Exposing Digital Content Piracy: Approaches, Issues, and Experiences

(Invited Paper)

Simon Byers*, Lorrie Cranor†, Eric Cronin‡, Dave Kormann*, and Patrick McDaniel§

*AT&T Labs – Research, Florham Park, NJ – email: {*byers,davek*}*@research.att.com*

†School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213 – email: *lorrie@cs.cmu.edu*

‡CIS Department, University of Pennsylvania, Philadelphia, PA – email: *ecronin@cis.upenn.edu*

§Systems and Internet Infrastructure Security (SIIS) Laboratory, Computer Science and Engineering
Pennsylvania State University, University Park, PA 16802 – email: *mcdaniel@cse.psu.edu*

Abstract— **The explosive growth of peer-to-peer networks has been cited as a cause of a host of social and economic ills. However, much of the assessment of content misuse in these networks occurs in a vacuum of information. In a previous study we showed that insiders were a major source of movies leaked onto peer to peer networks. Such results were deemed controversial by the press and the movie industry, and led to significant public debate. This paper revisits our study and proposes some future work. We conclude with some brief notes on the recent industry moves to mitigate insider and outsider piracy.**

I. Introduction

The Internet introduces new avenues for illegally sharing digital artifacts. The ubiquity of access and simplicity of peer-to-peer systems allows content to be shared globally with little effort. Such sharing has been cited as costing content providers and software manufacturers billions of dollars [1], [2]. Yet, the public has little independent data to show the existence, magnitude, and source of the problem.

This paper discusses our study of digital movie piracy on one peer-to-peer network [3]. An outline of an expanded study is given, and we propose a new *piracy analysis center* aimed at providing an ongoing, unbiased view of digital content misuse on the Internet. We conclude with a discussion of ways the industry has improved their processes and posit the long-term impact of this work.

Our study principally showed that more than half of popular movies were available in sharing networks, most in high quality DVD format. In a more surprising result, we found that 78% of the available movies appear to be leaked *by members of the movie industry itself* (henceforth called insiders). Because it places at least some of the blame of piracy at the door of the movie production and distribution industry, this latter point became the focal point of the many press stories.

The timing of movie leakage was also enlightening. One out of forty ($\approx 2.5\%$) of the movies we studied were leaked before they even reached the theater. These movies were often unfinished. In the case of the movie "The Hulk", the rip[1] contained few special effects [4]. Because of the nature of the movie, the lack of effects subjectively made for poor viewing, and was the subject of significant ridicule from bloggers. While the true impact of these statements is unknowable, at a minimum, the availability of the content on the Internet prior to theater release served to generate some measure of negative sentiment toward the film, although some have argued that such discussion raises the visibility of the movie and hence serves to actually *increase* revenue. We also found that the ripping of commercial DVDs represented a relatively minor factor in movie leakage, e.g., only 5% of movies were leaked for the first time via DVD. This result is in conflict with some prior industry speculation about the problem, and calls into question the benefits of DVD player oriented solutions for content piracy control.

Our study was peer reviewed and accepted for publication prior to its release on our web site and publicity in the press. The peer review process helped us improve the paper and added credibility to our results. The study was first reported in a *New York Times* article describing the study, the results, and the implications of the work to the industry [5]. This was succeeded by months of follow up stories debating the causes and results of movie piracy, a discussion in which we would become central figures. Since we released our study, the industry has employed a number of measures consistent with the study. We believe its wide reporting of the story helped to inform not only the effected content providers but also the public.

Our study has had a demonstrable affect on the industry and has helped shape the public discourse on digital content piracy. We have identified several important facets of the problem and proposed possible solutions. Hence, we have succeeding in our core responsibility as scientists: we have armed the public with data to begin to make informed judgments about piracy. This paper is considers how we can be build upon that effort.

II. Studying Piracy

This section outlines our study of insider movie piracy on the Internet. We begin by detailing the methodology and continue by describing our major findings.

A. Methodology

The central artifact we used to study movie piracy was the Internet itself. Using tools we built for this study, we were able to extract not only the movies, but also great deal of other pertinent information about their release onto the Internet. It

[1] A rip is the colloquial term used for a particular copy of a film. *Ripping* refers to the act of digitizing or copying a digital version onto a personal computer for later use or distribution.

0-7803-8622-1/04/$20.00 ©2004 IEEE

was this latter *movie meta-data*, detailed below, that was key to reconstructing the piracy puzzle.

The movie meta-data was collected as follows. Our suite of custom programs accessed publicly available movie and piracy-related web sites. Using this tool, we were able to compile lists of movies that were in the U.S. box office top 50 any time between January 1, 2002 and June 27, 2003. The tools automatically collected and organized meta-data including cinema release date, DVD release date, distributor, MPAA rating, box office take, and some crude popular ratings. We gathered statistics on 409 movies that met our initial criteria. From these, we removed 97 that were first screened or released outside the U.S. or for which the information was unavailable. The subsequent analysis was based on the study of the remaining 312 movies.

We used our software to search an *online content verification* site for unauthorized copies of the studied movies. Content verification sites act as indexes for movies shared on peer-to-peer networks, providing information such as file names, date of first appearance (on the verification site), file size, checksum,[2] and quality. The information on content verification sites is posted and maintained by volunteers, and may not be completely accurate. Furthermore, there is often a delay of several days to a few weeks from the time a movie is first made available on a peer-to-peer network until it is indexed on a content verification site. Note that this delay will cause piracy to be understated, since movies may be listed on the site strictly after they are available in sharing networks. However, use of the content verification site allowed us to obtain data for movies posted over more than an 18-month period without monitoring the peer-to-peer network for that entire period. We limited our search to a single content verification site; querying multiple content verification sites would likely have produced more hits. The content verification site we used usually does not index poor quality copies of movies. Again, this conservative approach tends to underestimate the amount of movie piracy.

The information obtained from the content verification site allowed us to automatically acquire a small part of each relevant copy. We downloaded, on average, about five percent of each movie, which typically represented the first 8 minutes. We further removed from our data set copies for which the content was no longer available, whose files were unusably corrupt, or were discovered to be foreign releases (typically with non-English subtitles). We successfully downloaded and viewed content corresponding to 285 relevant hits for the 312 movies we studied. These hits referenced online copies of 183 distinct movies representing 59% of the movies in our data set. The downloading process acquired about 18 gigabytes of data over about one week using a standard home cable modem.

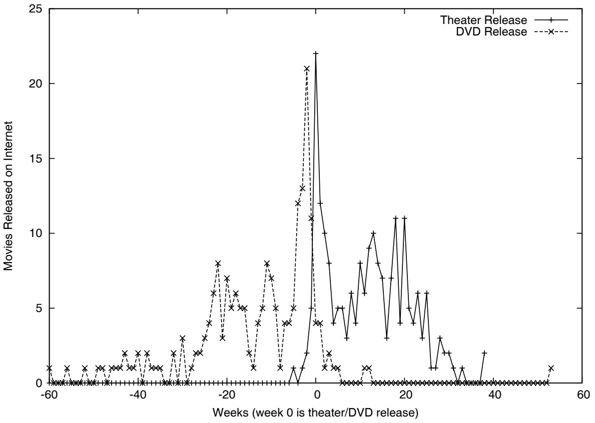

Fig. 2. Theater/DVD Internet release time lag for samples in our dataset. Week 0 represents the time at which the movie was initially released in U.S. theaters or sold in retail stores, respectively.

Once the samples were acquired an automated script delivered the samples to a pool of human observers for judgment, along with a form in which to enter various data. The data recorded included a judgment on video and audio quality along with the presence or absence of the various possible features of unauthorized copies. Some automated analysis methods were performed at this stage. In most cases it was straightforward for the observers to judge the audio and video quality. However, there were 38 samples for which no clear judgment was possible. In most cases their uncertainty was about audio quality, and in all cases we adopted a conservative approach, where it was deemed an outsider if the sample did not exhibit overtly insider sources.

We examined the interaction between release timing, quality, and attack point. For each movie we calculated the time lag between its theater release and its first appearance on the content verification site. If the movie had been released on DVD we also calculated the time lag between the DVD release date and its first appearance on the content verification site.

We classified the attack point as an insider (as opposed to outsider) if: (a) the movie was leaked prior to theater release, (b) the movie had editing room artifacts, e.g., boom-mikes in frame, program counters (see Figure 1), (c) through the air capture with digital audio[3], or (d) a DVD/VHS copy of the movie was leaked before the DVD release date. Any copy not falling into any of these categories was deemed the result of an outsider.

B. Results

A central result of our study showed that more that 58% (183 out of 312) of the movies in the top box office 50 for the studied period were available for download from file sharing networks. While our initial analysis uncovered no correlation between box office receipts for a movie and its availability on the Internet, we believe the question of whether more popular movies are more likely to appear on file sharing networks than

[2] The checksum provides an identifier for each unique copy of a movie uploaded to a peer-to-peer network. All identical copies of the same movie have the same checksum. The checksums are useful for identifying movies, and they allow for client software that can download different blocks of a movie from multiple sources simultaneously.

[3] In this case a cinema employee likely captured the audio directly from the projector and captured the video via a camcorder positioned in the projection booth or in an optimally located cinema seat.

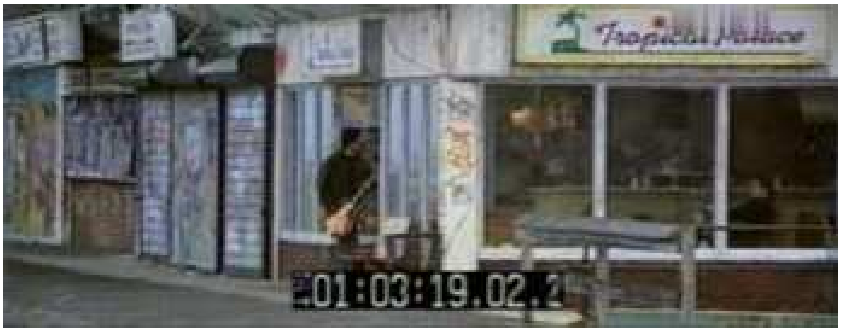

Fig. 1. Editing room artifact – program counter in center-bottom of frame.

less popular movies warrants further study. If popular movies are more likely to be available on the Internet, it might indicate the existence of a piracy economy, complete with its own laws of supply and demand. The more people want a movie, the more they are willing to work with those inside and outside the movie industry to obtain it.

The problem of leakage appears to be ubiquitous: an analysis of the industry showed that leakage occurred at every major movie studio study except Sony Classics.[4] Moreover, the study further showed that of the 285 movie samples we examined, 77% appear to have been leaked from sources within the movie industry (as determined by the criteria we outlined above). The movie samples were first indexed on the content verification sight an average of 100 days after theater release and 83 days before DVD release. While only 7 of these movies were indexed prior to their theater release date, 163 were indexed prior to their DVD release date. The study further showed that only 5% of the movies that had been released on DVD were first indexed after their DVD release date. Hence, this indicated (counter to previous assertions of some in the industry) that the sharing of consumer DVDs on the Internet seemed to play a minor role in movie piracy.

The two significant events in the lifetime of a movie with respect to digital piracy are the theater and DVD releases. Figure 2 shows the distribution of time lags between appearance on the content verification site and these two dates. Note that many movies appear on the Internet within three weeks of their theater release date. This can be attributed to movies being leaked during the production and cinema distribution process and to copies sent to critics and awards judges. Similar leakage occurs about one month before DVD release. Most of those leaks likely originate from DVD pressing plants, DVD distributors, retail employees, or Oscar reviewers; however, some may occur as a result of consumer ripping of DVDs purchased at stores that sell them before their official release date.

78% of the samples in our data set were DVD quality. DVD is the highest quality available for home use, and as such is likely to be the most damaging to the industry. Generally, rips with poorer quality had shorter lag times between theater release and Internet availability. Likewise, those with overt watermarks or textual markers also had considerably shorter lag times. Table I summarizes the movies in our data according to these lag times.[5] Note that many movies will have multiple rips with differing quality on the Internet, e.g., the movie "Behind Enemy Lines" was available in DVD and VHS formats on the Internet.

III. Study Extensions

While the previous study illuminated the problem of piracy, it is far from definitive. We believe a more exhaustive study is needed to better understand the causes and effects of content leakage and further help guide public discourse on the subject. The goal of a second study would be to expand the scope and breadth of our analysis. We propose the following areas for further study:

- Conversations with industry and anecdotal evidence suggest that there is a natural progression to movie leakage. In the absence of an insider pre-theater release, through the air rips might appear on sharing networks immediately after they are shown publicly.[6] Later, higher quality rips may become available from both insider and outsiders. A road map of this process would be helpful in understanding the realities of piracy.
- The analysis of a movie leakage life cycle will also lead to informative statistics of the rips themselves. That is, they will indicate how many versions of the movies are leaked and in what quality. Such statistics will be enormously helpful in understanding the scale of leakage, and indirectly show the ubiquity (or lack thereof) of the problem.
- An unanswered question is how fast content spreads. Recent stories have suggested that sophisticated users are able to obtain the content within hours of a release [8], but it is not clear if this is true of all users. This analysis will be extended to consider the expected timing of a

[4] Sony classics releases films that are not first run, and hence they may be less demand for them.

[5] Features marked with a † represent conclusive evidence by themselves of insider leakage, and those marked with a ⋆ provide conclusive evidence of outsider leakage.

[6] The movie industry strongly believes this scenario occurs with almost every major movie. As evidence of this, they have recently started coordinating the release of major movies worldwide [6], [7].

	Samples	**Theater Internet Lag (days)**	**DVD Internet Lag (days)**
Aggregate Sample Data			
Reviewed Samples	285	100	-83
Insider	220 (77%)	105	-79
Outsider	65 (23%)	86	-96
Sample Features			
Incomplete video editing†	4 (1%)	38	-192
Incomplete audio editing†	1 (<1%)	12	-362
Watermark or text marker†	35 (12%)	52	-141
VHS quality	6 (2%)	60	-149
DVD quality	223 (78%)	123	-62
Through-the-air video	46 (16%)	9	-171
Through-the-air audio*	39 (14%)	10	-171

TABLE I

ANALYSIS OF INDEXED MOVIES.

particular content quality to be made available on the net.

A second unexplored area that warrants study is an analysis of the consumers themselves. While it is clear what people want (the best content quality for free), it is not clear what they will tolerate. For example, would the average home consumer be willing to sit through a through-the-air rip of a movie? Do the viewers that find through-the-air rips acceptable represent more than a tiny fraction of the paying public? If not, then does it matter that such rips are available? Would an average user be willing to pay for a good quality rip and how much? It is answers to these questions that *should* be driving the efforts of the movie and scientific industry. In fact, an attempt to address content piracy either publicly or within the industry without a clear picture of how it is truly affecting consumer behavior seems, at best, premature.

Future studies may be designed to be more efficient. The original study required a significant amount of manual labor. Conversely, to support larger and ongoing analysis, future researchers may wish to automate much of the process. They could expand the scripting capability used in the initial study into a single interface (or applications programmer interface, API) supporting a large number of peer-to-peer clients. This API can be built using open-source client interfaces or through interface scraping tools [9]. The API will not only allow them to streamline the content acquisition process, it will enable the "programming" of other kinds of analysis. For example, it may be possible to "program" the probing of a specific geographic region under scrutiny.

A second avenue for automation is in content analysis. A major physical effort of the first study was simply viewing and evaluating content. Much of that effort can be avoided by the application of the appropriate technology. For example, if a tool were developed to perform optical character recognition (OCR) [10] on the first 10 minutes or so of a movie, the title credits would be captured and compared to preexisting databases. The quality and source of the content could also be automatically assessed. Because of artifacts of the ripping process, multimedia processing techniques can identify with high accuracy whether the audio and video is through the air, VHS, or DVD quality.

The cost of digital content piracy is as yet unknown. The original study shed light on the scope and source of movie piracy, but it represents a single data point. Because of the importance of the problem, we believe that a publicly or privately supported *Internet Piracy Analysis Center* should be established to track and evaluate the state of digital piracy worldwide. The function of that center would be to understand trends and problems with piracy, to advise public policy and industry of the threats and countermeasures to content loss. However, we assert that the parties must be independent from the industries that are served by them. Justified or not, the public has been suspicious of past studies of piracy sponsored by industry. Transparency is of paramount importance.

IV. CONCLUSIONS

The movie industry has initiated a number of programs in an effort to address content piracy. Efforts internal to the industry have focused on tighter control of high quality movies within the production and distribution process. In one sweeping effort, the MPAA initially banned the use of DVD or VHS screeners. Due to enormous pressure from critics and awards judges, this prohibition has since been partially lifted. Industry insiders suggest that these and other measures have been successful in practice. Those claims remain unconfirmed.

The industry has also actively pursed those perpetrating insider leaks. In one case, Carmine Caridi was found to have shared a huge number of movies with an associate Russell Sprague in Illinois. The apparent duplicator, Sprague, was arrested and charged with criminal copyright infringements. Caridi was subsequently expelled from the Academy.

Externally, the industry has sought to curb outsider piracy as well. Through documents such as the *Corporate Policy Guide To Copyright Use And Security On The Internet*, the industry informs enterprises of the legal risks of sharing content [11].

Other efforts have sought to show people the ethical risks and costs of movie piracy [12]. In a move reminiscent of the RIAA piracy stance, the movie industry has recently asserted it will aggressively find and sue individuals who illegally share content [13]. However, it is unclear if and how such legal countermeasures will have an impact on piracy. In particular, the content will remain available in regions where legal devices for pursing content sharing are impotent or non-existent.

In the end, studies such as the one we performed are an essential part of public discourse. Failure to provide independent, unbiased analysis of important social issues often leads to bad policy. Hence, we strongly feel that the trend towards the legislation and litigation of piracy issues should only be supported after a clear understanding of both the features of piracy and the effects of the countermeasures on the served community have been established.

ACKNOWLEDGMENTS

We would like to thank Mike Dickman, our press liaison at AT&T for his help in navigating the press. We would also like to thank Bill Aiello and Avi Rubin for their advice both in the construction and presentation of the study.

REFERENCES

[1] Deloitte and Touche. The impact of piracy on the film industry, June 2003.

[2] Richard Wray. Matrix downloaded: Net piracy could cost film business billions. *The Guardian*, 4 June 2003. `http://film.guardian.co.uk/news/story/0,12589,969754,00.html`.

[3] S. Byers, L. Cranor, E. Cronin, D. Kormann, and P. McDaniel. Analysis of Security Vulnerabilities in the Movie Production and Distribution Process. *Telecommunications Policy*, 28(8):619–644, August 2004.

[4] Lia Haberman. Hulk: It's not easy being CG. *E! Online*, 10 June 2003. `http://story.news.yahoo.com/news?tmpl=story&cid=794&ncid=799&e=1&u=/eo/%20030610/en_movies_eo/11951`.

[5] John Schwartz. Hollywood Faces Online Piracy, But it Looks Like an Inside Job. *The New York Times*, 15 September 2003.

[6] Steven Rea. Studios battling movie piracy. *The Philadelphia Inquirer*, 15 May 2003. `http://www.philly.com/mld/inquirer/news/frong/5864665.htm`.

[7] Andy Seiler and Mike Snider. The movie industry fights off the pirates. *USA Today*, 6 May 2003. `http://www.usatoday.com/tech/news/2003-05-06-movies-piracy_x.htm`.

[8] Amanda Ripley. Hollywood robbery: How does a hit movie go from the free market to the black market? time retraces the trail. *Time Magazine*, 18 January 2004.

[9] Jess Bisbal, Deirdre Lawless, Bing Wu, and Jane Grimson. Legacy information systems: Issues and directions. *IEEE Software*, 16(5):103–111, 1999.

[10] Shunji Mori, Ching Y. Suen, and Kazuhiko Yamamoto. Historical review of OCR research and development. pages 244–273, 1995.

[11] MPAA. Recording and Motion Picture Industries Provide Copyright Use and Security Guide Brochure for Fortune 1000 Companies . *Press release*, 13 February 2003.

[12] MPAA. MPAA Launches New Phase of Aggressive Education Campaign Against Movie Piracy. *Press release*, 14 June 2004.

[13] Associated Press. Movie trade group files suits over piracy. *The New York Times*, 17 November 2004.