, t_2, ..., t_k\}$, which we call *query neighborhood*. We estimate the number of seeders P_{q_t} for torrent q_t as a weighted average of the query neighborhood, as described in Algorithm 1. Formally,

$$P_{q_t} = \frac{\sum_{i=1}^{k} \sigma(q_t, t_i) * s_{t_i}}{\sum_{i=1}^{k} \sigma(q_t, t_i)}, \tag{2}$$

where s_{t_i} denotes the number of known seeders of torrent t_i. To determine the impact of query neighborhood size on the estimation quality, we performed an experiment where we varied the number of neighbors. We also examined the case of using a weighted sum of approximated values of known number of seeders s_{t_i}, based on a linear regression model, expressed as: $P'_{q_t} = \alpha s_{t_i} + \beta + \varepsilon$, where parameters α and β are determined by going over both feature vectors of the query and reference data,

Algorithm 1 Calculates the popularity of query torrent q_t

Input: m dimensional feature vector of query torrent q_t, D torrents with known number of seeders represented by their respective m dimensional feature vectors, and query neighborhood size k.
Output: Estimated number of seeders P_{q_t} for q_t.

1: D = total # of torrents, for which popularity is known.
2: **for** $(i = 0; i < D; i++)$ **do**
3: $\sigma(q_t, t_i) = \frac{\sum_m f_{q_t}(m) f_{t_i}(m)}{\sqrt{\sum_m f_{q_t}(m)^2}\sqrt{\sum_m f_{t_i}(m)^2}}$
4: **end for**
5: Rank torrents based on σ.
6: Choose top k torrents $\{t_1, t_2, ..., t_k\}$ with the highest similarity score.
7: $P_{q_t} = \frac{\sum_{i=1}^{k} \sigma(q_t, t_i) * s_{t_i}}{\sum_{i=1}^{k} \sigma(q_t, t_i)}$
8: **return** P_{q_t}

and ε is the regression model error. Unfortunately, this model did not fit our data well, hence we did not consider this model further.

Our approach can be viewed as a variation of collaborative filtering. In collaborative recommendation systems, users explicitly provide item ratings, which are usually bounded and discrete. Past users ratings are then used to estimate user preference for yet unknown, hence unrated, items to the users. Collaborative filtering systems are divided into memory-based and model-based systems [14]. Memory-based systems calculate similarity between all users and predict a missing rate for user u by aggregating the ratings of u's k nearest neighbors. Model-based systems assume that users cluster together based on similar ratings on common items [1]. Machine learning techniques are often used in this case to learn the model. The motivation for collaborative filtering and our approach is similar. However, collaborative filtering techniques have many prior evidence (i.e., ratings for numerous items) for each user in their disposal, whereas in our case each torrent attribute, number of seeders tuple is unique for each torrent. Further, attributes are not bounded in our case, a fact that constitutes our estimation problem even harder.

Depending on the application, estimation of the exact number of seeders may be undesirable. Instead, the order of magnitude of the number of seeders may be highly appreciated. For example, consider a memory pre-allocation application for which the order of magnitude of seeders is more crucial than the exact number itself. Estimating that the number of seeders for torrent t_i will be 30 as compared to the true number 31 may not be as valuable as estimating that the number of seeders will be in the range of $[0, 50]$. Further, due to the skewed distributions of torrent features and seeders values, we anticipated that exact estimation of the number of seeders may be impossible. To approximate the order of magnitude of the number of seeders instead of estimating the exact number, we split the range of possible seeder values into clusters, such that torrents with s_i number of seeders belong to cluster $C_k \iff s_i \in [C_{k-1}, C_k)$.

3.1 Evaluation Metrics

We use two error metrics to evaluate the accuracy of our estimator. Fist, we use a statistical accuracy metric, *Mean absolute Error* (*MAE*), which evaluates the deviation of our model's numerical estimation against the actual number of seeders for query torrents in our test datasets. We compute MAE by calculating the absolute error between the true value of seeders s_i and the estimated seeders value p_i for each query torrent t_i, sum these absolute errors over N query torrents and compute the average. The lower the MAE, the more accurate the estimation of the exact number of seeders. Formally, $e_{\text{MAE}} = \frac{1}{N}\sum_{i=1}^{N}|s_i - p_i|$.

Our second metric, *Mean Absolute Error with Clustering* (*MABC*), treats the estimation process as a binary operation: either the correct order of magnitude (cluster) is estimated for the number of seeders or not. Assuming that torrent t_i has true value of seeders s_i and that the estimated seeders value is p_i, then the absolute error can be computed as $e_{ABS}(s_i, p_i) = 1_{\{s_i \neq p_i\}}$. Assuming further that $s_i \in C_{s_i}$ and $p_i \in C_{p_i}$, the absolute error with clustering can be computed as $e_{ABS_C}(s_i, p_i) = 1_{\{C_{s_i} \neq C_{p_i}\}}$. We compute MABC by calculating the absolute error between the cluster C_{s_i} of the true value of seeders s_i and the cluster C_{p_i} of estimated seeders value p_i for each query torrent t_i, sum these absolute errors over N query torrents and compute the average. The lower the MABC, the more accurate the estimation of the order of magnitude of the number of seeders. Formally, $e_{\text{MABC}} = \frac{1}{N}\sum_{i=1}^{N} e_{ABS_C}(s_i, p_i)$.

3.2 Experimental Results

For our experiments, we conducted a cross validation by randomly dividing our two datasets, D_{TPB} and D_{AVG} into disjoint training and test sets and averaging over the MAE and MABC values for each dataset. In our experiments we consider 10,000 queries and examine the estimation power of four vector spaces of torrents, represented by their corresponding feature vectors, namely: (a) $[n_f, n_t, n_c]$, (b) $[n_f, n_t, n_c, n_p, n_n, n_a]$, (c) $[n_c, n_p, n_n]$, and (d) $[n_p, n_n, n_a]$.

Figure 15 shows the Mean Absolute Error for our two datasets. Exact estimation is not possible using either data-set, however, MAE is significantly small in both cases. On average we get a distance of $\approx$10 from the exact number of seeders, using dataset D_{TPB}. This result makes as conjecture that even though estimation of the exact number of seeders is impossible, a qualitative approximation can be made. D_{TPB} outperforms D_{AVG}, achieving better estimation accuracy by at least one order of magnitude. This fact can be explained as a result of averaging in D_{AVG}. Aggregating disperse values, especially after the border of $\frac{10^3}{2}$, about which we discussed in Sect. 2.5, mitigates heterogeneous feature values of individual torrents, thus restricting the estimation accuracy of this model. The size of

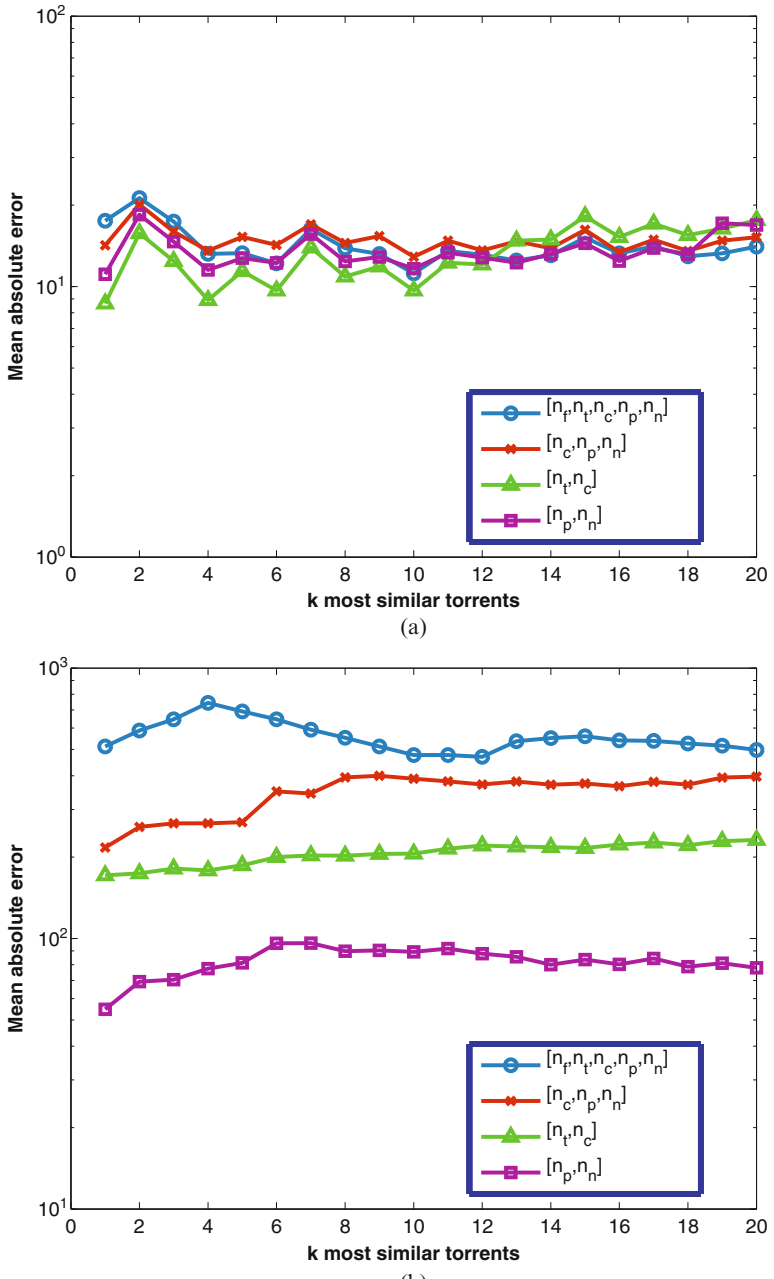


Fig. 15 Mean Absolute Error for (**a**) D_{TPB} and (**b**) D_{AVG} as a function of query neighborhood size

neighborhood does not quantitatively affect the estimation quality for D_{TPB}, even though small fluctuations of MAE can be observed. On the contrary, increasing values of neighborhood size seem to have a small effect in the case of D_{AVG}, where MAE slightly increases. Based on these observations, multiple k values provide the best accuracy for D_{TPB}, whereas $k = 1$ provides the best results for D_{AVG}.

Figure 16 shows the Mean Absolute Error with clustering for our two datasets, with fixed cluster size of 500, such that for a query torrent with true number of seeders $s(q_t)$ and estimated number $s'(q_t)$ we have zero absolute error when $|s(q_t) - s'(q_t)| \leq 500$. The error in this case is practically zero. To determine estimation accuracy sensitivity to cluster size, we carried out an experiment where we selected the best feature vectors for each of the datasets and we varied the value of cluster size. We considered six cases of decreasing granularity (by consequently increasing the cluster size) where the cluster size remained fixed, and two cases where we varied the cluster size to get a combination of fine and coarse granularities for different ranges of the seeders space. The first variable cluster size scheme progressively decreases in granularity from fine to coarse, while the second does the exact opposite: it begins with a coarse grained split of the space and progressively considers smaller clusters of the number of seeders.

Figure 17 demonstrates the impact of cluster size on estimation accuracy. We observe that estimation quality increases with increasing cluster sizes for both datasets. In practice, we can achieve great performance for any cluster size greater than or equal to 50. In other words, for any given query torrent with true number of seeders $s(q_t)$ and estimated number $s'(q_t)$ we can achieve close to zero absolute error when $|s(q_t) - s'(q_t)| \leq 50$. For variable cluster size, we can achieve great performance when the neighborhood size is less than 12. At that point we observe a waterfall effect, after which the estimation accuracy rapidly decreases for D_{TPB}. This effect is not visible for D_{AVG} due to mitigation of diverse feature values into their aggregated values.

4 Related Work

Even though it has been reported that peer-to-peer traffic was declining due to copyright laws [17], more recent work has argued that it is in fact increasing and that it constitutes a significant fraction of the total workload transmitted in the Web [17, 24]. Peer-to-peer network studies have thus far been focusing on topological characteristics of peer-to-peer networks [9, 10] or properties such as bandwidth rates, churn and overhead [8], download and upload times [5, 18], content availability [4], and deviant users identification [23, 25].

Han et al. [13] conducted an empirical study of TPB focusing on "how prevalent bundling is and how many files are bundled in a torrent, across different types

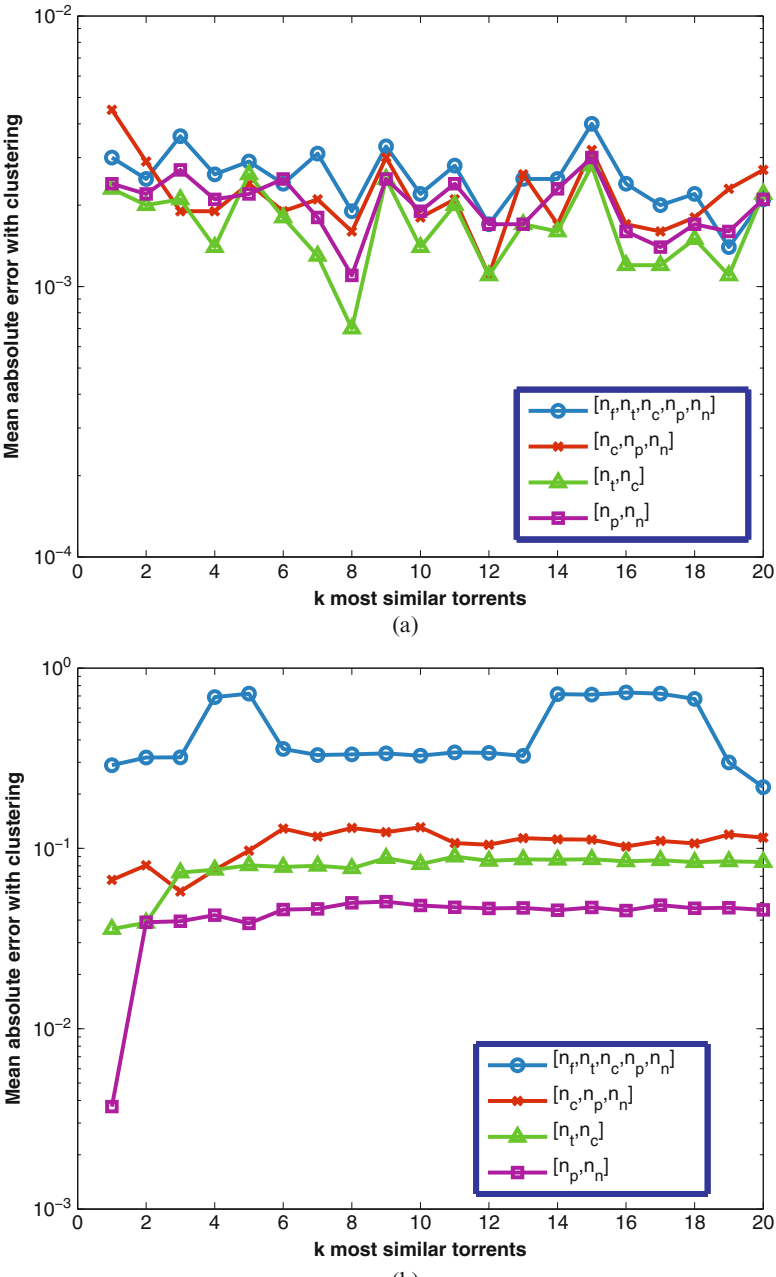
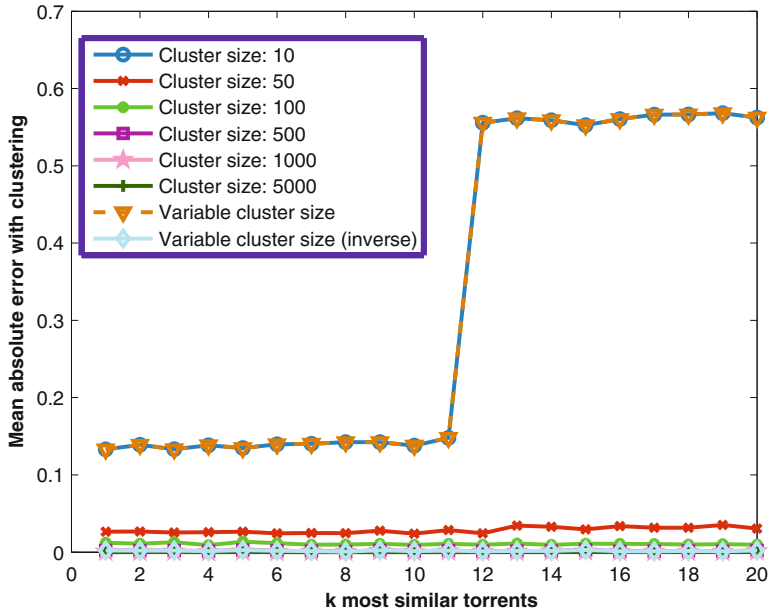
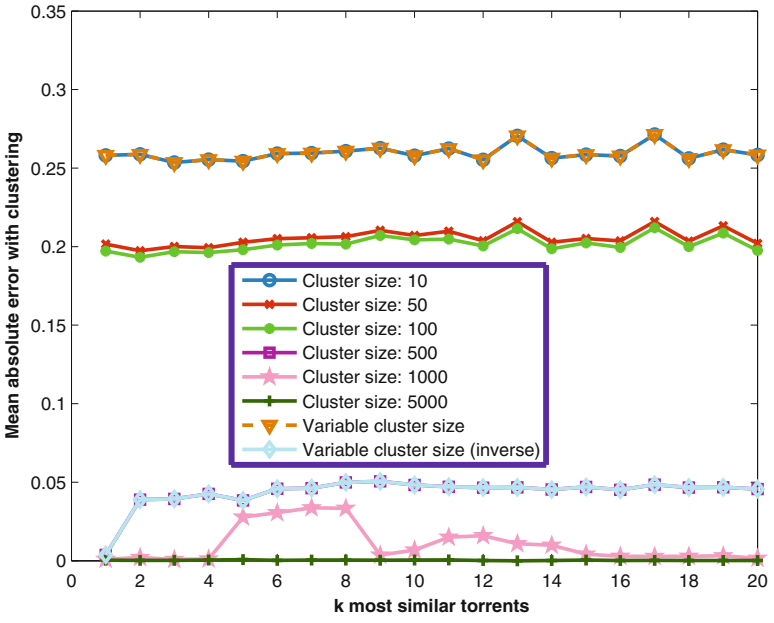


Fig. 16 Mean Absolute Error with Clustering for (**a**) D_{TPB} and (**b**) D_{AVG} as a function of query neighborhood size



(a)



(b)

Fig. 17 Mean Absolute Error with Clustering for (**a**) D_{TPB} and (**b**) D_{AVG}, with different clustering sizes

of contents shared: Movie, Porn, TV, Music, Application, E-book, and Game" [13]. Bieber et al. [15] examined the relationship between number of seeders and bandwidth utilization, and between site attributes and number of seeders. More specifically, [15] examined the correlation between number of seeders and (1) the effect of having to register to use a web site, (2) whether a web site has a prominent reminder for users to leave their client open, (3) whether a network distributes niche-content (e.g., only anime, hip-hop, or Linux files) or general content, (4) whether a site posts the identities of their top 10 and bottom 10 seeders, and (5) whether a site sells site-specific merchandise.

Prior art (e.g., [15]) has not performed a systematic examination of torrents with respect to their features so as to understand correlations between torrent features and examine the impact of such features to torrents quality, popularity, and user feedback. To the best of our knowledge no effort has ever been conducted to estimate torrent popularity based on torrent features.

5 Conclusion

We conducted a thorough empirical analysis of The Pirate Bay torrents with respect to their features, identified correlations between features, and provided insights into the use of such features for estimation of torrent popularity and quality. We defined a robust measure to compute similarity between torrents and applied this measure as accurate estimator of the number of seeders for previously unseen torrents. We showed that estimating the exact number of seeders is difficult, partially due to large variances and skewed distributions of torrent features, and we argued that linear regression does not perform well in this case. We developed a vector space model which provides close to perfect accuracy as a function of estimation granularity, when the order of magnitude of the number of seeders is preferable to the exact number.

To the best of our knowledge, our work is the first to study torrent popularity and propose an accurate estimator of the number of seeders based on a small set of publicly available metadata associated with the actual pirated content. Based on this work, we identify several directions, including: (a) further examining popularity over time (for example, people may be interested in a current season of a TV show as compared to older seasons of the same show, (b) the effect of user comments on long term torrent popularity, (c) exploring the predictive power of additional features such as the credibility of the user that uploaded the torrent (e.g., how many popular torrents this user has shared in the past) since torrent downloaders often trust torrents that have been posted by highly rated users for downloading, and box-office data, (d) the co-evolution of popularity of similar torrents in time, and (e) automatically discover correlation of features in a latent space.

References

1. Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE Trans. Knowl. Data Eng. **17**(6), 734–749 (2005)
2. Chatzopoulou, G., Sheng, C., Faloutsos, M.: A first step towards understanding popularity in youtube. In: INFOCOM IEEE Conference on Computer Communications Workshops, 2010, pp. 1–6. IEEE, New york (2010)
3. Cheng, J., Adamic, L., Dow, P.A., Kleinberg, J.M., Leskovec, J.: Can cascades be predicted? In: Proceedings of the 23rd International Conference on World Wide Web, WWW '14, pp. 925–936 (2014)
4. Christin, N., Weigend, A.S., Chuang, J.: Content availability, pollution and poisoning in file sharing peer-to-peer networks. In: Proceedings of the 6th ACM conference on Electronic commerce, EC '05, pp. 68–77. ACM, New York (2005)
5. Di, W., Dhungel, P., Xiaojun, H., Chao, Z., Ross, K.W.: Understanding peer exchange in bittorrent systems. In: IEEE Tenth International Conference on Peer-to-Peer Computing (P2P), pp. 1–8 (2010)
6. Ding, W., Shang, Y., Guo, L., Hu, X., Yan, R., He, T.: Video popularity prediction by sentiment propagation via implicit network. In: Proceedings of the 24th ACM International on Conference on Information and Knowledge Management, CIKM '15, pp. 1621–1630 (2015)
7. Fabio, H., Thomas, B., David, H.: The pirate bay 2008-12 dataset. http://www.csg.uzh.ch/publications/data/piratebay/
8. Falkner, J., Piatek, M., John, J.P., Krishnamurthy, A., Anderson, T.: Profiling a million user dht. In: Proceedings of the 7th ACM SIGCOMM conference on Internet measurement, IMC '07, pp. 129–134. ACM, New York (2007)
9. Farzad, A., Rabiee, H.: Modeling topological characteristics of bittorrent-like peer-to-peer networks. IEEE Commun. Lett. **15**(8), 896–898 (2011)
10. Fletcher, G.H.L., Sheth, H.A.: Unstructured peer-to-peer networks: topological properties and search performance. In: Third International Joint Conference on Autonomous Agents and MUlti-Agent Systems. W6: Agents and Peer-to-Peer Computing, pp. 14–27. Springer, Berlin (2004)
11. Gibbs, S.: Swedish police raid sinks the pirate bay. The Guardian (2014)
12. Guo, R., Shaabani, E., Bhatnagar, A., Shakarian, P.: Toward order-of-magnitude cascade prediction. In: Proceedings of the 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining 2015, pp. 1610–1613. ACM, New York (2015)
13. Han, J., Chung, T., Kim, S., Kwon, T.T., Kim, H.c., Choi, Y.: How prevalent is content bundling in bittorrent. In: Proceedings of the ACM SIGMETRICS Joint International Conference on Measurement and Modeling of Computer Systems, SIGMETRICS '11, pp. 127–128 (2011)
14. Herlocker, J.L., Konstan, J.A., Borchers, A., Riedl, J.: An algorithmic framework for performing collaborative filtering. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '99, pp. 230–237. ACM, New York (1999)
15. Justin, B., Michael, K., Nick, T., Cox, L.P.: An empirical study of seeders in bittorrent. Technical report, Duke University (2006)
16. Karaganis, J., Renkema, L.: Copy culture in the US & Germany (2013)
17. Karagiannis, T., Broido, A., Brownlee, N., Claffy, K., Faloutsos, M.: Is P2P dying or just hiding? [P2P traffic measurement]. In: Global Telecommunications Conference, 2004. GLOBECOM '04, vol. 3, pp. 1532–1538. IEEE, New York (2004)
18. Liu, Z., Dhungel, P., Wu, D., Zhang, C., Ross, K.W.: Understanding and improving ratio incentives in private communities. In: Proceedings of the 2010 IEEE 30th International Conference on Distributed Computing Systems, ICDCS '10, pp. 610–621. IEEE Computer Society, Washington (2010)

19. McKelvey, F.: We like copies, just dont let the others fool you the paradox of the pirate bay. Telev. New Media **16**(8), 734–750 (2015)
20. Menczer, F.: Lexical and semantic clustering by web links. J. Am. Soc. Inf. Sci. Technol. **55**, 1261–1269 (2004)
21. News, B.: The pirate bay 'breaches' bt's ban of the filesharing site (2012)
22. Pinto, H., Almeida, J.M., Gonçalves, M.A.: Using early view patterns to predict the popularity of youtube videos. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 365–374. ACM, New York (2013)
23. Ripeanu, M., Mowbray, M., Andrade, N., Lima, A.: Gifting technologies: A BitTorrent case study. First Monday. **11**(11), (2006)
24. Saroiu, S., Gummadi, K.P., Dunn, R.J., Gribble, S.D., Levy, H.M.: An analysis of internet content delivery systems. SIGOPS Oper. Syst. Rev. **36**, 315–327 (2002)
25. Siganos, G., Pujol, J., Rodriguez, P.: Monitoring the bittorrent monitors: a bird's eye view. In: Moon, S.B., Teixeira, R., Uhlig, S. (eds.) Passive and Active Network Measurement. PAM 2009. Lecture Notes in Computer Science, vol. 5448, pp. 175–184. Springer, Berlin (2009)