Piracy, Public Access, and Preservation: An Exploration of Sustainable Accessibility in a Public Torrent Index

John D. Martin III
School of Information and Library Science
University of North Carolina at Chapel Hill
me@johndmart.in

ABSTRACT

The present study explores the feasibility of using torrent networks as potential repositories for popular cultural materials and historical primary source data. Members of the Pirate Parties International have claimed that torrent networks function in this manner, even potentially as replacements for public libraries. *The Pirate Bay*, the world's most popular public index for tracking and downloading torrents is conceived as a potential repository. Using an open-sourced dataset based on Karel Bílek's 2012-2013 snapshot of *The Pirate Bay,* metadata from 2.1 million torrents are categorized by media type and the robustness of given torrents was assessed. Trends over time for features of each media type, such as number of uploads, size, and volume were also assessed. The study finds that, in fact, relatively few torrents exhibit long-term survivability, even though the overall volume of torrents in the index shows continuous increase.

Keywords: media piracy, P2P file-sharing, torrents, *The Pirate Bay*, digital repositories.

INTRODUCTION

One of the easiest vectors for access to pirated media content is the torrent public index, a website which tracks a set of BitTorrent files—or torrents—and then provides links to them so that content may be downloaded through client software. Public indices for torrents serve several purposes. They provide downloaders with an easily searchable database of media content available through associated trackers. They provide uploaders with a forum for advertising content which they wish to make available to a wide audience. Most of the content is not licensed for redistribution—in other words, it is pirated. For the last ten years, *The Pirate Bay*, originally a Swedish site, has been among the most popular sites to index torrents. *The Pirate Bay*, along with its associated tracker, does not represent the totality of the BitTorrent-based P2P network that exists in the wild, but it has been the top torrent public index for 7 of the last 10 years of its existence. Additionally, it has been one of the most durable, resisting domain seizures, blocking, raids and seizures of equipment.

ASIST 2016 October 14-18, 2016, Copenhagen, Denmark

The Pirate Bay as library

In 2012 Zacqary Green wrote a blog post proposing that *The Pirate Bay* the most popular public index for BitTorrent users, is effectively the world's most efficient public library (2012). The post opens with the following bait-and-switch:

> The way digital piracy works is that one person or group purchases a work, and then shares it with millions of other people. This supposedly deprives the author or artist of those millions of people's money. One group has acquired over 50 million media items, and makes each of them available to approximately 20 million people which must be a tremendous hit to creative professionals' wallets. This notorious institution is called the New York Public Library.

Green ignores the mechanisms through which the New York Public Library acquires its content and the huge licensing fees built into such systems. Adrian Johns discusses the possibilities presented by Google's book scanning projects for creating a massive digital library to be made globally available. Instead of becoming a resource for the dissemination of literature and text, the majority of the scanned documents remain hidden from view and the corpus only available to sufficiently qualified researchers (Johns, 2009, 512-513). Lawrence Lessig addresses the prohibitive expense associated with the collection of television, film and music in analog form in the ninth chapter of his book, *Free Culture*, titled "Collectors." He identifies that books and print media have the possibility of a "second life" once their commercial value has been exhausted, which is usually very quickly, in real temporal and economic terms. The same is not true for "the most important components of popular culture in the twentieth and twenty-first centuries." For these, Lessig argues:

> television, movies, music, radio, the Internet—there is no guarantee of a second life. For these sorts of culture, it is as if we've replaced libraries with Barnes & Noble superstores. With this culture, what's accessible is nothing but what a cer-

> tain limited market demands. Beyond that, culture disappears (Lessig, 2004, 113).

Green was writing for the website of Swedish Pirate Party founder Rick Falkvinge. The site covers information piracy, privacy, security and policy issues and is oriented plainly toward an anti-copyright, pro-piracy worldview—or at least one which advocates radical copyright reform. Falkvinge followed Green's post a week later with one of his own listing four more reasons that *The Pirate Bay* is effectively acting in the capacity of a public library.

Falkvinge (2011) previously elaborated on the laws that were put in place to give copyright exception to libraries because of the function that they serve. Engström and Falkvinge (2012) proposes similar set of exceptions be made for peer-to-peer file-sharing and other services and activities that serve to extend the non-commercial dissemination of cultural production.

If *The Pirate Bay* is to be rendered analogous to a public library, then it is necessary to evaluate its content and use in terms very different from the current dominant treatment of its use and maintenance as a purely criminal enterprise. Such a shift in conception raises a number of questions about the medium itself (i.e., the public torrent index or torrent tracker) as it relates to the enterprise of digital piracy as a whole as well as to its development, both historical and projected. There are several questions of interest related to this conception of torrent networks.

Research questions

The present study aims to address three questions regarding digital media piracy:

a. What is the shape of *The Pirate Bay* as a repository in terms of media types represented by percentage?

b. What trends in media uploading and sharing can be identified in the *The Pirate Bay*?

c. How robust is a torrent-based network in terms of preserving media and making it available as represented by *The Pirate Bay*?

LITERATURE REVIEW

Literature on digital piracy tends to focus on the implications of copyright violation and control or deterrence rather than attempting investigate the networks and mechanisms through which the activity occurs (Higgins, 2007; Danaher, Smith, & Telang, 2014; Driouchi, Wang, & Driouchi, 2014; Yoo, Sanders, Rhee, & Choe, 2014). Theoretically, the literature tends to be oriented toward deterrence and business/information technology (IT) ethics or toward economic modeling of the effects of piracy on legitimate markets (Marx, 2013; Miyazaki, Rodriguez, & Langenderfer, 2009). Several projects in recent years have begun to address copy culture, sharing, and the motivations underlying the circumvention of licit markets for media distribution (Karaganis, 2011; Karaganis & Renkema, 2013; Nguyen, 2013; Schwarz & Eckstein, 2014; Yar, 2005). Study of the content of the systems that support digital piracy is quite infrequent. Karaganis (2011) found that price points are among the most problematic aspects of the phenomenon of media piracy and that very few if any of the mitigation strategy put in place by governments and media industry groups are effective in combating the behavior. Karaganis and Renkema (2013) analyzed media piracy in the United States and Germany and found that nearly half of the sampled population had copied or shared media and software illegally. There is a great deal of ambiguity with regard to the boundaries of what qualifies as private copying and copying for redistribution.

Price (2013) attempted to estimate the size of the torrent "universe," using a public torrent tracker to estimate percentages of content by type. The most interesting finding from their analysis is that approximately 99.75% of content tracked by their selected public tracker, PublicBT, was found to be infringing content.

METHODS

Media types were manually identified within a sample of the data and then terms intended to be matched for extraction were identified using word frequencies of each target media type. The most frequent terms that did not overlap with the term groups associated were assigned to a given group. Each group was matched using a simple string matching with regular expressions.

Data

The data utilized in the present study was gathered in early 2013 by Karel Bílek. The archive contains over two million records associated with individual torrent files that are being tracked by *The Pirate Bay*. Bílek made the data available via *The Pirate Bay* so that it could be used for research.

Extraction

Using some of *The Pirate Bay's* browse categories, match-term lists were created for the following mutually exclusive categories: *a*) audio, *b*) video, *c*) books, *d*) images, *e*) software, *f*) games, and *g*) other.

There is some legitimately-distributed content indexed by *The Pirate Bay* that had to be addressed. For the purposes of the present study Linux images represented in the dataset were used as a proxy for content legitimately distributed through through a public torrent index. Known free-culture films and other media were also searched and included in searches for non-infringing content.

Verifying Extraction

In order to estimate the relative accuracy of the counts associated with the categories listed above, the data was bootstrapped to estimate sample means to be compared with counts on the full dataset by percentage of total, as well as the estimation of standard deviation and confidence intervals around the counts. 10,000 resamples of 2,500 records were drawn at random with replacement from the full dataset (2,142,134 total records). Descriptive statistics are shown in Table 1.

Estimating usage and robustness of The Pirate Bay

In the data available, there are several possible avenues for estimating usage: *a*) volume of data uploaded; *b*) total number of uploads for a given time period; and *c*) number of peers available (divided into seeders and leechers). A "peer" in the context of BitTorrent is an individual computer using a program which advertises that it either has or is trying to get a specific torrent. "Seeders" are individual computers uploading the files associated with a given torrent at any given time. "Leechers" are individual computers downloading the files associated with a given torrent at any given time. All peers start off as leechers and then become seeders when they have completely downloaded a file from other peers. Both leechers and seeders upload, but only leechers download.

Because the number of peers changes over time, and given that the dataset in question was aggregated over the course of multiple months, the number of peers associated with a given torrent cannot be taken to be an exact measurement of the quality of the network or even the quality of a given torrent. In aggregate, however, the number of peers can give useful information about whether a given torrent is available or not at the time of the snapshot. The existence of a torrent is reckoned as any torrent with one or more seeders associated with it. Leechers are disregarded for this purpose because they represent incomplete copies of the files associated with a given torrent. A robust torrent is defined as any torrent with two or more seeders. A dead torrent is defined as any torrent with no active seeders at the time of snapshot.

Volume of data uploaded and number of torrents uploaded in a given time period will be examined below. Both of these will be broken down categorically as well.

RESULTS

Table 1 shows the percentages of the data that were categorized using the entire dataset and the estimates for percentages based on the bootstrap described above

Categories of media

The records categorized from the entire dataset all fall within the confidence interval with non-significant values by percentage at $\alpha = 0.10$. Since the counts reported in the first column of Table 1, by percentage of full dataset, are not significantly different from the parameter estimates in the second column, by percentage of sample, the counts can be reliably used to estimate the usage trends in the next section. The t scores and confidence intervals reported in in Table 1 were computed by comparing the mean counts from samples drawn in the bootstrap tested against the full dataset count. Extremely low p-values indicate that it is unlikely that these estimates would have been arrived at by chance. Standard deviations and ranges for the distribution of mean counts in the samples are reported in columns 3-5 in in Table 1.

Just slightly less than half (48.6%) of the total torrents in the dataset matched as video torrents, followed by slightly more that a quarter (27.4%) matched as audio. The other categories comprised a considerably smaller block of the data. 8.5% of the torrents represented in the data did not match on any of the categories included in the analysis. The remaining unmatched torrents comprise a range of different media types, mostly from the "other" category in the internal browse categories used by *The Pirate Bay.*

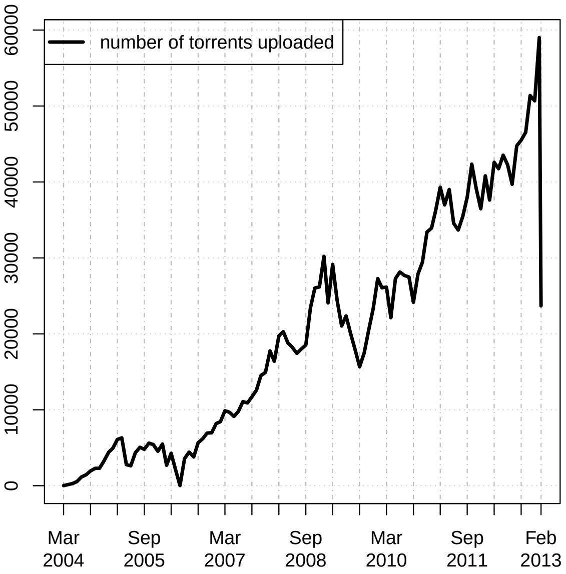

Figure 1: Number of uploads per month

Usage Trends

The figures show the trends in overall usage from the the opening of *The Pirate Bay* index in 2004 to the time of the data collection in early 2013. Figure 1 shows the number of torrents uploaded to *The Pirate Bay* each month from March 2004 to mid-February 2013. The general trend sees a five-fold increase in number from below 10,000 torrents per month in the first three years of its existence to well over 50,000 torrents a month in the most recent period in the dataset. Since number of torrents does not necessarily mean anything with regard to the volume of the data being represented, it was important to consider overall increase in monthly volume uploaded as well. Figure 3 shows the monthly volume of data represented in the torrents uploaded in TB. This volume has has climbed alongside the number of torrents uploaded, with the most recent monthly total represented in the data at just below 80 TB of new data uploaded per month. The total size of the data represented by all uploads is shown in Figure 4 in PB. The size of the total historical uploaded data and the number of torrents (not pictured) both exhibit exponential growth, as can be seen in Figure 4. The total data represented by all torrents uploaded to *The Pirate Bay* tops out at the end of the data collection period at 2.5 PB. Figure 2 displays the mean size (in MB) of data represented by torrents uploaded by month

	All data	Bootstrap estimates				Test statistics		Confidence Interval	
	($N = 2,142,134$)	($n = 2,500 \times 10,000$ resamples)				($df = 9999$)		($\alpha = 0.10$)	
Category	%	Mean %	Std. Dev.	Min	Max	t score	p-value	Lower	Upper
Audio	27.407	27.421	0.901	24.280	30.960	-301091.90	< 0.001	27.407	27.435
Video	48.608	48.606	0.999	45.080	52.400	-481773.40	< 0.001	48.589	6.633
Books	6.644	6.641	0.496	4.680	8.880	-333924.00	< 0.001	6.633	6.649
Software	2.277	2.278	0.299	1.320	3.480	-75423.96	< 0.001	2.273	2.283
Games	1.283	1.283	0.225	0.560	2.160	-56402.36	< 0.001	1.279	1.286
Image	5.380	5.379	0.449	3.600	7.200	-118584.50	< 0.001	5.371	5.385

Table 1: Percentage of media types and estimates from bootstrap

for the period in the dataset. As shown, the average size of upload varies considerably for the first four years of the existence of *The Pirate Bay*, but largely remains below the 1 GB mark except for a few spikes. However, after about June 2009, the average size increases to over 1 GB and never returns to a smaller size.

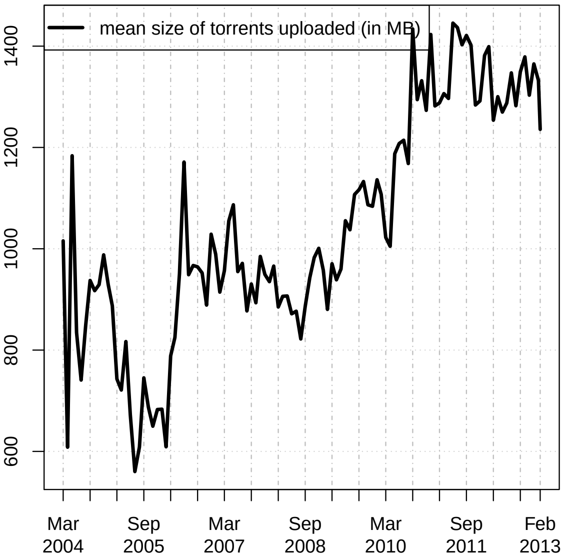

Figure 2: Mean size (in MB) of upload by month

Non-infringing Content

Both estimates and full dataset counts for non-infringing content were extremely low at 0.02601% and 0.02586% respectively.

Robustness of The Pirate Bay

Table 2 describes the state of *The Pirate Bay* in terms of what scale of media is actually available when compared to all uploads of torrents to the index. If torrents that are no longer downloadable are considered to be "dead," or having zero full copies available in the torrent networks, then just over a third of the data (36.35%) falls into this category. Torrents with only one available copy, or seeder, were considered to be extant but at the brink of nonexistence. These make up just under a quarter of the total data at 23.27%. Only 40.38% of the total torrents historically uploaded to *The Pirate Bay*, as represented in the available data, can be considered to be robustly available; that is, having two or more copies available at the time of the snapshot. As mentioned above, these percentages are estimates, given that the number of peers varies drastically from hour to hour withing torrent networks. They do, however, help to give a sense of the capacity of torrent networks, as represented by *The Pirate Bay* for preservation of media and the provision for access thereto. The fact of the total number of available torrents being so small means that the volume of data is considerably smaller than otherwise estimated as well: 1.14 PB down from 2.52 PB, or 45.2%.

DISCUSSION

While it is clear that video dominates torrent networks by a wide margin as the most shared medium, audio is in a reasonable second place. These two media have risen both in terms of individual uploads as well as total size of data over the nine years represented in the available data by wide margins over other data types. There are several possible explanations for this phenomenon.

First, the price of storage capacity has dropped considerably over the last decade as has the relative cost of high-bandwidth transfer globally. Second, video and audio devices have increased their storage capacity and capacity for resolution apace with the decrease in cost of storage capacity and data transfer. An general increase in demand for high resolution video spurred by the wide saturation of electronics markets with high-definition (HD) video devices may also have the effect of increasing demand within torrent networks for HD video. This is particularly likely when HD video services like Hulu, Netflix, Amazon Instant Video, etc. are encumbered by licensing obstacles for making the most current high-resolution media content available at prices con-

Category	Definition	# torrents	% of records
Dead	0 seeders	778,655	36.35
Extant	only 1 seeder	498,573	23.27
Robust	2 or more seeders	864,906	40.38
Total		2,142,134	

Table 2: Robustness of torrents in dataset

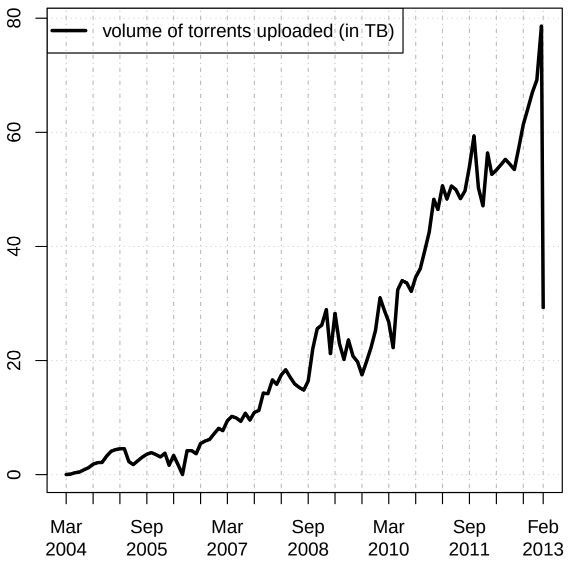

Figure 3: Volume (in TB) of uploads per month

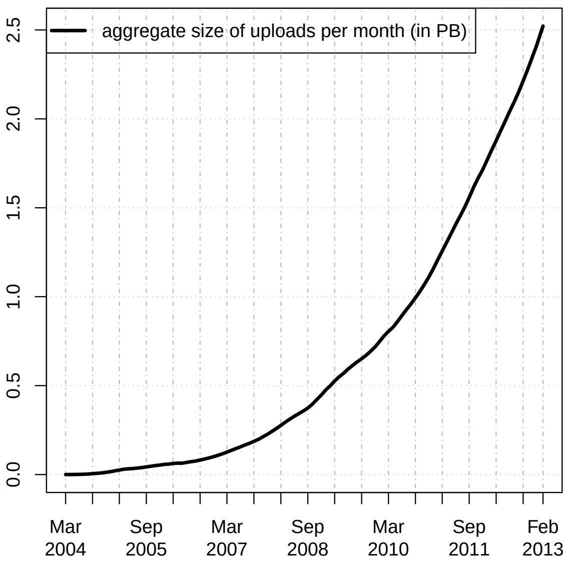

Figure 4: Aggregate size (in PB) of uploads per month

sumers are willing to pay.

Similarly for audio, the increase in desire for high-resolution audio may be expressed in *The Pirate Bay* in the form of an increase in the average size of audio files being uploaded, as shown in the data. There is still a dearth of services available for streaming of high resolution audio in the market. Torrent networks, on the other hand, yield a huge number of available media encoded using "lossless" audio formats, such as the Free Lossless Audio Codec (FLAC).

The percentages for video are similar to those reported in Price (2013, 29) with film and TV combined at 48.7% of their sample from PublicBT. However, they also included pornography in their analysis as a category mutually-exclusive to others which comprised 30.3%. While Price used a public tracker, PublicBT, and only selected 12,500 torrents at random from those being tracked by the system, the present study considers the entire snapshot of all torrents ever uploaded to a public index, *The Pirate Bay*. All torrents in PublicBT must have been active at the time of sampling or they would not have been able to access them through the tracker. The present study utilized records associated with 2,142,134 individual torrents uploaded to *The Pirate Bay* over the course of nine years. Additionally, the present study estimated parameters around the counts derived from the individual categories in order to lend some reliability to the analytic strategy. Previous studies have not given this kind of analysis.

More important, however, than the vast increases in torrent usage over time is the relative availability of that content. If content is being uploaded, utilized for a short time, and then abandoned by users, then the increases in usage and uploads are somewhat misleading. Instead it means that content is actually being lost. As shown above, that loss translates into the availability of only a small proportion of content relative to total uploads.

This finding challenges the popular notion in Pirate Party political discourse that a torrent network might serve as some sort of repository for preserving content and maintaining access. In reality, it is a fragile ecosystem that relies on individual users to maintain connections in order to distribute content. Issues of copyright aside, this is not a good strategy for preservation or the provision of access. Content moves on and off the network ephemerally. Continued efforts on the part of media lobby organizations to target torrent networks and indices for legal action only serve to make the ecosystem more fragile. In the face of all such obstacles, use of torrent networks remains ceaseless. This alone is not enough to preserve the content that the networks may have once contained.

CONCLUSION

Returning to the claims made by piracy activists as identified in the introduction, the more than doubling in the use of *The Pirate Bay* in under two years does indeed indicate that there are appeals to using this system for the distribution of digital media. Such an increase also means, as argued by Lessig, that the cultural data being accessed through piratical means has increased. Since the robustly available torrents found in the present study sit at approximately 40% of the total of everything that has ever been uploaded to *The Pirate Bay*, it seems likely that as the total aggregate size of the network grows, so will the aggregate loss. This warrants further investigation through scraping, logging and monitoring *The Pirate Bay* and other public indices to develop at least cross-sectional if not longitudinal knowledge about changes in the shape of torrent networks as they continue to mature and develop.

The problem, as identified in the present study, is not an increase in content pirated, but rather really an increase in the total content we are losing by not preserving what is being made available via torrent networks. This runs counter to the arguments of piracy and free culture advocates such as Falk-vinge, Lessig, and Green that these systems are currently being used to support a grand subaltern repository for access and preservation. Instead, torrent networks and other channels for piratical transmission of media content represent vast content graveyards, where metadata about what was once available is preserved forever in public and private indices, even though the actual media objects themselves may no longer be available in any form. This problem has manifold causes, one of which is that the general consensus is that since this mode of distribution violates copyright and is therefore illegal, it has no value, and therefore the data in it have only a negative economic value. On the contrary, the variety and richness of the content found in torrent networks may eventually represent a tragic loss of cultural data, should the systems that support it every be completely disabled. For the time being though, there is no sign of that on the horizon.

References

Bílek, K. (2013, February). *Complete Pirate Bay archive (february 2013) [Data File].* Retrieved from http://thepiratebay.se/torrent/8163015

Danaher, B., Smith, M. D., & Telang, R. (2014). Piracy and copyright enforcement mechanisms. *Innovation Policy and the Economy*, *14*(1), 25–61. Retrieved from http://dx.doi.org/10.1086/674020 doi: 10.1086/674020

Driouchi, A., Wang, M., & Driouchi, T. (2014, 17 June). Determinants of software piracy under risk aversion: a model with empirical evidence. *European Journal of Information Systems*, *24*(5), 519–530. Retrieved from http://dx.doi.org/10.1057/ejis.2014.14 doi: 10.1057/ejis.2014.14

Engström, C., & Falkvinge, R. (2012). *The case for copyright reform.* Pirate MEP Christian Engström with support from the Greens/EFA-group in the European Parliament. Retrieved from http://www.copyrightreform.eu/sites/copyrightreform.eu/files/The_Case_for_Copyright_Reform.pdf

Falkvinge, R. (2011, February 8). *History of copyright, part 4: The US and libraries.* Retrieved from http://falkvinge.net/2011/02/08/history-of-copyright-part-4-the-us-and-libraries/

Falkvinge, R. (2012, December). *Four more reasons The Pirate Bay is effectively a public library—and a great one.* Retrieved from http://falkvinge.net/2012/12/13/four-more-reasons-the-pirate-bay-is-effectively-a-public-library-and-a-great-one/

Green, Z. A. (2012, December 7). *The Pirate Bay is the world's most efficient public library.* Retrieved from http://falkvinge.net/2012/12/07/the-pirate-bay-is-the-worlds-most-efficient-public-library/

Higgins, G. E. (2007). Digital piracy, self-control theory, and rational choice: An examination of the role of value. *International Journal of Cyber Criminology*, *1*(1), 33–55. Retrieved from http://www.cybercrimejournal.com/georgeijcc.pdf

Johns, A. (2009). *Piracy: the intellectual property wars from Gutenberg to Gates.* University of Chicago Press.

Karaganis, J. (Ed.). (2011). *Media piracy in emerging economies.* New York, New York: Social Science Research Council. Retrieved from http://piracy.americanassembly.org/wp-content/uploads/2011/06/MPEE-PDF-1.0.4.pdf

Karaganis, J., & Renkema, L. (2013). *Copy culture in the US and Germany* (Tech. Rep.). American Assembly/Columbia University. Retrieved from http://piracy.americanassembly.org/wp-content/uploads/2013/01/Copy-Culture.pdf

Lessig, L. (2004). *Free culture.* New York, New York: Penguin.

Marx, N. (2013). Storage wars: Clouds, cyberlockers, and media piracy in the digital economy. *Journal of E-Media Studies.* Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.672.1985&rep=rep1&type=pdf

Miyazaki, A. D., Rodriguez, A. A., & Langenderfer, J. (2009). Price, scarcity, and consumer willingness to purchase pirated media products. *Journal of Public Policy & Marketing*, *28*(1), 71–84. Retrieved from http://dx.doi.org/10.1509/jppm.28.1.71 doi: 10.1509/jppm.28.1.71

Nguyen, L. U. (2013). *Networks at their limits: Software, similarity, and continuity in vietnam* (Doctoral dissertation, University of California, Los Angeles). Retrieved from http://search.proquest.com.libproxy.lib.unc.edu/pqdtglobal/docview/1412732361/fulltextPDF/D8F25D8D5E1946D2PQ/1?accountid=14244

Price, D. (2013, September). *Netnames piracy analysis: Sizing the piracy universe* (Tech. Rep.). NetNames/Envisional.

Schwarz, A., & Eckstein, L. (2014). *Postcolonial piracy: Media distribution and cultural production in the global south.* London: Bloomsbury Publishing.

Yar, M. (2005, 1 September). The global epidemic of movie piracy: crime-wave or social construction? *Media Culture & Society*, *27*(5), 677–696. Retrieved from http://mcs.sagepub.com/content/27/5/677.abstract doi: 10.1177/0163443705055723

Yoo, C.-W., Sanders, G. L., Rhee, C., & Choe, Y.-C. (2014, 1 November). The effect of deterrence policy in software piracy: cross-cultural analysis between Korea and Vietnam. *Information Development*, *30*(4), 342–357. Retrieved from http://idv.sagepub.com/content/30/4/342.abstract doi: 10.1177/0266666912465974