UNIVERSITY OF CALIFORNIA, SAN DIEGO

A New Look at the DMCA Section 512 Takedown Process

A Thesis submitted in partial satisfaction of the
requirements for the degree of
Master of Science

in

Computer Science

by

Gautam Akiwate

Committee in charge:

Professor Geoffrey M. Voelker, Chair
Professor Stefan Savage
Professor Lawrence K. Saul

2015

The Thesis of Gautam Akiwate is approved and is acceptable in quality and form for publication on microfilm and electronically:

____

____

____

Chair

University of California, San Diego

2015

DEDICATION

To my family. It’s done.

EPIGRAPH

...
Answers in the unknown will I find?
Is the world black and white
Or is it a murky grey sight?
Morals and ideals are yours or the world's?
Perception is but an idea of yours!
Reality is but the thoughts in your head
So is bad good and good bad instead?
The world is who decides these laws
...

Shades of Grey, *Shivani Malpani*

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

No journey can ever be completed without the support of countless people throughout that journey, not all of whom may be apparent. This is no exception. While I cannot hope to enumerate all those who supported me since there are bound to be omissions, I'll be remiss if I didn't mention a few at least.

First, I would like to express my sincere thanks to my advisor Geoffrey Voelker for being an immense source of inspiration and encouragement all along. His outlook towards research has helped me develop valuable insights and perspectives for my research. I could not have imagined having a better advisor for my research work. Besides my advisor, I would also like to thank Stefan Savage and Lawrence Saul for their guidance and insightful comments.

I would like to thank Mohit Kothari, who also worked on this project for being a great friend and also an excellent roommate. Bouncing around ideas and working with him never ceased to be exciting. Thanks to Rishi Kappor, Sen Zhang and Long Jin - my office mates, for being ever helpful.

I cannot even begin to properly thank my family whose support and advice has been instrumental to everything I have done and will do. My friends, who have always been there, despite my griping, to help and encourage me. And Purvi Desai, who patiently helped me revise my thesis.

I would also like to thank Tristan Halvorson for introducing me to Spark and also helping me understand how the crawler works. Finally, I would like to thank the 'Systems and Networking' group for providing an encouraging environment for research while cultivating an environment of openness.

All chapters, in part are currently being prepared for submission for publication of the material. Akiwate, Gautam; Akiwate, Gautam S. The thesis author was the primary investigator and author of this material.

ABSTRACT OF THE THESIS

A New Look at the DMCA Section 512 Takedown Process

by

Gautam Akiwate

Master of Science in Computer Science

University of California, San Diego, 2015

Professor Geoffrey M. Voelker, Chair

The Internet has become an engine for growth and innovation and has been responsible for the heretofore unprecedented scale of sharing and access to information. However, today the Internet is also used for the illegal exchange of copyrighted materials. The fast changing nature of the Internet meant that the traditional copyright law was ill-equipped to handle this infringement. The Digital Millennium Copyright Act was introduced as an attempt to address the changing needs of the copyright law. One of the salient additions of the law was an *extra-judicial* mechanism to remove allegedly infringing content.

This thesis examines the takedown process in practice in a structured fashion. As part of this work, we analyze the takedown notices made available to the public, in particular the Chilling Effects repository. We also run active measurement studies to understand the actions and reactions of the system as a whole. In addition to this, we address questions about the DMCA Section 512 process and the entities involved in this process. Finally, we also make specific observations and recommendations in an attempt to structure the future work in addition to summarizing the results of this work. In particular, we find that while the process is not able to take down *all* the infringing content, it is able to make it less accessible to a casual user.

Introduction

The Internet is an essential engine for growth and innovation. The expansion of the Internet has led to a heretofore unprecedented scale of sharing and access to information. However, the use of the Internet to illegally share copyrighted songs, movies, TV shows, software and books prompted changes to the existing law in an effort to curb this sharing since this illegal sharing is perceived as a major threat by copyright holders [21][25]. This illicit sharing may take many forms like streaming, file downloads or file sharing through peer-to-peer (P2P) protocols.

One response to address the issues with the traditional copyright law was the Digital Millennium Copyright Act (DMCA). Specifically, the DMCA Section 512 process was introduced for taking down allegedly infringing content. This process uses takedown notices as the principal mechanism for removing illicitly shared copyright content on the Internet in the United States [21]. The process is novel since it bypasses judicial oversight over these takedowns. This extra-judicial process leads to obvious concerns of abuse. Further, given the influence of the DMCA on the laws passed in other countries [22], the study of a process that so affects the inherent fabric of the Internet is important.

This thesis examines the effect that the DMCA Section 512 process has had on the Internet in practice. Expanding upon previous work in the area [16, 18, 22], we conduct an empirical analysis on the takedown notices obtained from the Chilling Effects repository. In addition, we also perform an active measurement study to supplement our understanding of the takedown process. To address the challenge of the growing size of the dataset we run our analysis as MapReduce tasks on a Hadoop cluster. Additionally, we also extend our analysis to the URL takedown requests and use this to make indirect inferences about takedown notices submitted to other service providers.

On the whole we find that the Section 512 process is not effective in taking down content since sites have evolved strategies to keep their content available. However, the joint effort of making this illegally shared content less accessible by Google and copyright holders is still effective. The analysis of takedown notices shows three main trends: increasing number of notices submitted, increasing number of URL takedown requests and the increasing number of URL takedown requests filed in a single notice. Additionally, our work shows that there are a few major copyright holders who dominate the takedown process and account for the majority of the takedown notices and the URL takedown requests submitted. Further, we also discuss the importance of expanding the analysis to the URL takedown requests and the issues faced during the analysis of these URL takedown requests. In particular, for links that were requested to be taken down, the determination of the *liveness* of these links becomes surprisingly involved and complicated. Finally, we discuss our concerns about the process before summarizing and briefly discussing issues that need to be tackled as future work. Specifically, we discuss the need for designing scalable processes for analysis along with larger scale active measurement studies to supplement our understanding of the takedown process.

The thesis is organized into eight chapters. Chapter 1 briefly introduces the DMCA Section 512 process and its provisions. Chapter 2 describes the datasets used in the thesis and previous work related to our study. Chapter 3 details my active measurement study and its results. Chapter 4 and 5 describe the results from the analysis of the datasets, while Chapter 6 lists the issues faced during the analysis and the active measurement study. Chapter 7 describes some key observations and patterns that arise from the analysis. Finally, Chapter 8 summarizes the work and concludes with a brief discussion.

Chapter 1

A Song of Ice and Fire

The Digital Millennium Copyright Act, commonly referred to as DMCA, was signed into law in 1998, while the law came into effect in 2000 [2, 26]. DMCA was primarily aimed at updating the U.S. Copyright Law to address the changing needs of the copyright law given wide-spread use of the Internet, as the advent of the Internet had complicated the application of the traditional copyright law [2].

The DMCA also addresses other significant copyright-related issues and is divided into five titles [26]. The focus of this thesis is the Online Copyright Infringement Liability Limitation Act (OCILLA), the Title II of DMCA. The Title II of DMCA modified the U.S. Copyright Act to add a new section, Section 512 [26], which puts into place limitations on the liability of Online Service Providers (OSPs) for copyright infringement while also laying down conditions for the OSPs to qualify for this *safe harbor*.

This chapter provides a high-level overview to the relevant sections of the DMCA legislation before discussing how it is currently being *used or misused*. However, before we discuss the law we first present the reasons and the motivations behind the law.

1.1 The Background

Primarily, the need to amend and update U.S. Copyright Law arose due to the changing nature of technology. In particular, the advent of the Internet had changed how people viewed and shared information. Perhaps what was even more important was that, not only did the Internet change the way people viewed and shared information, but also how easy and accessible all of it was made. The ease of copying and sharing information in the digital world meant that the old copyright laws had a hard time keeping up.

The ease of sharing information invariably led to sharing of copyrighted material. Further, given the involved nature and the involvement of multiple parties in the eventual sharing of the copyrighted information via digital media the question of liability quickly became complicated. As per traditional copyright laws the Online Service Providers (OSPs), which included the Internet Service Providers (ISPs), search engines and others, were faced with secondary liability based on their customers' copyright infringement [1, 16]. Before DMCA was adopted, the OSPs faced a high degree of uncertainty as a result of the conflicting interpretations of when they could be held liable [16].

Thus, on the one hand OSPs lobbied for a safe harbor from the secondary liability while on the other hand the copyright holders wanted the OSPs to be held responsible for their networks and software and in effect police them to be qualified for any sort of safe harbor. In fact, the Working Group on Intellectual Property Rights that was responsible for the initial recommendations opined that the OSPs should be considered responsible for policing information and that the business relationship between OSPs and Internet users would justify the risk and cost of liability on OSPs [16].

Thus, the Online Copyright Infringement Liability Limitation Act (OCILLA) was in essence a compromise between the OSPs and copyright holders [16]. The legislation tried to strike a balance between the concerns of the copyright holders and that of the OSPs. The result was the creation of a process that allowed the copyright holders to request the rapid removal of allegedly infringing material while guaranteeing safe harbor to OSPs that complied. The resultant process thus created is the primary focus of this work.

Given its history, it is not surprising that the legislation has been received with mixed reviews since its inception. However, there seems to be consensus that the safe harbor provisions, even though debated, have been essential to the growth of the Internet and integral to keeping it the engine for innovation and free expression that it represents today [7].

1.2 OCILLA: An Overview

The Online Copyright Infringement Liability Limitation Act (OCILLA), as mentioned before, was codified as Section 512 of the US Copyright Act [16, 26].

1.2.1 Liability Limitation

The legislation creates four new limitations on liability for copyright infringement by OSPs. The limitations are based on the service that the service provider offers. The four categories are:

1. **Transitory Communications**

   This includes service providers who act as a data conduit by providing transmission, routing or connections for data at the request of a person other than the service provider. An example of a service provider that could qualify under this category would be an Internet Service Provider (ISP).

2. **System Caching**

   This includes service providers who cache information to improve network performance, provided the information stored is unmodified. An example of a service provider that could qualify under this category would be a Content Delivery Network (CDN) like Cloudflare or Akamai.

3. **Storage of Information on Systems or Networks at directions of users**

   This includes service providers who provide storage to users, provided that they are not aware of the infringing material and do not directly benefit from the infringing activity. An example of a service provider that could qualify under this category would be a website hosting service.

4. **Information Location Tools**

   This includes service providers who provide services and tools to locate information on the Internet. This primarily covers search engines like Google and Bing who link and refer users to content.

Limitation of Liability: Eligibility

To be eligible for any of the liability limitations, the service providers in addition to qualifying as a *service provider*, must satisfy two additional conditions [17, 26].

1. A service provider must adopt and reasonably implement a policy of terminating the accounts of subscribers and account holders of the service provider's system or network who are repeat infringers under appropriate circumstances.

2. A service provider must accommodate and not interfere with standard technical measures used by copyright holders to identify and protect copyrighted works.

1.2.2 Service Provider: A Formal Definition

For the purposes of this work it is beneficial to formalize the definition of a service provider as used in the legislation.

A service provider is interpreted and defined differently based on the service that the service provider offers. For transitory communications, a service provider is defined in Section 512(k)(1)(A) as *"an entity offering the transmission, routing, or providing* *of connections for digital online communications between or among points specified by a user, of material of the user's choosing, without modification to the content of the material as sent or received."* For the other limitations, the definition is broader and is defined in Section 512(k)(1)(B) as *"a provider of online services or network access, or the operator of facilities therefor"* [17, 26].

1.3 The DMCA Section 512 Process

The Section 512 was responsible for creating an extra-judicial process to allow for expeditious take down of infringing material. Section 512 was a compromise between copyright holders and the OSPs. The prevalent concern of the copyright holders was the co-operation of the OSPs and hence sought to have enough incentive for the OSPs to remove the infringing material. As a result, the *safe harbor* is only granted to OSPs in exchange for the expeditious takedown of allegedly infringing material upon receipt of a notice from the copyright holder [16].

The legislation creates several categories of protection based on the service provided by the OSP as mentioned in Chapter 1.2.1. For the service providers that fall under the transitory communications category, the protection as provided by Section 512(a) is the broadest. Since these service providers act as simple conduits of information there is no requirement to remove infringing material [16].

On the other hand, OSPs that qualify under the system caching category limitation liability provided by Section 512(b), that is service providers who cache information to improve network performance are required to respond and remove or disable access to allegedly infringing material when certain conditions are met. For these conditions to be met, the material from the originating site must have been removed or at least must have been ordered to be removed. It is expected that the copyright holder filing the notice must also give a notification confirming the same to the service provider [16, 17].

Finally, the OSPs that qualify under the storage of information and information locating categories as provided by Section 512(c)-(d), that is primarily hosting services and search engines, are required to respond *expeditiously* to notices of copyright infringement by removing hosted content or links to content upon receipt of notice.

1.3.1 Replacement of Removed or Disabled Material

The Section 512 process is in essence an extra-judicial process. Hence, both universities and libraries were concerned that this extra-judicial process for removing users' material from the Internet violated constitutional provisions for due process. Another concern was that the process gave the OSPs strong incentive to maintain safe harbor while few incentives to question the takedown. In an attempt to avoid this, procedural protections were added in the form of Section 512(g) which created a counter notice procedure for users to challenge the removal of their material. Additionally, it also requires that the OSPs who qualify under Section 512(c), which are primarily hosting services, to establish and maintain a process by which on receipt of a statutorily compliant notice the OSP must not only take down the material but also notify the alleged infringer that the material has been removed. Additionally, the OSP needs to also forward any counter-notices from alleged infringers back to the complainant. Finally, if no action has been taken by the complainant within 10-14 days after the counter-notice has been filed, the OSP has to reinstate the material.

OSPs that provide search engine services are required to *expeditiously* remove links leading to allegedly infringing materials. However, they are not required to notify the alleged infringer of the removal since the service provider likely has no service relationship with the alleged infringer.

The OSPs are exempt from the liability, as per Section 512(g)(1), of a mistaken yet good faith removal of material based on a notice from a copyright holder.

Other Relevant Details

The DMCA Section 512(h) authorizes copyright holders to request the clerk of any United States district court to issue a subpoena to a service provider to identify an alleged infringer. The OSP is expected to *expeditiously* disclose a user's personal identifying information in accordance with a subpoena as long as it is accompanied by a valid notice [16, 17].

Interestingly, the subpoena process does not apply to the service providers who qualify under the Section 512(a) due to a ruling made in the RIAA vs Verizon case in 2003 [4].

Finally, an alleged infringer can file a counter-notice to the service provider upon receipt of a notice. To be considered valid the counter notice must include the following:

1. A physical or electronic signature.

2. Identification of the material removed and its former location.

3. Statement under penalty of perjury that the user has a *good faith belief* that the material was mistakenly removed.

4. The user's name, address, and phone number.

5. Consent to the jurisdiction of Federal District Court.

Chapter 2

A Feast for Crows

The DMCA Section 512 is a simple extra-judicial process that affords the copyright holders to remove allegedly infringing material on the Internet [16]. Surprisingly, there is very little research done on the takedown process given that it affects the Internet in a real way. One reason for this could be that, since takedown notices are not required to be part of public record [2, 17], meaningful research on the takedown process becomes hard. Thus, an effective study of the process is contingent upon getting a useful dataset to use as a ground truth from which we can base our analysis or experiments.

In this chapter, we briefly discuss the related work done on takedown notices and use it to motivate our study. We then discuss the datasets that were used in this work, what they include and how we use them.

2.1 Related Work

To date, there have been three studies examining takedown notices. The first was by Marjorie Heins and Tricia Beckles on 153 notices in the Chilling Effects repository submitted up to 2004 [18]. The second was the seminal study by Jennifer Urban and Laura Quilter [16]. This study undertook the analysis of 876 notices, in the Chilling Effects repository, submitted up to August 2005. The third study by Daniel Seng analyzes over 539,000 notices, in the Chilling Effects repository, submitted between January 2001 and December 2012 [22]. This work undertakes the analysis of the notices that were submitted from January 2010 to September 2014.

While the first two studies concerned few notices, one of the pressing concerns of the 2012 study was the sheer volume of the notices, which made a rigorous empirical study of these takedown notices hard [22]. To overcome this issue, the 2012 study used crawlers and automated parsers in conjunction with a MySQL database to store the data. Unfortunately, between 2012 and late 2014 the number of notices submitted jumped to more than 2.5 million notices. Hence, for our study to process this increased volume we use a Hadoop cluster to run analyses.

Apart from this, there are three major points that set this work apart from this previous work. First is the use of active measurement to supplement our observations from the passive empirical analysis of the notices themselves. Secondly, we try and expand our analysis to indirect inferences about the takedown process to other service providers. Additionally, we also expand our analysis to the individual URL takedown requests.

2.2 Chilling Effects

Chilling Effects, a project of the Berkman Center for Internet & Society at Harvard University, aims to address the gaps in how the process is being used [10]. It is a collaboration among law school clinics and the Electronic Frontier Foundation. The project collects takedown notices from a variety of sources. Note that for the rest of the thesis we use the terms *takedown notices* and *notices* interchangeably.

2.2.1 Sources of Data

The Chilling Effects project collects data from various sources. Figure. 2.1 shows the number of notices that were submitted over months to the Chilling Effects repository

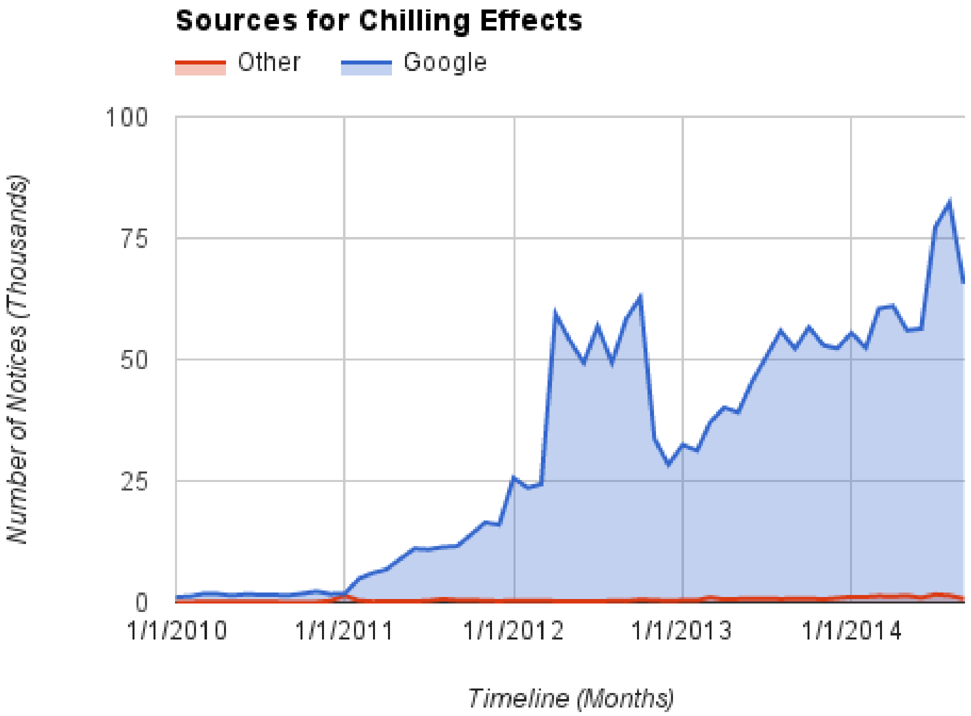


Figure 2.1. Sources of Notices for Chilling Effects

by Google as compared to other sources. From Figure 2.1 it is evident that Google accounts for majority of the notices. Twitter, Yahoo, Digg have been other sources of notices over time. Figure 2.2 shows the takedown notices submitted over months by Twitter, Yahoo, Digg and other sources excluding Google.

2.2.2 Crawling the Notices

To analyze the takedown notices, we crawled the Chilling Effects repository. The crawler runs on a Hadoop Cluster and stores the notices as records in the Hadoop Distributed File System.

Note that after 2014, the Chilling Effects project started using a different API. However, for the purposes of this thesis we restrict analysis to the takedown notices that were submitted using the old API and go all the way till September 2014. Additionally, there were a few notices which were ignored since they did not have a date associated

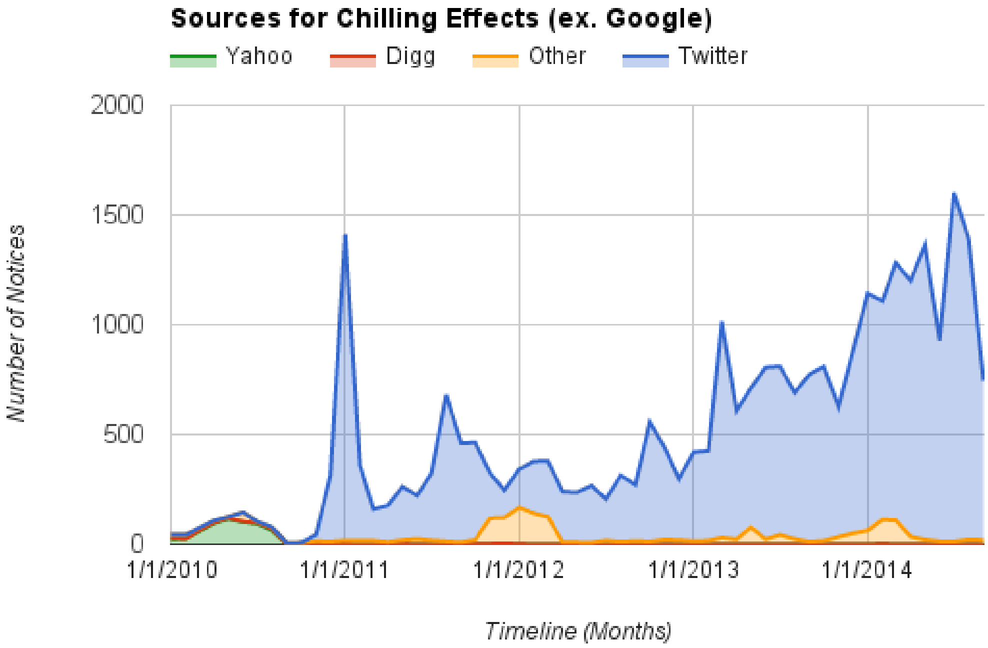


Figure 2.2. Sources of Notices for Chilling Effects (Excluding Google)

with it.

Table 2.1. Summary of Chilling Effects Dataset

	All	**2010-2014 (%)**
# of Takedown Notices	1,896,287	1,807,757 (95%)
# of URL Takedowns	493,559,330	484,557,918 (98%)

Chilling Effects allows the takedown notices in their repository to be viewed and downloaded in various formats. We chose the JSON format since its structured format makes it easier to write analysis scripts. We crawled a total of 2,056,606 notices. However, not all were DMCA takedown notices and are excluded from the analysis. Table 2.1 summarizes the number of notices and the accompanying number of URL takedown requests in the entire Chilling Effects repository and those submitted from 2010 through 2014. The time frame from 2010 through 2014 effectively captures 95% of the notices and 98% of the URL takedown requests and is the focus of this thesis.

2.2.3 Analysing the Notices

One of the prime considerations was the large number of notices. Given the convenient JSON format, we decided to use Spark which is a fast and general processing engine that is compatible with Hadoop data [23]. Given that our crawler already stores data on HDFS, running Map-Reduce style jobs on the data seemed like an obvious solution. Spark also offered an easy interface to write analysis scripts.

2.2.4 Real World Data: Issues

One major problem dealing with real-world data is that of *incomplete* or *corrupted* data. In this case too we faced similar issues since takedown notices in the repository sometimes had missing fields or insensible data. Given the scale of the data we are processing we omit notices that have *corrupted* fields. Note, we only omit a takedown notice if the field we are interested in is corrupted.

2.2.5 Structure of a Takedown Notice

The contents of a valid takedown notice were briefly discussed in Chapter 1. The elements of a valid takedown notice are laid out in Section 512(c)(3)(A).

We briefly discuss the structure of a takedown notice as it appears in the Chilling Effects repository. Listing 1 shows the structure of typical takedown notice.

A brief explanation for the important fields is as follows:

1. **Principal Name**

   Organization or person that holds the copyright.

2. **Sender Name**

   Organization or person that sent the notice.

3. **Recipient Name**

   Service Provider to whom the notice was sent.

4. **Works**

   These contain the *claims* which contain the list of URL takedown requests. There may be multiple claims per takedown notice.

5. **Infringing URLs**

   These are the URLs that the copyright holder is requesting to be taken down.

2.3 Google Transparency Report

The Google Transparency Report is an effort by Google to document the information specified in requests received from copyright holders through their Web form to remove search results that link to allegedly infringing content [14]. The Google Transparency Report refers to the Chilling Effects repository for the actual notices while it stores metadata, annotations and the particular actions taken for each of the notices. This additional information becomes key to fully understanding the information from the takedown notices that we analyze from Chilling Effects.

The data includes information for more than 95% of the copyright removal requests that Google has received for Google Search since July 2011. However, it does not include:

1. Requests submitted by means other than their web form, such as fax or written letter

2. Requests for products other than Google Search (e.g; requests directed at YouTube or Blogger).

3. Requests sent to Google Search for content appearing in other Google products (e.g; requests for Search, but specifying YouTube or Blogger URLs).

For the purposes of this thesis, we utilize the Google Transparency Report as an additional resource rather than as a dataset. For example, we use it to choose domains from which to monitor links in the active measurement setup as seen in the next chapter. However, this is still an incredibly rich resource which could potentially be leveraged for a lot of interesting analysis in the future. The main value addition of this lies in the additional annotations and notes that Google adds with each request. Following is a summary of the additional information that Google typically adds to each notice:

1. **Percentage of Domain specified by request**

   Ideally, we could use these domains to understand how domains react to takedown notices. Specifically, domains that have a substantial portion of their domain taken down by these notices. Potentially, we might want to look at these domains and understand what type of site they are in addition to also monitoring how long the domain stays active.

2. **URLs from this domain not removed**

   This lists the URL takedown requests from notices that Google did not comply with. Essentially, if a link is not taken down, it could possibly mean three things:

   (a) Duplicate Request

   (b) Link not in Google Index

   (c) Invalid Takedown Request

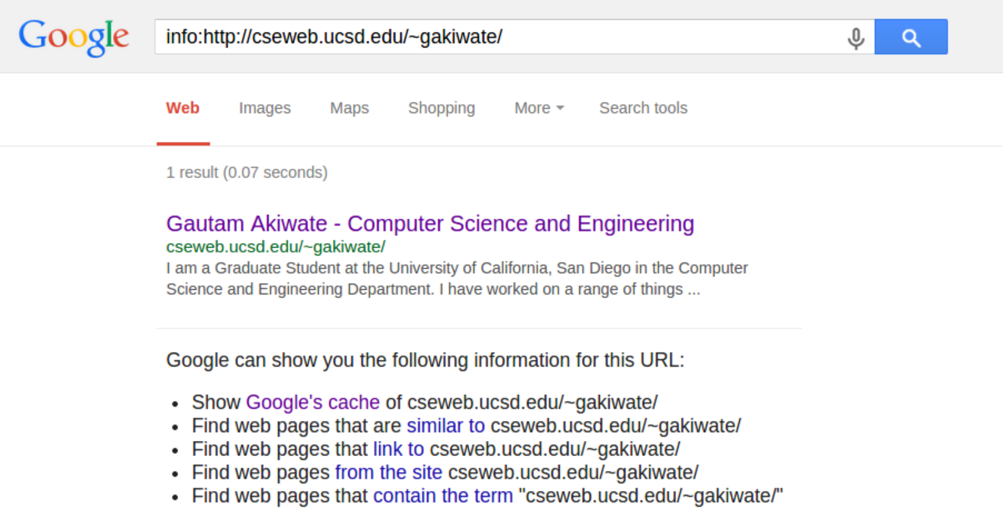
2.4 Search Engine Index

The Chilling Effects database is primarily dominated by notices from Google. Given that notices sent to Google form the bulk of our dataset, we would ideally like a way to inspect the Google search index. Specifically, we would like to know if a particular link is in the index or not. However, as mentioned before we would also like to make indirect inferences about notices being served to other search engines, particularly Bing.

Considering the number of links that we would like to study and the fact that it is an experiment that we need to repeat over multiple days, we wanted to automate the whole pipeline. To achieve this we re-purposed infrastructure that had been built for another project which could query both Google and Bing and fetch the results [5]. However, in this case we were only interested in determining if the said link existed in the index or not. Refer to Appendix A.1 for a detailed explanation as to how we determine if a link is in the index or not for each of the search engines.

2.4.1 Searching the Google Search Index

Provided a link, the *Google Search* keyword *info* returns information that Google has on the link. Figure 2.3 shows an example of how an *info* query might look. The assumption is that if the search with the link prefixed by the *info* keyword returns a result then it exists in the index. If no result is returned, there are two possibilities. First, it was never indexed before. Second, it had been indexed but has since been removed from the index. Hence, knowing when the content became live in the search index is a helpful aid to make a distinction between the two.



info:http://cseweb.ucsd.edu/~gakiwate/

1 result (0.07 seconds)

Gautam Akiwate - Computer Science and Engineering
cseweb.ucsd.edu/~gakiwate/
I am a Graduate Student at the University of California, San Diego in the Computer Science and Engineering Department. I have worked on a range of things ...

Google can show you the following information for this URL:

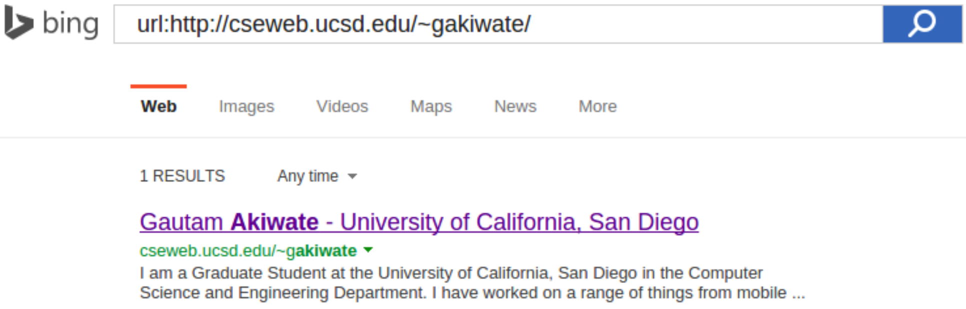
- Show Google's cache of cseweb.ucsd.edu/~gakiwate/
- Find web pages that are similar to cseweb.ucsd.edu/~gakiwate/
- Find web pages that link to cseweb.ucsd.edu/~gakiwate/
- Find web pages from the site cseweb.ucsd.edu/~gakiwate/
- Find web pages that contain the term "cseweb.ucsd.edu/~gakiwate/"

Figure 2.3. Example of the Google Search keyword info

2.4.2 Searching the Bing Search Index

This measurement setup was similar to the Google index search. To search the Bing index we used the Bing API along with the search keyword url which worked similar to the search keyword info in Google.

Figure 2.4 shows an example of how an url query might look. Similar to Google, we make an assumption that a query prefixed by the url keyword returns a result only if the link exists in the search index.



url:http://cseweb.ucsd.edu/~gakiwate/

1 RESULTS

Gautam **Akiwate** - University of California, San Diego
cseweb.ucsd.edu/~**gakiwate**
I am a Graduate Student at the University of California, San Diego in the Computer Science and Engineering Department. I have worked on a range of things from mobile ...

Figure 2.4. Example of the Bing Search keyword url

```
{
    "dmca": {
        "id": "<Reference ID>",
        "type": "<DMCA>",
        "title": "<Title>",
        "body": "<Body>",
        "date_sent": "<Date Sent>",
        "date_received": "<Date Received>",
        "topics": [
            "<Copyright>",
            "<DMCA Safe Harbor>"
        ],
        "sender_name": "<Organization that filed the notice>",
        "principal_name": "<Copyright Holder>",
        "recipient_name": "<Service Provider>",
        "works": [
            {
                "description": "<Description of work that was infringed>",
                "infringing_urls": [
                    {
                        "url": "url #1"
                    },
                    {
                        "url": "url #2"
                    }
                ],
                "copyrighted_urls": "<URL to original content>"
            }
        ],
        "tags": [ ],
        "jurisdictions": "<Jurisdiction>",
        "action_taken": "<Yes/No>",
        "language": "<Language>"
    }
}
```

Listing 1. Structure of a DMCA Takedown Notice in Chilling Effects

Chapter 3

A Game of Thrones

The Section 512 process is the principal mechanism for copyright enforcement on the Internet [21]. Hence, understanding the use of the mechanism in the real world is of particular interest. Chapter 1 discussed the law and the rationale behind it. However, while understanding the law and rationale behind it is important, it is perhaps even more important to understand how the process is actually being used today, and contrast with its use over the years.

A particularly challenging problem is that any measurement or analysis can only be done on takedown notices which were made available after the action had been taken. Hence, while we could analyze the notices, observations or analyses based on the URL takedown requests in the notices was significantly tougher. By the time the notices were released to the public domain, action had been taken on the notices and hence, the early dynamics of the process were missed. Since one of the goals of the work was to understand how the entire process works and how the users and copyright holders react to takedown notices, we also performed an active measurement case study.

Ideally, we would like an active measurement study to shed some light on the following questions:

1. How responsive are copyright holders? How long does it take them to start sending

takedown notices?

2. How responsive is Google in taking down links after they receive a takedown notice?

3. What about other search engines? Could we make indirect inferences about how copyright holders targeted search engines other than Google?

4. Similarly, how responsive are content hosting services? Since we cannot see notices that are served to these service providers, can we estimate the responsiveness to a fair degree of certainty?

5. If the content remains up, what could the possible reasons be?

6. On average, how long does content usually remain up before it gets taken down?

7. How has the Internet been affected by these notices, if at all?

While these are questions that we would *like* to answer, given the nature of the process and our limited pool of information it should be understood that an active measurement, while giving interesting insights into the process, might not help us get representative figures or numbers for most of the questions. Nevertheless, this study promised a new perspective to begin answering, if not completely answer, the questions posed above.

Given the nature of the problem, it is infeasible to exhaustively monitor everything. Hence, the active measurement was based on a set of links that led to infringing material. Additionally, to make reasonable conclusions and observations about responsiveness, we also know the particular date on which the links went up. In this case study, our active measurement methodology consists of choosing a set of links and then determining on a daily basis if the links are indexed by Google or Bing. We also crawled the HTML page of the links to keep track of the changes in content and availability of the link. This was then correlated to the takedown notices being submitted for the chosen set of links.

It is important to note that the measurement is contingent on the fact that the links that we monitor eventually get acted upon. Ideally, we would like to *predict* beforehand the links that would be taken down. However, since we do not have access to an *oracle*, we make do with links that have a high probability of being taken down.

3.1 The Game of Thrones Piracy Resurgence

Game of Thrones is said to be the *most pirated TV show* and has broken records in the piracy world [12]. On April 12th, 2015 just before the *Season 5* of *Game of Thrones* was to be premiered, the first four episodes of the season were leaked online.[1] Given that, we considered this leak to be a tailor made situation for the application of the Section 512 takedown process.

More importantly, this was an ideal event to base the active measurement study around. The popularity of *Game of Thrones* meant that we would have a chance to monitor links that shared the content through multiple ways. Given the publicity that the leak received it was evident that *Home Box Office*, the copyright holder for the *Game of Thrones*[2] series, would pursue this leak and take down infringing URLs.

3.1.1 The Measurement

The goal of the measurement was straightforward: observe the takedown process in work. Note that a link getting *taken down* in the context of a search engine means that the link was *removed from the search index*.

[1] http://www.bbc.co.uk/newsbeat/article/32275599/game-of-thrones-season-five-leaked-online

[2] http://www.hbo.com/game-of-thrones/

The Measurement Approximation

The leak of the *Game of Thrones* episodes provided an excellent opportunity to perform active measurements. Since, exhaustively monitoring all possible infringing links was infeasible for the purposes of the measurement we chose a set of representative 30 links that had content which *potentially* could be taken down.

To determine a set of links we used two main guidelines. First, to ensure that the links covered the three popular methods of sharing content on the Internet, namely:

1. File Sharing Links / Cyberlocker
2. P2P File Links
3. Streaming Links

Second, since our experiment is contingent upon action being taken against the links monitored, we need to ensure that there is a high probablity that the links get taken down. To ensure this, we sought links in two categories.

1. Links through searches. This was a way to mimic how a user would typically obtain links. We used 20 links at the top of search results since it implied that they had a high probability of being taken down.

2. Links belonging to top-level domains that have been the subject of multiple takedown notices by Home Box Office as submitted to Google as per the Google Transparency Report [14]. We obtained another 10 links that led to *allegedly* infringing material, specifically the leaked episodes from two such top-level domains.

The Measurement Elucidation

After finalizing the set of 30 links the question then becomes, what should the scope of the measurement be. A straightforward one is that we get the HTML pages that the links lead to. Since our only major source of notices is the takedown notices sent to

Google, another obvious choice was to track the *liveness* of the link on the Google search index. Here, by *liveness* we mean if the link is in the search index or not.

3.1.2 The Google Search Manifestation

The main task was to see if a link was in the Google search index or not. The high level idea is that, given a set of links, hopefully most of which are indexed, the goal then would be to see when these links are removed from the search index. More accurately, can we track the liveness of the links over days and then, once they are *de-indexed* (assuming they are *de-indexed* due to the takedown process), find the relevant takedown notice that took down the link from the Chilling Effects repository.

Searching the Index

Starting April 12th, 2015, we checked daily the liveliness of the 30 links in the Google search index. We then tracked the links over time to trace their liveness. At a high level, we were interested if they become indexed in case they were not initially indexed, if Google removed them from the index, or if Google re-indexed them after having removed them from the index. Figure 3.1 shows the state of the set of 30 links over time. For the purposes of the measurement here, since the links could only have come up on April 12th, 2015, we assume that all the links were indexed on that day.

For the purpose of analysis, we break down the links into the following categories:

(a) **Initial**

Links that were indexed when we made our first measurement.

(b) **Never Indexed**

Links that were not indexed from the day of the first measurement to the day of the last measurement.

(c) **New**

Links that were not initially indexed but were newly indexed for the first time.

(d) **Removed**

Links that stop showing up in the search index *after* having shown up in the index previously.

(e) **Re-Indexed**

Links that were *removed* but then for whatever reason were re-indexed and showed up again.

The Link Attenuation

As can be seen from Figure 3.1, action begins to be taken on the links on April 14th, 2015. In just 10 days, a total of 21 links were removed from the search index. As of April 30th, 2015 a total of 26 links were removed and 1 link was never indexed. Table 3.1 provides a brief summary of the takedown notices that were responsible for taking down the links. It should be noted that, a link being *taken down* from the search index has no bearing on the site or the content on the site. However, it is reasonable to assume that the copyright holder will also submit a similar takedown notice to the content hosting service provider that hosts the site when it submits a takedown notice to Google.

The Reindexed Links Conundrum

Figure 3.1 shows links being re-indexed after having been taken down. Why do these specific links get re-indexed? What could have led to the crawler re-indexing the page after it had been explicitly removed? Moreover, we are certain that these are the links that were taken down since we can verify this using the takedown notices submitted to Chilling Effects by Google.

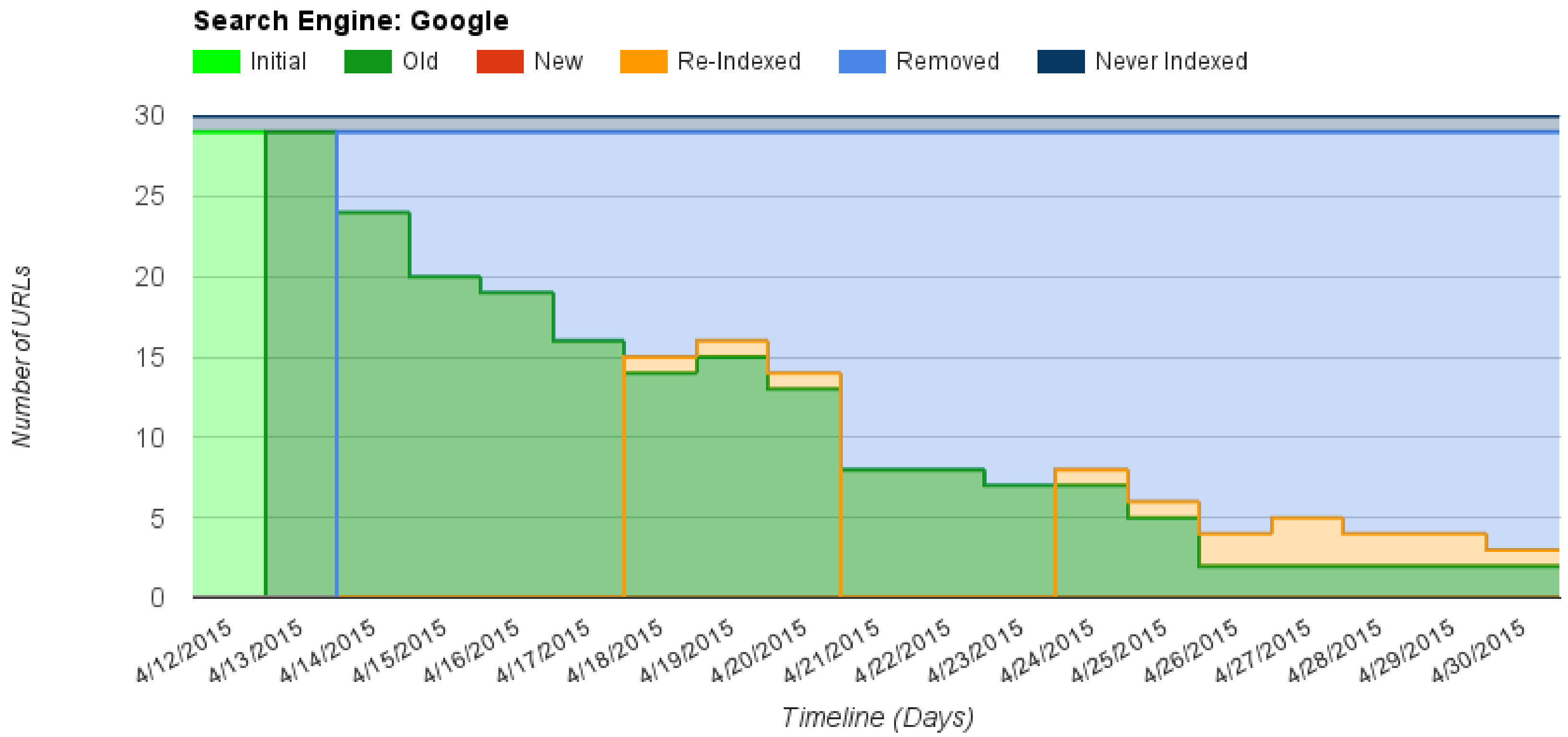


Figure 3.1. Google Search Index: Link Status

Table 3.1. Summary of Takedown Notices Against Monitored Links

URL	Notice ID	Date	Sender Name
1	10664731	April 15th	MarkMonitor AntiPiracy
2	10660600	April 14th	MarkMonitor AntiPiracy
3	10704954	April 26th	MarkMonitor AntiPiracy
4	10668599	April 16th	MarkMonitor AntiPiracy
5	10660600	April 14th	MarkMonitor AntiPiracy
6	10660600	April 14th	MarkMonitor AntiPiracy
7	10690259	April 22nd	MarkMonitor AntiPiracy
8	10712053	April 28th	MarkMonitor AntiPiracy
9	10661834	April 14th	IP-Echelon Pty Ltd
10	10661834	April 14th	IP-Echelon Pty Ltd
11	`<dns error>`	-	-
12	10663526	April 14th	IP-Echelon Pty Ltd
13	10682726	April 20th	MarkMonitor AntiPiracy
14	10704954	April 26th	MarkMonitor AntiPiracy
15	10704954	April 26th	MarkMonitor AntiPiracy
16	10682726	April 20th	MarkMonitor AntiPiracy
17	10682734	April 20th	MarkMonitor AntiPiracy
18	10686672	April 21st	MarkMonitor AntiPiracy
19	10660600	April 14th	MarkMonitor AntiPiracy
20	10686672	April 21st	MarkMonitor AntiPiracy
21	10686672	April 21st	MarkMonitor AntiPiracy
22	10663035	April 14th	IP-Echelon Pty Ltd
23	10676134	April 17th	IP-Echelon Pty Ltd
24	10664211	April 14th	IP-Echelon Pty Ltd
25	10664211	April 14th	IP-Echelon Pty Ltd
26	10672948	April 17th	IP-Echelon Pty Ltd
27	`<never indexed>`	-	-
28	`<re-indexed>`	-	-
29	`<still indexed>`	-	-
30	`<still indexed>`	-	-

Notes:

1. The Notice ID refers to the ID of the notice in the Chilling Effects repository.

2. Home Box Office, Inc is the copyright holder for all the notices.

3. URLs #6 and #7 were re-indexed and were taken down again. URL #28 was re-indexed. No notice for its initial takedown.

4. URL #22 has no takedown notice linked to it. The notice listed is filed for a domain that was a mirror of the original link domain.

Table 3.3 lists an example of the re-indexing we observe. This specific link led to a *streaming* site that streamed *allegedly* copyright infringing material. As we can see from Table 3.3, this link listed in Table A.1 as URL #6 was re-indexed multiple times. Interestingly, it is not the *same* link that gets indexed but a variant of the original link. However, the search for the original link returns this slight variant of the original link if the link with the slight variation is indexed and the original link is not indexed. Note, that opening the *original* link in the browser redirects to the variant which appears to have the same content.

Table 3.2. Re-Indexed Link Variants

Type	Link
Original	*http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come/*
Variant #1	*http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come-2/*
Varaint #2	*http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come-3/*

3.1.3 The Search Engine Equivalency

Presently, Google is the only major search engine that discloses the takedown notices it receives. One question that we wanted to answer was if other search engines like Bing also got the same notices. Since none of the other search providers disclose takedown notices, we can only indirectly infer the takedown notices they receive. Our approach here is that, if a copyright holder submits a takedown request to Google then it is reasonable to assume that it would submit a similar request to other search providers at the same time if it intends to do so.

Hence, if copyright holders were to serve takedown notices to search engines other than Google this would be reflected by the links being removed from other search engines roughly around the same time they are removed from the Google search index.

Table 3.3. Re-Indexed Link Snapshot
As an example of the behavior of links being re-indexed we provide a link that we were monitoring and how it appears to be re-indexed multiple times and its status over a period of approximately three weeks.

Date	Status	Link Indexed
April 12th, 2015	Indexed	http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come/
April 13th, 2015	Indexed	http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come/
April 14th, 2015	Removed	`<not indexed>`
April 15th, 2015	Removed	`<not indexed>`
April 16th, 2015	Removed	`<not indexed>`
April 17th, 2015	Removed	`<not indexed>`
April 18th, 2015	Re-Indexed	http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come-2/
April 19th, 2015	Re-Indexed	http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come-2/
April 20th, 2015	Re-Indexed	http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come-2/
April 21st, 2015	Removed	`<not indexed>`
April 22nd, 2015	Removed	`<not indexed>`
April 23rd, 2015	Removed	`<not indexed>`
April 24th, 2015	Removed	`<not indexed>`
April 25th, 2015	Removed	`<not indexed>`
April 26th, 2015	Removed	`<not indexed>`
April 27th, 2015	Re-Indexed	http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come-3/
April 28th, 2015	Removed	`<not indexed>`
April 29th, 2015	Removed	`<not indexed>`
April 30th, 2015	Removed	`<not indexed>`

Notes:

1. For all the entries above we ran the same search query: `info:http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come/`
2. As of the final measurement done on May 6th, 2015 we have not seen this link re-indexed again.

Google Search Index and Re-Index
April 2015

Searching the Bing Index

Similar to the Google measurement, starting April 12th, 2015, we checked which of the 30 links appear, in the Bing search index every day. Figure 3.2 shows the state of the set of 30 links over time in the Bing search index.

The Bing measurement was interesting in a number of ways. Here is a short summary on some of the salient observations:

1. **Smaller Index**

   A quick comparison between Figure 3.1 and Figure 3.2 shows that Bing at its peak never indexed all of the 30 links.

2. **Re-Indexed Links**

   Bing seems to have an issue too with indexing links that have been previously removed. However, in case of Bing the re-indexed links seem to have no apparent variations.

3. **Disparity in Removed Links**

   The disparity in the removed links strongly suggests that not all the takedown notices submitted to Google are submitted to Bing.

3.1.4 The Measurement Hypothesis

The main goal of the measurement was to examine a few of the factors which we usually cannot examine from a passive analysis of the notices.

Responsiveness

Copyright Holders

It appears that the copyright holders are respsonsive and start issuing takedown notices for the links leading to infringing content within a few days. From the results we can see

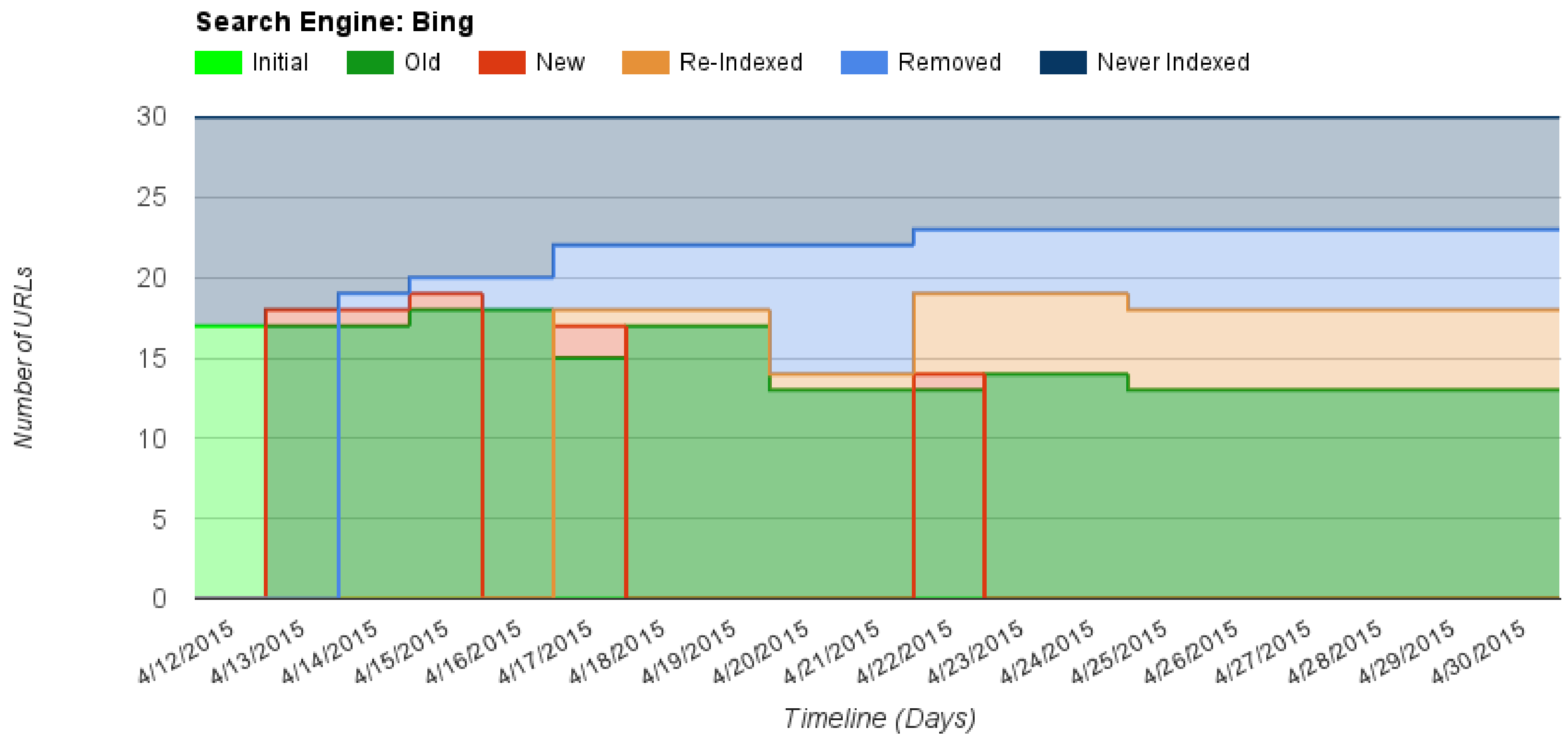


Figure 3.2. Bing Search Index: Link Status

that they first begin taking down content two days after it was released. In a week there were only 15 links in the index and only 4 at the end of two weeks.

Google

Google is extremely responsive and takes down the links the same day they receive the notice. This can be cross-verified by the takedown notices made available through Chilling Effects. Google claims that they process the notices in around 6 hours of receiving the takedown notice [14].

Content Hosting Services

It is still hard to gauge the responsiveness of the content hosting services. In most cases, the domains of the links we monitor do not themselves store the allegedly infringing content. Thus, without doing a deep crawl it is hard to gauge the *liveness* of a link and hence the responsiveness of the content hosting service that hosted the infringing content. We briefly discuss this issue in Chapter 6.2 and discuss potential solutions for it.

Takedown Notices to Bing

Looking at the data, we see that not all notices are served to Bing. However, why Bing chooses to re-index certain links after removing them is an open question.

Reaction to Notices

Perhaps the most interesting aspect of this measurement was how the sites whose content was being flagged and *taken down* adapted. The use of variants of the original link which bypasses the de-indexing of the original link is one such adaptation.

Effectiveness

In this case, there seems to be essentially two aspects to making the entire process effective. Note that we make this observation from the point of view of *finding* things using search engines. The first aspect is the copyright holders being pro-active and flagging content as soon as they can. The second aspect is the search engines themselves. Google is *the gateway* for many of the people to the sites containing infringing content. To counter this, Google updated their search algorithms to down rank domains that have had valid takedown notices issued against it [15].

Thus, it seems the hope and focus is on common users not being able to effectively locate the content, in addition to taking it down.

Chapter 4

A Storm of Swords

One of the main goals of this work was to analyze the takedown notices in the Chilling Effects repository. Specifically, we wanted to observe the trends over time. For the purposes of this work, we mainly focus on the timeframe from 2010 through late 2014. As mentioned in Chapter 2, we do not include the takedown notices submitted using the new API.

There are two primary trends that we are interested in. The first trend is how the use of takedown notices has changed over time. The second trend is how the takedown notices themselves have changed over time. Note that this analysis is done on the takedown notices from the Chilling Effects repository. Additionally, we do not distinguish between notices that are sent under the different provisions of Section 512. The reason is two fold. First, the distinction is not made clear in the notice and is based on interpretation. Second, for this work we are interested primarily in the general trend rather than the specific per-provision trends.

Finally, we note that the takedown notices are filtered before being analyzed. We ensure that the notices with *corrupted* fields are excluded. The corruption could be that there is no value specified. Alternatively, the value specified could be *insensible*. For example, there are notices that have dates from the future. However, we do include notices with no *infringing URLs* listed. Note that these considerations may result in notice counts that could potentially be different from those in previous works.

Table 4.1. Takedown notices by recipient and year

Year	Google	Twitter	Yahoo	Digg	Others	Total
2001	26	0	0	0	21	47
2002	70	0	2	0	137	209
2003	145	0	2	0	135	282
2004	263	0	0	0	815	1,078
2005	685	0	0	0	2,564	3,249
2006	1,285	0	6	0	2,133	3,424
2007	770	0	1	4	1,997	2,772
2008	2,038	0	4	31	142	2,215
2009	8,147	1	84	35	163	8,430
2010	18,952	328	582	23	156	20,041
2011	119,579	4,676	1	13	398	124,667
2012	525,866	3,362	0	2	562	529,792
2013	546,319	8,209	1	0	366	554,895
2014	567,066	10,349	2	0	408	577,825

4.1 Over the Years

Before we consider the notices post 2010 in more detail we briefly analyze the notices received over the years. Table 4.1 is an extended tabular version of Figure 2.1. Table 4.1 lists the number of notices submitted by major sources from 2001 through late 2014. The increase in the number of notices over the years is clear. Another interesting observation is the sudden increase of takedown notices submitted in 2011 and 2012, which we discuss below.

4.2 Over Months

The trend of increasing notices over the years is undeniable. To better understand the increase shown in Table 4.1 we look at the number of notices filed each month post 2010 as shown in Figure 4.1. The increase in takedown notices in 2011 and 2012 seen in

Table 4.1 is reflected in Figure 4.1.

4.2.1 SOPA/PIPA Effect

In January, 2012 there is a 60% increase in the number of notices submitted, whereas in April 2012 there is a 141% increase in the number of notices submitted. This could indicate an interesting correlation to the tabling of SOPA[1] and PIPA[2] by Congress. Stop Online Piracy Act (SOPA) and Protect IP Act (PIPA) were laws that had provisions much desired by major copyright holders. After these laws were tabled, it seems that copyright holders decided to use existing laws more actively, which could explain the increase. Interestingly, this is then followed by a sudden dip around November 2012. We hypothesize that this is because copyright holders started listing more URL takedown requests in a notice than previously.

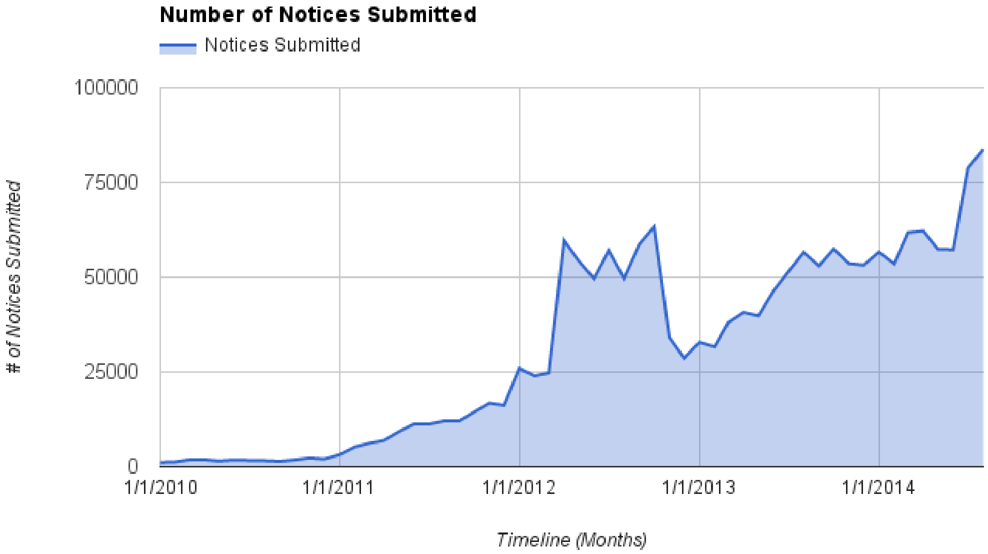


Figure 4.1. Number of Notices submitted per month.

[1] http://en.wikipedia.org/wiki/Stop_Online_Piracy_Act

[2] http://en.wikipedia.org/wiki/PROTECT_IP_Act

4.3 URL takedown requests

The other related trend to takedown notices is the number of URL takedown requests that these notices contained. Figure 4.2 shows the number of URL takedown requests for each month post 2010 along with the trendline. It can be seen that the number is increasing at a significant rate over the years. In a period of one year from January 2011 to January 2012, the number of URL takedown requests increased by 870%. Additionally, corresponding to the dip in the number of notices at the end of 2012, one would expect to see a similar dip in the number of URL takedown requests. However, surprisingly that does not happen. We discuss this anomaly later in the chapter.

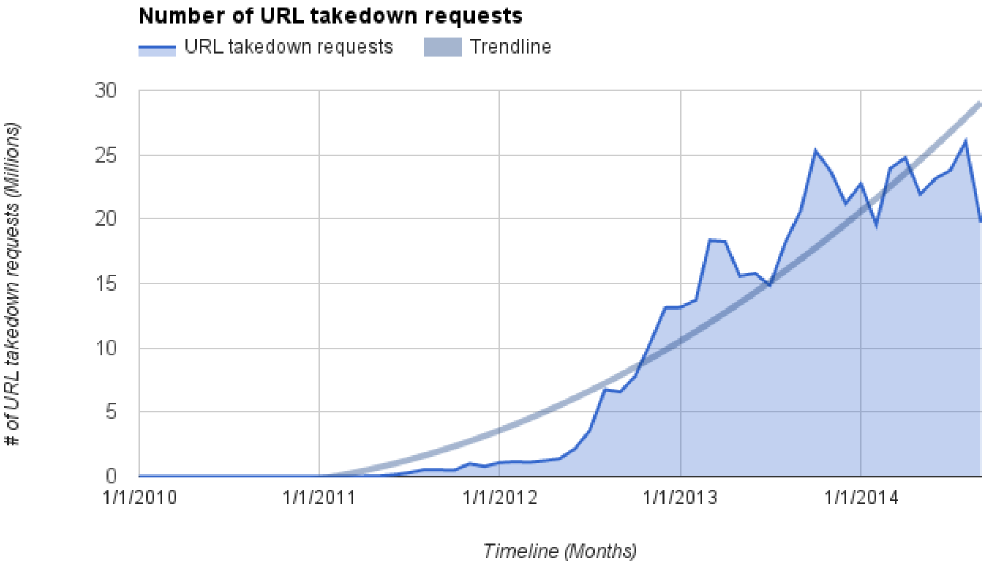


Figure 4.2. Number of URL takedown requests submitted per month.

4.4 URL takedown requests in Notices

To explain and understand the sudden increase of notices followed by a dip as seen in Figure 4.1, we look at the change in the size of notices over time and see if we can correlate that to our observation. Note that here by *size* of a notice we mean the number of URL takedown requests in the notice.

Figure 4.3 is a stacked graph showing the percentage of notices that have fewer than 10 URL takedown requests along with the percentage of notices that have 0 or 1 URL takedown requests for each month post 2010. Additionally, to understand the distribution of the notice sizes over months we plot a boxplot as seen in Figure 4.4. The boxplot follows the convention that the bottom of each box is the 25th percentile while the top is the 75th percentile. The line in the middle is the 50th percentile. The whiskers at the end indicate the maximum and minimum value excluding the outliers which are not part of the plot here.

Figure 4.4 shows that until late 2011/early 2012, takedown notices were relatively small. For the purposes of this work we define *small notices* as *notices that have less than 10 URL takedown requests*. Figure 4.3 reflects this and we can see that until roughly mid 2011 approximately 90% of the notices have fewer than 10 URL takedown requests. This trend changes around 2012 where we see a large increase in the number of notices while the number of small notices continues to decrease. Thus, the dip in number of notices at the end of 2012 and the lack of a corresponding dip in the number of URL takedown requests could be attributed to the gradual shift from smaller notices to larger notices. Note that the end of 2014 data is not exact since not all notices submitted during 2014 appear in our dataset and could explain the dip at the end of the timeline.

4.5 Claims in Notices

Takedown notices can contain multiple claims or *works* as labelled in the notices. A claim contains a list of URLs that are being requested for the takedown. One possible reason for the general increase in the number of URL takedown requests per notice could also be because copyright holders chose to pack multiple claims into a single notice. One of the motivations for doing this seems to be that Google may process notices in the

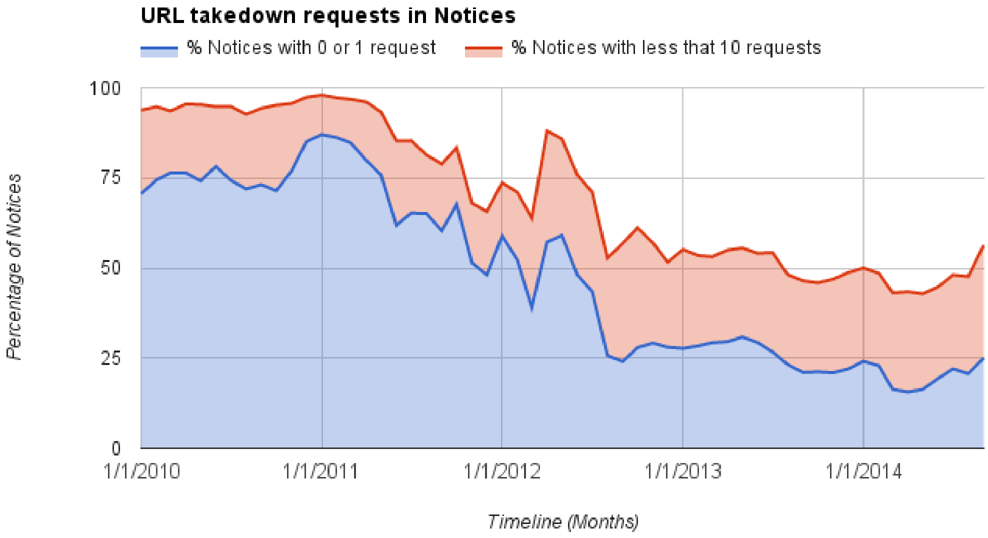


Figure 4.3. URL takedown requests in Notices

order in which they are received and hence packing as many URL takedown requests as possible in a single notice will guarantee better turnaround time than if multiple notices were filed [22]. Until recently, Google had a cap of 10,000 requests per notice [11].

Figure 4.5 is a boxplot similar to Figure 4.4 and shows the distribution of the number of claims in notices for each month post 2010. From Figure 4.5 we can see that until 2011, a single claim per notice was the norm. This started changing towards the end of 2012 which coincides with the increase in the URL takedown requests submitted per notice. However, there are concerns about whether the current use of multiple claims in a notice is consistent with the intended use as per the law [22].

4.6 Summary

The main observations of this chapter could be summarized as follows:

1. The trends of increasing number of notices and the accompanying number of URL takedown requests continue to grow.
2. There has been a gradual shift from smaller notices to larger notices and filing of multiple claims in a single notice.

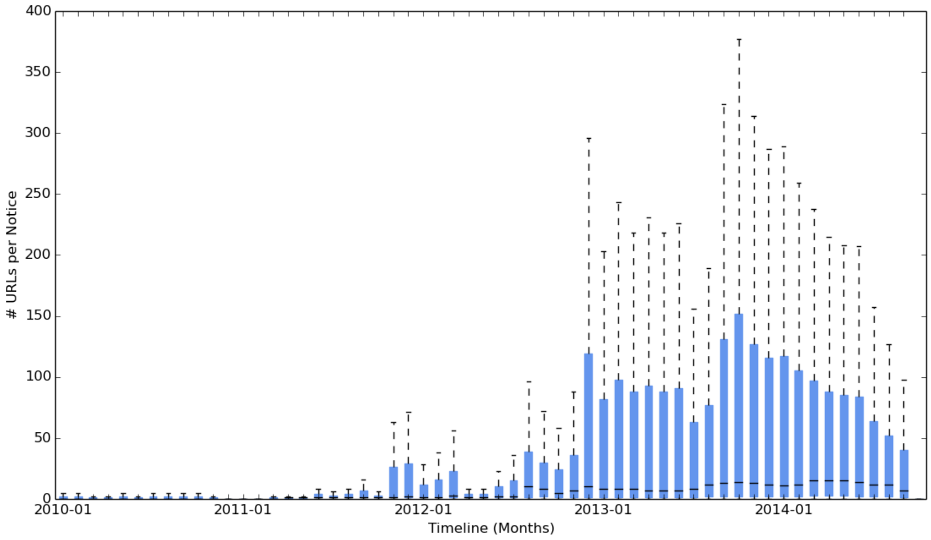


Figure 4.4. Boxplot of URL takedown requests submitted per notice

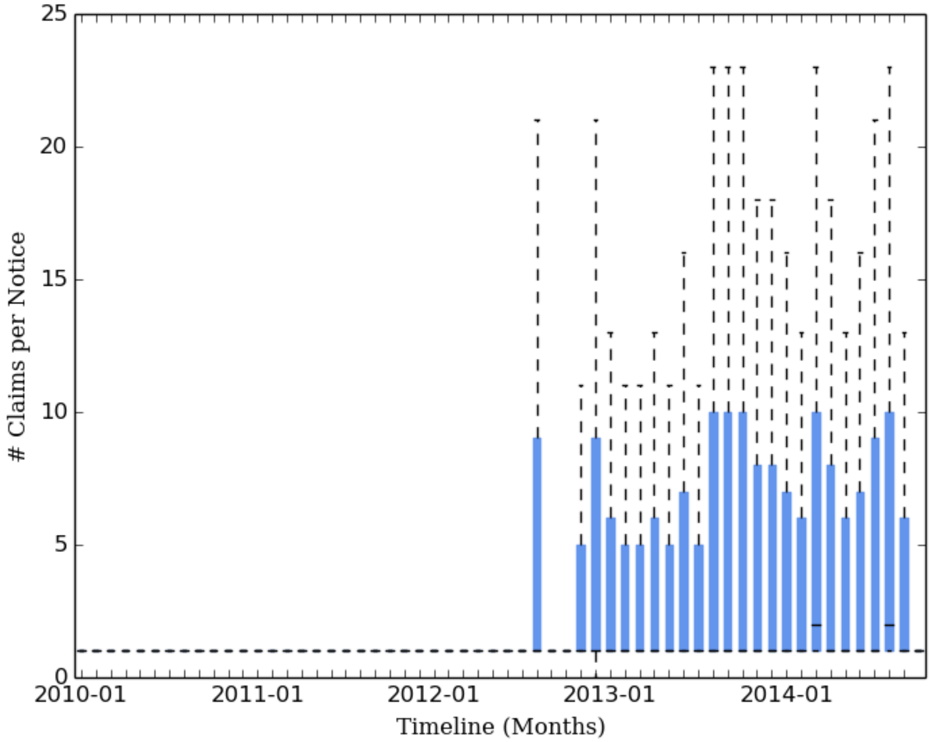


Figure 4.5. Boxplot of Claims submitted per notice

Chapter 5

A Clash of Kings

Given the general trend of takedown notices submitted over time, more than a few questions arise about the copyright holders who file these notices. Who are the major copyright holders who file these takedown notices? What industries do these organizations represent? Are takedown notices always filed by the copyright holders or are there industry representatives also filing notices too? If not, who are these other organizations who file notices on behalf of the copyright holders?

We start by looking at the top 5 copyright holders by the number of URL takedown requests. Table 5.1 shows copyright holders who have submitted the largest number of URL takedown requests according to the Google Transparency Report [14]. We use this to motivate our discussion and analysis of copyright holders from the Chilling Effects repository.

Table 5.1. Top 5 Copyright holders with most URLs submitted (Google Transparency Report)

Copyright Holder	**Industry**	**# URLs Submitted**
BPI LTD Member Companies	Music	137,500,891
RIAA Member Companies	Music	64,069,009
Froytal Services Ltd	Adult	48,638,597
FOX	Movies	40,380,026
MG Premium Ltd	Adult	25,166,387

The first question that we would like to answer is what industries do these copyright holders represent? On a high level there seem to be 5 major industries that most of the copyright holders belong to: Music, Adult Entertainment, Movies, Software and Books.

5.1 Breakdown by Industry

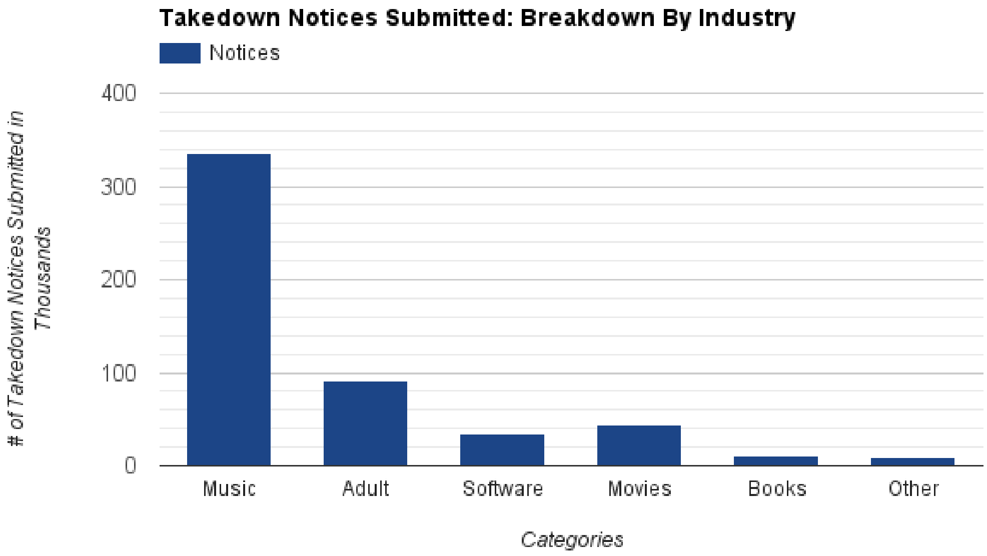


Figure 5.1. Notices Submitted: Breakdown by Industry for Top 20

We are first interested in which industries are responsible for the notices. Figure 5.3 shows the CDF plot of takedown notices and URL takedown requests submitted by copyright holders. As seen in Figure 5.3 there is a long tail for copyright holders. Hence, it is infeasible to break down notices by the industry they represent for all the notices since not all the entities are known. To make the analysis more tractable, we limit the analysis to the submissions of the top 20 copyright holders by the number of notices and URL takedown requests submitted respectively. The top 20 copyright holders account for 28.3% of the notices and 71.3% of the URL takedown requests.

Figure 5.1 shows a breakdown of takedown notices submitted by the top 20

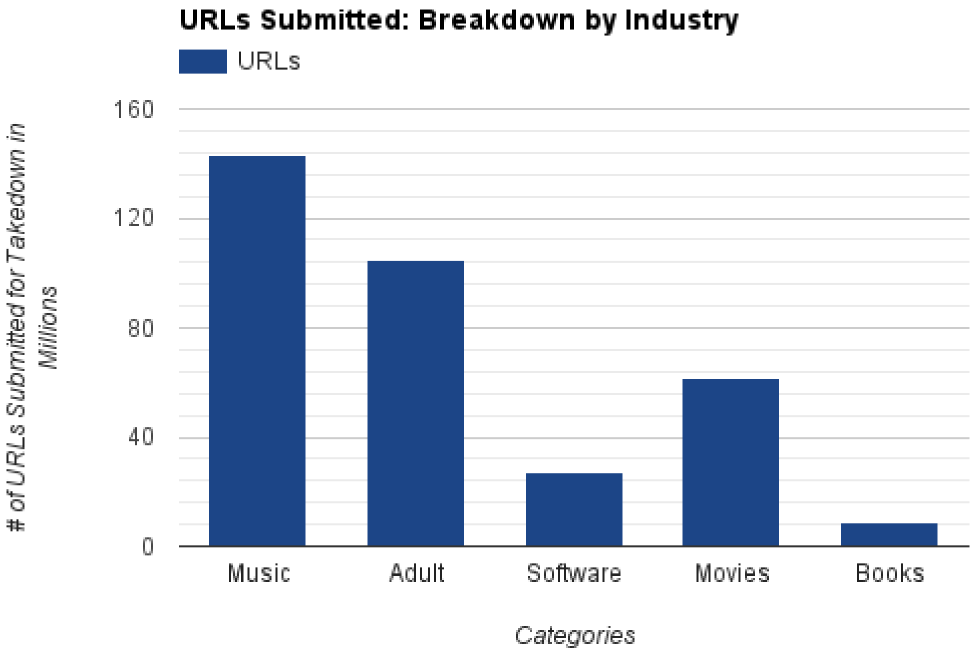


Figure 5.2. URLs Submitted: Breakdown by Industry for Top 20

submitters of notices by industry. Music and Adult Entertainment are the two largest contributors to takedown notices and account for 63.8% and 17.6% of the takedown notices submitted (by the top 20) respectively.

Additionally, Figure 5.2 shows the breakdown by industry, of the top 20 copyright holders by the number of URL takedown requests. Music and Adult Entertainment are again the two largest contributors. However, interestingly Music now only accounts for 41.4% while Adult for 30.3% of the URL takedown requests as opposed to 63.8% and 17.6% of the takedown notices submitted.

5.2 Notices and URL takedown requests

This discussion presents an interesting question. What is a better measure for *enforcement activity*? Is the number of notices filed by a copyright holder or the number of URL takedown requests submitted a better measure?

It is evident from Figure 5.3 that a few copyright holders account for the majority of the URL takedown requests. This again ties into the increasing sizes of notices we have seen in recent years. As a quick comparison, the top 50 copyright holders account for 80% of the URL takedown requests submitted but only for around 35% of the notices submitted. Table 5.2 and Table 5.3 lists the top 10 copyright holders by the number of notices submitted and by the number of URL takedown requests respectively. Table 5.2 and Table 5.3 confirms that there is indeed a disconnect between the number of notices and the URLs. For example, VIZ Media which produces *manga* magazines does not show in the top if we only consider the number of notices submitted. Thus, it seems that the number of URL takedown requests is more indicative of the *enforcement activity* of an organization than the number of notices submitted.

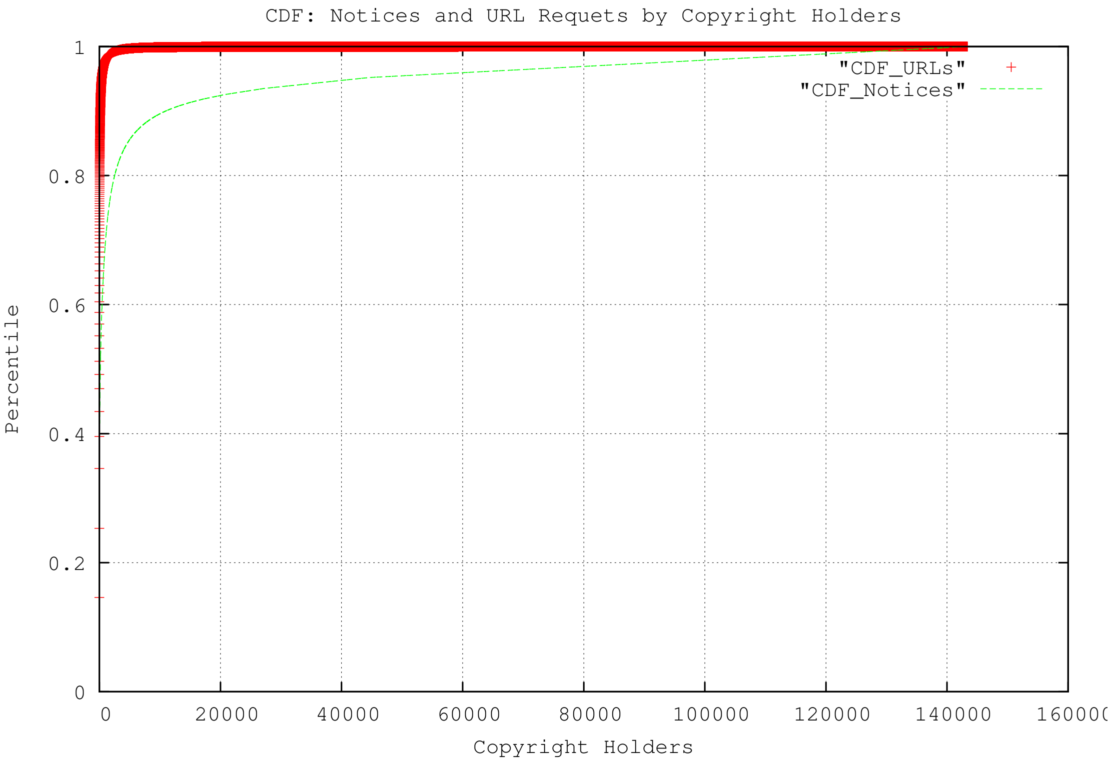


Figure 5.3. CDF: Notices and URL Requests Submitted by Copyright Holders

Note that, in Table 5.2 and Table 5.3, in the case of FOX we have included notices

Table 5.2. Top 10 Copyright holders with most notices submitted (Chilling Effects)

Copyright Holder	**Industry**	**# Notices Submitted** (%)
BPI LTD Member Companies	Music	266,110 (14.4%)
Froytal Services Ltd	Adult	36,466 (1.97%)
Microsoft	Software	27,380 (1.48%)
CA Co, Ltd.	Adult	19,609 (1.05%)
IFPI	Music	17,368 (0.94%)
RIAA	Music	17,243 (0.93%)
Fox+	Movies	14,973 (0.81%)
Intellectual Property Promotion Association	Adult	14,074 (0.76%)
RK NetMedia Inc	Adult	13,140 (0.71%)
Paramount	Movies	12,421 (0.67%)

from Twentieth Century Fox, FOX and related organizations.

Table 5.3. Top 10 Copyright holders with most URL requests submitted (Chilling Effects)

Copyright Holder	**Industry**	**# URLs Submitted** (%)
BPI LTD Member Companies	Music	89,575,023 (18.2%)
RIAA	Music	53,614,200 (10.9%)
Froytal Services Ltd	Adult	45,925,788 (9.31%)
Fox+	Movies	25,657,004 (5.20%)
Microsoft	Software	20,941,445 (4.24%)
RK NetMedia Inc	Adult	19,145,509 (3.88%)
Hydentra L.P	Adult	10,946,334 (2.18%)
DMM.com Labo, Ltd	Adult	10,089,750 (2.04%)
NBC Universal	Movies	9,928,891 (2.01%)
VIZ Media	Books	8,926,177 (1.80%)

5.3 Reporting Organizations

Section 512 provides that a takedown notice may be filed by *a person authorized to act on behalf of the copyright owner*; the copyright holder need not be the one submitting the takedown request. This has led to a rise of entities who specialize in detecting online infringement and sending takedown notices on behalf of the actual copyright holder. Essentially, the entity that files the notice need not have the license to use the content, but needs to be authorized to send the takedown notices on behalf of the copyright holder [17, 22].

On the whole, an entity that submits the notice can be of three types: a Collective Management Organization (CMO), an agent or a copyright holder.

Table 5.4 and Table 5.5 lists the top 5 entities by the number of takedown notices and the number of URL takedown requests submitted respectively from the dataset we analyze. It is clear that agents have become the norm. The Music industry is well represented by CMOs like BPI and RIAA who generate their own takedown notices. Agents being listed as the top reporters could largely be accounted to small copyright holders since anecdotally we found that smaller copyright holders tended to use agents to submit takedown notices.

Table 5.4. Top 5 Reporters that submitted most takedown notices (Chilling Effects)

Reporter	**Type**	**# Notices Submitted** (%)
AudioLock	Agent	357,585 (19.3%)
BPI Ltd	CMO	265,520 (14.3%)
Degban	Agent	120,898 (6.52%)
MUSO	Agent	102,905 (5.55%)
Digimarc	Agent	59,307 (3.20%)

Table 5.5. Top 5 Reporters that submitted most URL takedown requests (Chilling Effects)

Reporter	**Type**	**# URLs Submitted** (%)
Degban	Agent	93,689,172 (19.0%)
BPI Ltd	CMO	87,776,895 (17.8%)
DtecNet	Agent	55,445,658 (11.2%)
RIAA	CMO	53,382,892 (10.8%)
Takedown Piracy LLC	Agent	26,247,778 (5.32%)

5.4 Copyright Holders: A Monthly View

Do the same *major* copyright holders dominate takedown notice submissions over time? Specifically, do copyright holders that have many takedown notices overall dominate the monthly takedown notice submission as well? If so, do these *major* copyright holders submit takedown notices continuously or do so in batches, say every quarter?

To answer this question, we look at the top 5 submitters of takedown requests for each month in 2013. The year of 2013 is well suited for this analysis since the trends for that year are reflective of the trends we see today. Moreover, the dataset has complete data for the year of 2013.

Table 5.6 shows the top 5 copyright holders for every month by the number of takedown notices submitted in that month in the year of 2013. It shows that the *major* copyright holders consistently submit notices and are often the top submitters every month. BPI (British Recorded Music Industry) Ltd is the top submitter for 10 out of the 12 months. However, we can also see smaller copyright holders appearing once in a while. One possibility is that while major copyright holders continuously file takedown notices, the smaller entities do so in batches.

5.5 Summary

The main observations of this chapter could be summarized as follows:

1. Few major copyright holders dominate the takedown process.
2. The Music and Adult Entertainment industry account for the majority of the takedown notices and the URL takedown requests.
3. The number of URL takedown requests submitted is a better measure of the *enforcement activity* of a copyright holder than the number of notices submitted.

Table 5.6. Notices submitted in 2013 by Month: Top 5 (Chilling Effects)

2013			
January	**February**	**March**	**April**
BPI	BPI	BPI	BPI
Froytal Services Ltd	CA Co. Ltd.	CA Co. Ltd.	Fox
IFPI	Froytal Services Ltd	Froytal Services Ltd	Froytal Services Ltd
Paramount	IFPI	IFPI	IFPI
RIAA	RIAA	RIAA	RIAA
May	**June**	**July**	**August**
BPI	BPI	BPI	BPI
CA Co. Ltd.	CA Co. Ltd.	Froytal Services Ltd	ISO
Froytal Services Ltd	Froytal Services Ltd	Murator	Random House
IFPI	RIAA	RIAA	RIAA
RIAA	YALA.FM	YALA.FM	Shueisha
September	**October**	**November**	**December**
BPI	Armada Music	Armada Music	BPI
Froytal Services Ltd	BPI	BPI	DMM
ISO	ISO	DMM.com	ISO
RIAA	RIAA	INgrooves	RIAA
Spinnin' Records	Spinnin' Records	RIAA	Spinnin' Records

4. Agents have become the norm and account for a large fraction of the notices and URL takedown requests submitted.

5. Major copyright holders tend to continuously submit notices as opposed to smaller copyright holders who submit notices in batches.

Chapter 6

A Dance with Dragons

The trend of increasing URL takedown requests continues to grow. Currently, Google processes around 7 million URL takedown requests a week [14]. Given the sheer number of these takedown notices and the even larger accompanying number of URL takedown requests it is imperative to come up with new ways to analyze the data. One potential approach is to analyze this growing corpus of data using machine learning techniques. A related effort incorporated machine learning into the analysis of this dataset. In this chapter, we focus on issues we face that could potentially use machine learning techniques and also discuss some initial results.

6.1 Search Index Liveness

One question that we would like to eventually answer is, do copyright holders send takedown notices to search engines other than Google? We partially answered this question in Chapter 3. However, in the future a more thorough analysis can be done by seeing if the links given in the takedown notices to Google are still live in the search index of Google, Bing and other search engines.

6.1.1 Issues

The Cost of Crawling

Most of the search engines rate limit the number of searches that can be done in a period of time. This became a crucial hurdle when doing this work: *crawling a link is significantly cheaper than crawling for the same link in the search engine index*.

Links after Action Taken

The other class of issue arises since we get these notices only after action has been taken. This usually means that the link has potentially been removed from the Google search index. However, when we crawl the link itself there are three possibilities:

1. The link does not lead to a valid page as it has been taken down, presumably, as a result of a notice to the content hosting provider.

2. The link leads to a valid page but the *content* has been taken down. Again, presumably as a result of a notice to the content hosting provider.

3. The link leads to a valid page *with the content*. In these cases, anecdotally we found that these domains do not fall under US law.

Limiting the Search Space

Given the cost of crawling a search index for a link, it makes sense that we weed out links that will potentially not help us determine anything about takedown notices to other search engines like Bing.

Let us walk through each of the above scenarios and then discuss what it could mean for the liveness of a link in the search index. First, if we were to do a search for a link in the Bing search index *after* the page that it leads to itself has been taken down, it is not possible to distinguish the exact reason for the link to not be in the index. It could be because the link was never indexed or that it was removed from the index after the page content was taken down or as a result of a takedown notice. Thus, there is little benefit in checking for liveness for these links.

To put things in perspective, nearly 65% of the links return an error when crawled using a corpus of more than 3 million URL takedown requests crawled from around 10,000 notices in 2014.

Ideally, we would have liked to search for every link in the Bing search index. However, given that we are rate limited we should prioritize links that have a higher likelihood of being in the index. One solution could be to restrict searches to links that return valid pages. However, just because a link leads to a valid page does not mean the content is available. Further, we would also like to prioritize links that have the content available since it has a greater chance of being in the index. However, this turns out to be a difficult problem since it is hard to automate distinguishing between an error page and a content page.

6.1.2 Solution: An Attempt

Clearly, if we are able to distinguish between error pages and content pages we will be able to effectively prioritize our crawls and hence maximize returns from the crawl (there is a small caveat here which we detail later). Thus, an attempt was made to automate this classification using machine learning. As a part of work related to this project, varied approaches like K-Means and hierarchical clustering were used. Note that this was done on a per domain basis since we would expect an error page of a single domain to be similar.

To test the liveness in the Bing search index, we chose a set of 300 URLs that appeared to have the content available with a high probability. The URLs selected were from the corpus of 3 million URL takedown requests crawled from around 10,000 notices in 2014. Out of 300 links we found that 13 links were still live on the Bing index.

That said, this is one anecdote and we do not know how many links were indexed by Bing in the first place. Note that this is still in its early stages and improving the classification is one major goal for the future. However, this is further complicated since *content liveness* is not as straightforward to determine as one would think as we shall see later in this chapter.

6.2 Link Liveness: A Rabbit Hole

On the face of it, determining if a link is live seems to be a straightforward problem. However, as it turns out, as with most real world measurement, the issue of *link liveness* becomes surprisingly involved and complicated quickly. Before we start, we briefly discuss what we exactly mean for a link to be *live*. In the context of these notices, we say a *link is live if the content on the page loaded by visiting the link is still available*. For example, if the link were to lead to a streaming site then we would say the link is live if the video content is available. Hence, the question then becomes gauging the availability of the content.

6.2.1 Issues

So why exactly is gauging the content availability hard? The issue lies in how most of these sites serve the content. It turns out that sites of most of the links that are mentioned in the takedown notices sent to Google do not *host* any content on their servers. One reason to do this is to avoid dealing with DMCA and the fallout thereof. Thus, most of these links embed content that points to some other site that actually hosts the content. This is where things become difficult. When we do a crawl, we only crawl the link and do not do a deep crawl. Hence, even if we are able to crawl the link and get a valid page it does not imply that the content was available.

Further, it is not always straightforward to get the *exact* link that leads to the content from the page given the number of links embedded in a typical page of this nature. For example, one link we crawled as part of our active measurement had 164 embedded links in its HTML page. Thus, given the nature of these sites it is infeasible to crawl all the links from the page. Further, we are not always certain of the exact content (video, file, torrent file) to be expected, which adds an additional hurdle.

6.2.2 Potential Solutions

One brute force solution could be to manually create a *whitelist* and *blacklist* of domains. The *whitelist* could list the domains that usually tend to actually host the content, while the *blacklist* lists domains that do not so that when we deep crawl we can narrow the pool of links to be crawled. However, given the flux in these domains, it is impractical to manually populate these lists.

Hence, if instead of manually creating a *whitelist* and *blacklist* we can instead using machine learning to classify domains automatically based on a few samples we potentially can to solve the problem of liveness.

Chapter 7

The Winds of Winter

The DMCA has been in effect for the last 15 years since it came into effect in 2000. The DMCA safe harbor provisions, in spite of having its detractors, has been considered essential to the growth of the Internet [7]. Further, the DMCA safe harbor has been the basis for similar provisions in similar laws in other countries [22]. Given that, it becomes even more important to understand how the law is being used. More specifically, to understand if the law is being used adhering to the DMCA requirements. One of the long term goals of this project is to be able to identify patterns of use over the years.

7.1 Mega Takedown Notices

One prominent trend was the increasing number of claims and takedown requests packed into takedown notices. Given the sheer number of requests, one conjecture is that a large number of these requests are generated using automated scripts [6]. One of the prime requirements for a takedown notice to be valid under the DMCA is the fact that it must be based on a *good faith belief* that the material identified actually infringes a copyright held by the copyright holder submitting the notice. However, if the takedown requests were in fact generated by automated scripts, it remains to question if anything additional needs to be done for the *good faith belief* requirement to be met.

Concerns about the erroneous infringement claims are acknowledged but not considered a major problem. Google complying with the takedown of 97 % of the URL takedown requests is used as evidence that *inaccurate notices* are a small fraction of the actual notices. Given that, and the paucity of takedown notices made publicly available from service providers other than Google, essentially means the extent of these *inaccurate notices* is *subject to dispute* [6, 22].

7.1.1 Issues with Automated Scripts

For now, it seems that takedown notices generated using automated scripts, even those without human review, have gained tacit approval [6]. However, organizations like EFF have expressed the opinion that they consider this a likely violation of the DMCA process [8, 9].

7.2 The Rise of P2P

Perhaps one of the most disruptive change for DMCA as a law has been the rise of Peer-to-Peer (P2P) file sharing. P2P file sharing can be used for transferring files between clients using a distributed algorithm without a centralized server. A popular implementation of P2P file sharing is the *BitTorrent* protocol.

The lack of a central server in P2P file sharing makes it difficult to stop sharing of the content. Also, recent improvements to the protocols means that popular *BitTorrent indexer sites* no longer have to store any copyright infringing files. The Pirate Bay, touted to be the most popular *torrent* sharing site, claims:

“No torrent files are saved at the server. That means no copyrighted and/or illegal material are stored by us. It is therefore not possible to hold the people behind The Pirate Bay responsible for the material that is being spread using the site” [3]. In these cases, copyright holders then usually resort to a lawsuit against the *indexer* sites. A recent

lawsuit resulting in key Pirate Bay domains being seized is an example of such an action [13].

DMCA predates the rise of P2P, which is why some of the provisions are not as effective against P2P as copyright holders would like. For example, in the *RIAA vs Verizon* case [4], the RIAA wanted ISPs to turn over identities of subscribers who they believed were infringing their copyright. However, since Section 512(a) did not require ISPs to divulge the details of their subscribers the RIAA lost.

That said, the copyright holders can still send DMCA notices to ISPs who can then forward the notices to the subscriber. In an effort to protect content, copyright holders police P2P networks to collect evidence of infringement and then issue a takedown notice against allegedly infringing users. However, a study done by the University of Washington uncovered that, at the time, copyright holders used inconclusive methods to identify infringing users which led to spurious takedown notices [21].

All in all, a continued study of how copyright holders deal with alleged infringement leading from P2P could potentially be interesting since the use of DMCA for this purpose is still evolving.

7.3 The Search URI Issue

DMCA requires that the takedown notice should specify a URI that identifies and locates the content uniquely. However, a subtly arises in the form of search results that then link to the actual infringing content. These URIs will *not* resolve to a specific location with infringing content. Instead, they will resolve to a page that lists these allegedly infringing resources, some of which may not be infringing. In a court case, Perfect 10 vs Giganews the court observed that *“while a web search may find a number of results, the search itself does not actually locate the items found; the search engine just presents its search results in a list, and any item in that list is not located until its*

URL is extracted" [22, 24].

Thus, one interpretation could be that search terms ideally should not be accepted as valid URIs for infringing material in a takedown notice. In fact, Google has rejected almost all of the URIs that take form of Google search results [22]. This then raises the question, should search results from other sites be accepted by Google as legitimate takedown requests or not? To be clear, since these URIs themselves point to a resource in the Google search index they are still valid takedown requests. However, if or not these take downs are legitimate is the question.

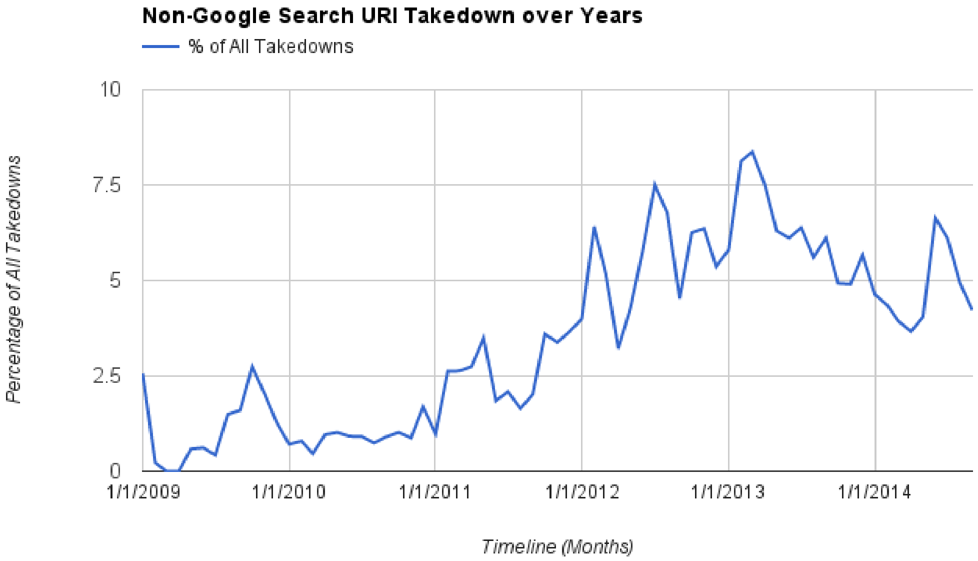


Figure 7.1. Non-Google Search Terms Takedown

Figure 7.1 shows the number of URL takedown requests over time that are likely search results from other sites. To avoid false positives we simply count the requests that have the term *search* in them on a per month granularity. Admittedly, this likely provides a lower bound of the actual URL takedown requests that relate to search terms. However, the main goal here is to question the fact whether these URL takedown requests are legitimate or not and hence even a lower bound, as determined by the measurement, is acceptable. Figure 7.1 clearly shows the increasing trend in the number of URL takedown requests that relate to search terms, and it will be interesting to see how this trend continues in the future.

7.4 DMCA Safe Harbor: The Other Side

Most of the focus while studying DMCA has been on copyright holders. However, there have been multiple instances in which the OSPs abused the safe harbor provisions of DMCA [19, 20]. Given this, in the future it will also be important to focus on the OSPs as well.

Chapter 8

A Dream of Spring

The Internet plays an extremely important role in the world today. Perhaps it is fair to say that the Internet derives its greatest strengths from the sharing of knowledge and information. That the Section 512 process affects the Internet is a given. However, how to objectively quantify this effect has been a challenge for a long time. This work is an attempt to start down the path of *objectively quantifying* the effect that the DMCA Section 512 process has had on the Internet, and this chapter summarizes the lessons learnt during the duration of the work. Finally, we will briefly talk about the potential issues and directions that related future work might want to consider.

8.1 Evolution of Use

As far as laws go, the DMCA is a fairly young law. A fallout of this is that there are still many areas that are not clear from the legal viewpoint. As a result, the use of the law also shows continuous changes over time. An example of this would be the changing nature of the takedown notices themselves. Over time, the copyright holders and OSPs have evolved amongst themselves an effective way of filing takedown notices. This evolution can be gauged from the changes in how the takedown notices themselves have changed over time. This is also reflected in how Google accepted takedown notices before and now, specifically the cap that Google had on the number of URL takedown requests per notice [11].

Evolution of the use of the law is certain. Given that, it is imperative that periodically we ascertain if the current use of the law is not just *fair* but also not detrimental to the health of the Internet as a whole.

8.2 Reliance on Google Takedown Notices

The takedown notices that are sent as part of the DMCA Section 512 process are not part of public record [22]. This has greatly limited the sources of information to completely understand the takedown process. While repositories like Chilling Effects greatly help, as Figure 2.1 suggests, the repository is dominated by notices from Google which may skew our results.

Additionally, this means that all conclusions are based on this dataset, which may or may not reflect the common case. For example, the Internet Policy Task Force tasked by the Department of Commerce concludes that the *incorrect* or *abusive* notices are a small fraction of the actual notices [6]. This is based on the Google Transparency Report, which claims that it complied with 97% of all the requests made [14]. Whether this conclusion is justified is moot since this is the only major source of data at hand. Finally, how Google makes a determination if a said link is in fact legitimate or not is not clear. For example, non-Google search URIs are considered legitimate takedowns when one could argue about their legitimacy.

On the whole, drawing conclusions based on a singular source of data should be done with some care.

8.3 Flawed Notices

One of the other issues has been flawed notices: notices that have missing data or even outright insensible data. Around 10% of the notices had one or more of the important fields corrupted. While this does not fundamentally prevent analysis, issues like this are a major hindrance in analyzing the notices.

8.4 Scaling the Analysis

Our analysis shows the increasing use of the DMCA. Currently, Google receives around 7 million URL takedown requests every week [14]. Even our current analysis involved roughly 2.5 million notices that were responsible for the takedown of around 494 million links. Clearly, traditional empirical analysis is going to become harder and harder at this rate.

Hence, it is imperative that we use scalable processes to analyze these notices. For example, for this work we decided to run the analysis as MapReduce tasks rather than using a database. Given the scale, *empirical* analysis based on manually determined *heuristics* [22] is no longer a feasible option. Thus, incorporating machine learning techniques to make analysis of the notices more tractable is going to be crucial.

Another potential direction that we would like to scale the analysis is to extend the analysis to the links. One of the major concerns with an extra-judicial process is abuse of the process. However, unless we extend our analysis to include the links the chances of spotting a pattern of abuse are limited.

8.5 Active Measurement Studies are Important

The Game of Thrones active measurement study was a departure from the traditional empirical analysis of the takedown notices. This study helped us understand the process on a deeper level than what would have been just possible by passive analysis of the notices. It helped us look at the takedown notices in a new light. More importantly, we were able to see how the process actually works.

Additionally, the active measurement study allowed us to make indirect inferences about other service providers. Given the lack of actual takedown notices this indirect inference methodology might be a meaningful way of understanding the process from the point of view of other service providers.

One of the biggest advantages of the active measurement is that we can see the events unfold. The passive analysis constraints the extent of our analysis. Hence, one potential direction of work could be to try and design larger scale active measurement studies.

8.6 The Road Ahead

This work sought to start understanding the DMCA Section 512 process with a goal to be able to objectively quantify its effect on the Internet. Thus, a major part of the work was to try and understand the challenges and identify the problems that needed to be solved to keep up with the increasing number of notices and the change in the analysis approach it will bring around.

We conclude with a list of issues that need to be tackled as part of future work.

1. **Link Liveness aka Links with useful information**

   As mentioned before, to further understand the process we must also include the links that get taken down into the analysis. However, since we only see these links after action has been taken on them, not all these links remain useful. Hence, one of the more important goals in the future should be to effectively determine if a link has useful information that could indicate to why it was requested to be taken down. Given the nature of the problem this is well suited for a machine learning approach to solve the problem.

2. **Deep Crawling**

One by-product of how these sites usually serve information is that just crawling the links is not sufficient. Hence, one of things we need to think about is to crawl a level deeper, that is crawl links that appear in the page we originally crawled. As discussed in Chapter 6.2, this problem can quickly become intractable and hence we need to have an effective way of pruning away the noise.

3. **More Active Measurements**

   One of the main lessons of this work was that passive analysis can only get us so far. If we are to have meaningful insights then we need to combine the passive analysis with active measurement studies.

4. **Other Sources for Takedown Notices**

   Takedown notices sent to Google dominate the Chilling Effects repository and hence dominate any results. Thus, it would potentially make sense to explore takedown notices sent to recipients other than Google separately. In addition to Chilling Effects which has data from Twitter, Yahoo and Digg companies like Github make the DMCA notices[1] they receive available too.

5. **Abuse of Safe Harbor**

   Most of the focus as a result of the takedown notices is on the copyright holders. However, there is also the issue of abuse of the safe harbor provisions by the OSPs. Hence, in the future it would also be desirable to have studies focusing on the OSPs to paint a more complete picture of how the entire system works.

[1] https://github.com/github/dmca/

Appendix A

Active Measurement Study

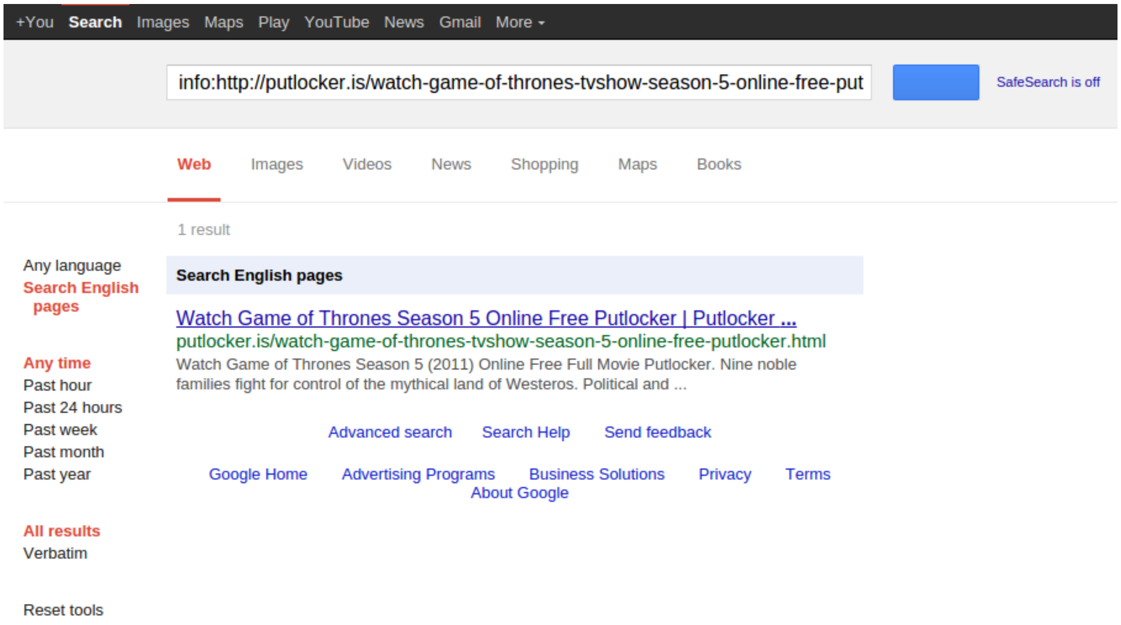
A.1 Search Engines: Link Indexed or Not?

To examine if links were indexed by a search engine we re-purposed some work related to search engines done earlier [5].

A.1.1 Google

For Google we run the query with the search keyword *info* which returns a result only if the link exists in the index. In particular, for Google we retrieve the HTML document for the search result and parse the HTML document to determine all the results returned. Since we use the keyword *info* if a link is in the search index then it will return a single result. Hence, if the HTML document contains a result in it we determine that the link is in the index.

We can also visually verify since we get the HTML document. Figure A.1 shows an example of a result when we determine the link is in the index. Figure A.2 shows an example of a result when we determine the link is *not* in the index. As a check, if we run this query on a browser we get the result back with a *DMCA complaint* snippet as seen in Figure A.3.



info:http://putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-put

1 result

Search English pages

Watch Game of Thrones Season 5 Online Free Putlocker | Putlocker ...
putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.html
Watch Game of Thrones Season 5 (2011) Online Free Full Movie Putlocker. Nine noble families fight for control of the mythical land of Westeros. Political and ...

Figure A.1. Example of a Link in the Google Search Index

A.1.2 Bing

For Bing we use the Bing API. The Bing API allows us to query for search results and choose between different formats for the results to be returned in. For this work, we chose JSON since it is structured and much easier to parse than HTML.

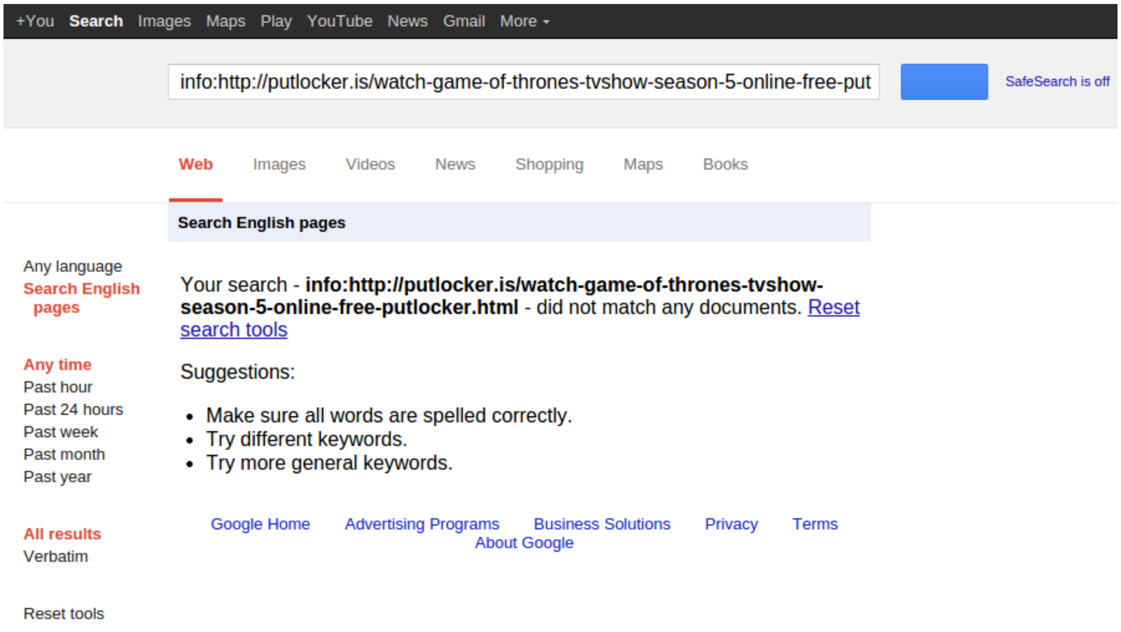
In this case, we determine a link is not in the index if an empty JSON object is returned as a search result. Listing 3 shows an example of the empty JSON object returned when we determine the link is not in the index. On the other hand, we determine that a link is in the index if a JSON object as shown in Listing 2 is returned.

A.2 Monitored Links

For the active measurement study we chose a set of 30 links that we then monitored for a period of 20 days starting April 12th, 2015. Table A.1 lists the URLs for the links that we monitored along with the numeric ID that they were identified with.

Table A.1. Links Monitored: Game of Thrones Study
For the active measurement we monitored a set of 30 links which spanned streaming, cyberlocker, torrent sites. This table lists the URLs for each of the 30 links. This table can be used to cross reference the actual URLs when we talk about the links using the link number

#	URL
1	*http://putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.html*
2	*http://putlocker.is/watch-game-of-thrones-tvshow-season-5-episode-1-online-free-putlocker.html*
3	*http://putlocker.is/watch-game-of-thrones-tvshow-season-5-episode-2-online-free-putlocker.html*
4	*http://putlocker.is/watch-game-of-thrones-tvshow-season-5-episode-3-online-free-putlocker.html*
5	*http://putlocker.is/watch-game-of-thrones-tvshow-season-5-episode-4-online-free-putlocker.html*
6	*http://streamtuner.me/game-of-thrones-season-5-episode-1-the-wars-to-come/*
7	*http://streamtuner.me/game-of-thrones-season-5-episode-2-the-house-of-black-and-white/*
8	*http://streamtuner.me/game-of-thrones-season-5-episode-3-high-sparrow/*
9	*http://torrentdownload.co/Game-of-Thrones-S05E01-WEBRip-XviD-FUM%5Bettv%5D/008B8E8633010064DDBD78609002E5895A18CA64*
10	*http://torrentdownload.co/Game-of-Thrones-S05E03-WEBRip-XviD-FUM%5Bettv%5D/C1F9F6BADE11A4C46028B118053452460AB94BE1*
11	*http://kickass.hid.im/search/game%20of%20thrones%20season%205/*
12	*http://kickass.hid.im/game-of-thrones-hdtv-season-5-ep-1-4-kaugip-t10495684.html*
13	*http://rapidgator.net/file/382aa32709ebe80f0337d271a62920f2/Game.of.Thrones.S05E01.HDTV.x264-Xclusive.mp4.html*
14	*http://rapidgator.net/file/3a571f50b58396ee48d38ab9fb41b49a/Game.of.Thrones.S05E02.HDTV.x264-Xclusive4iPT.avi.html*
15	*http://rapidgator.net/file/773b2c8cf0140c97a033271d41f9ce81/Game.of.Thrones.S05E03.HDTV.x264-Xclusive4iPT.mp4.html*
16	*http://rapidgator.net/file/d2f794e2a029d4d3b07f93fa5c023a6b/Game.of.Thrones.S05E04.HDTV.x264-Xclusive4iPT.mp4.html*
17	*http://rapidgator.net/file/6252815b89a4d87db6f46925069ba553/Game.of.Thrones.S05E02.WEBRip.XviD-FUM.avi.html*
18	*http://rapidgator.net/file/640a13709e0038855499d55bf5e5fa9f/Game.of.Thrones.S05E03.WEBRip.XviD-FUM.avi.html*
19	*http://rapidgator.net/file/ffc5e0ef08e5df9402e4ed152d6aa564/Game.of.Thrones.S05E04.WEBRip.XviD-FUM.avi.html*
20	*http://rapidgator.net/file/56fd8496d19cd8e25c4e254a30f0c784/*
21	*http://rapidgator.net/file/020a73ea5ce999f953a3443135e70d19*
22	*http://kickasslink.com/game-of-thrones-season-5-episode-1-the-wars-to-come-mp4-t10494255.html*
23	*http://torrentz.eu/eeaf9109dd6a44297f7303480220b74bba3501c0*
24	*http://torrentz.eu/1ff137dd8a31b665dc98167a7116293d85dd9bc3*
25	*http://torrentz.eu/7089350db39eda93bc703371f83c18ba15453652*
26	*http://torrentz.eu/search?q=game+of+thrones+season+5+episode+1*
27	*http://torrentz.eu/search?q=games%20of%20thrones%20season%205*
28	*http://torrentz.eu/sea/season+5+game+of+thrones-q*
29	*http://torrentz.eu/searchA?f=game+season++5*
30	*https://torrentz.eu/search?q=game+of+thrones+season+5+e02*


+You Search Images Maps Play YouTube News Gmail More
SafeSearch is off
Web Images Videos News Shopping Maps Books
Any language
Search English pages
Any time
Past hour
Past 24 hours
Past week
Past month
Past year
All results
Verbatim
Reset tools
Google Home Advertising Programs Business Solutions Privacy Terms
About Google

info:http://putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-put

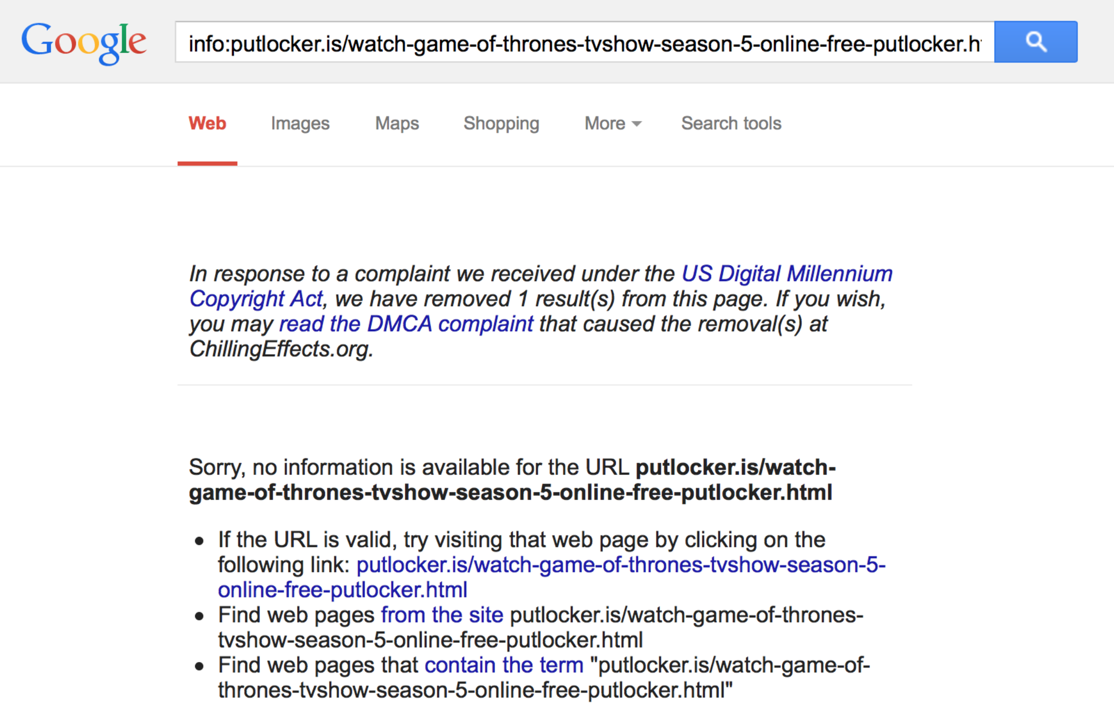
Search English pages

Your search - **info:http://putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.html** - did not match any documents. Reset search tools

Suggestions:

- Make sure all words are spelled correctly.
- Try different keywords.
- Try more general keywords.

Figure A.2. Example of a Link Not in the Google Search Index



info:putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.h

In response to a complaint we received under the US Digital Millennium Copyright Act, we have removed 1 result(s) from this page. If you wish, you may read the DMCA complaint that caused the removal(s) at ChillingEffects.org.

Sorry, no information is available for the URL **putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.html**

- If the URL is valid, try visiting that web page by clicking on the following link: putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.html
- Find web pages from the site putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.html
- Find web pages that contain the term "putlocker.is/watch-game-of-thrones-tvshow-season-5-online-free-putlocker.html"

Figure A.3. Search Query with the DMCA complaint snippet

```
{
  "d": {
    "results": [
      {
        "__metadata": {
          "uri": "<search_query>",
          "type": "WebResult"
        },
        "ID": "9068ab9b-c94e-4291-8a3b-f9163554faea",
        "Title": "Game.of.Thrones.S05E03. ...",
        "Description": "Game.of.Thrones.S05E03. ...",
        "DisplayUrl": "<url>",
        "Url": "<url>"
      }
    ]
  }
}
```

Listing 2. Example of a Link in the Bing Search Index (trimmed)

```
{
  "d": {
    "results": []
  }
}
```

Listing 3. Example of a Link not in the Bing Search Index

Bibliography

[1] Arnold P. Lutzker. *Primer on the Digital Millenium*. Lutzker & Lutzker LLP, http://goo.gl/ISqamh, March 1999.

[2] American Library Association. DMCA: The Digital Millennium Copyright Act. http://www.ala.org/advocacy/copyright/dmca.

[3] The Pirate Bay. About The Pirate Bay. https://thepiratebay.se/about.

[4] Electronic Privacy Information Center. RIAA v. Verizon. https://epic.org/privacy/copyright/verizon/, April 2004.

[5] David Y. Wang, Stefan Savage, Geoffrey M. Voelker. Cloak and dagger: dynamics of web search cloaking. In *Proceedings of the 18th ACM conference on Computer and Communications Security*, 2011.

[6] Internet Policy Task Force Department of Commerce. Copyright Policy, Creativity, And Innovation in the Digital Economy. http://www.uspto.gov/sites/default/files/news/publications/copyrightgreenpaper.pdf.

[7] EFF. Digital Millennium Copyright Act. https://www.eff.org/issues/dmca.

[8] EFF. In Hotfile Docs, Warner Hid References to Robots And Its Deliberate Abuse of Takedowns. https://www.eff.org/deeplinks/2014/10/hotfile-docs-warner-hid-references-robots-and-its-deliberate-abuse-takedowns, October 2014.

[9] EFF. Absurd Automated Notices Illustrate Abuse of DMCA Takedown Process. https://www.eff.org/deeplinks/2015/02/absurd-automated-notices-illustrate-abuse-dmca-takedown-process, February 2015.

[10] Chilling Effects. About Chilling Effects. https://www.chillingeffects.org/pages/about.

[11] Torrent Freak. Anti-Piracy Groups Want Google to Lift DMCA Takedown Cap. http://torrentfreak.com/anti-piracy-groups-want-google-to-lift-dmca-takedown-cap-130219/, February 2013.

[12] Torrent Freak. 'Game of Thrones' Most Pirated TV-Show of 2014. https://torrentfreak.com/most-pirated-tv-show-of-2014-141225/, December 2014.

[13] Torrent Freak. Key Pirate Bay domains must be seized, court rules. https://torrentfreak.com/key-pirate-bay-domains-must-be-seized-court-rules-150519/, May 2015.

[14] Google. Google Transparency Report. http://www.google.com/transparencyreport/.

[15] Google. An update to our search algorithms. http://insidesearch.blogspot.com/2012/08/an-update-to-our-search-algorithms.html, August 2012.

[16] Jennifer M. Urban and Laura Quilter. Efficient Process or Chilling Effects - Takedown Notices under Section 512 of the Digital Millennium Copyright Act. *Santa Clara Computer & High Tech Law Journal*, 621(22), 2005.

[17] Legal Information Institute, Cornell Law University. 17 U.S. Code Section 512. https://www.law.cornell.edu/uscode/text/17/512.

[18] Marjorie Heins and Tricia Beckles. Will Fair Use Survive? *Brennan Center for Justice*, 2005.

[19] Daniel McKenzie. Hotfile is Out Cold, But Googles DMCA Safe Harbor Debate Is Heating Up. *http://techcrunch.com/2012/03/25/hotfile-google-safe-harbor/*, 2011.

[20] Peter S. Menell. Jumping the Grooveshark: A Case Study in DMCA Safe Harbor Abuse. *http://ssrn.com/abstract=1975579*, 2011.

[21] Michael Piatek, Tadayoshi Kohno, Arvind Krishnamurthy. Challenges and Directions for Monitoring P2P File Sharing Networks or Why My Printer Received a DMCA Takedown Notice. In *HotSec*, 2008.

[22] Daniel Seng. The State of the Discordant Union: An Empirical Analysis of DMCA Takedown Notices. *Virginia Journal of Law and Technology, Forthcoming*, 2014.

[23] Spark. Spark, Lightning-fast cluster computing. https://spark.apache.org/faq.html.

[24] Western Division United States Distruct Court, Central District of California. Perfect 10 vs Giganews. http://goo.gl/Stn8Wn.

[25] NBC Universal. In the Matter of Broadband Industry Practices. Comments, Federal Communications Commission, 2007.

[26] U.S. Copyright Office, http://www.copyright.gov/legislation/dmca.pdf. *The Digital Millennium Copyright Act of 1998 Summary*, December 1998.