To my

Beloved father and grandfather who left me during the writing of this thesis, may their souls rest in peace; also to my beloved mother, wife and lovely children, from whom I have been separated for the past year of work on this thesis.

AutoURL: Automatic URL Tracking

to identify rogue advertising

This thesis is submitted in fulfillment of the requirements for the degree of

MASTER OF INFORMATION SCIENCES in

SOFTWARE ENGINEERING

School of Engineering and Advanced Technology (S.E.A.T)

at Massey University, [Albany],

New Zealand.

by

Mohammad Rashid Ourahman, BEng (Massey University)

January 2015

Acknowledgements

I express my deepest appreciation to Prof. Paul Watters for giving me the opportunity to complete this thesis through helping, guiding and motivating me during the entire processes of researching and completion. Furthermore I also acknowledge with much appreciation the love and support of family and friends who were not given their rights to their affection and left alone for almost one year. Last but not least, many thanks go to my friends and staff at Massey University who shared lovely moments and didn't hesitate to show me the right directions whenever I asked them any question during the entire process of writing this thesis.

Abstract

Maintaining Cyber Security has been one of the biggest challenges of a modern era which has seen the extensive emergence of internet advertisers, and in which some promote their malicious contents through rogue websites.

Internet rogue advertisers penetrate through cybercrime in various forms of advertisement banners which are displayed within any parts of a website. Tracking these rogue advertisers is important to the Cyber Security cause, where in an ideal scenario individuals are exposed to correct information as is their basic right, along with their reaction toward the sensitivity of any content.

In the past manual tracking has been the commonest method of checking but in some cases manual tracking could fail, other than time parameters the accuracy is also questionable, the solution to this the concept of Automatic URL Tracking.

This thesis represents an analytical method of Automatic URL Tracking, according to this approach, where various pages are checked for advertising banners, these are clicked until the final URL or its destination is reached.

To achieve various concrete results a significant work has been done to develop an Automatic URL Tracking Software which is run when connected through internet while holding the reported URLs databases where each of these are tracked to its final destination.

The Automatic URL Tracking Software was run for the total of 2500 URL samples, upon manually tracking these URLs the two processes showed 87.7 % agreement which can be reliable result considering the presence of various blocking techniques adopted by hosting sites and site developers but there are chances for further development where the application is enhanced specifically to overcome these obstacles.

Automatic URL Tracking overcomes the difficulties and challenges of manual tracking, allowing larger data volumes to be tested, identified and verified, but having said that it also comes with the challenges of rapidly changing internet technologies, in which more comprehensive strategies need to be built to overcome this challenge.

Table of Contents

List of Figures

List of Tables

CHAPTER ONE

Introduction

The current digital era of information technology and emerging internet usage and its penetration within social, economic and political aspects of human life and process has brought many security risks which could not have been thought of or evaluated a few years ago (Paris, 2001).

The integration of Information Technology to various aspects of life has a revolutionary impact and has also increased the dependency on social inter-relations of daily life, but it has also brought with it various challenges because of which IT users feel insecure and discuss their future in doubtful terms (Gerrard et al, 2006).

This rapid IT integration has widened the use of the internet, rapidly expanding it into people's daily routines. This has resulted in opportunities for the criminals to establish criminal hubs online, where various versions of thefts are carried out online. These include identity theft and stealing personal information to carry out online hacking, stealing money through internet banking, denying legitimate services, virus attacks, high risk advertising, copyright violations and selling legally challenged products through rough sites which are directed and forwarded through fake advertisement links (Felt et al, 2011). The damage is not only restricted to the financial side of society but it also creates a disturbance through the exposure of sensitive material to the wrong audiences, which may be outside the appropriate age range or may be influenced wrongly, which results in social disorders and increases in various other harmful elements that can directly influence socio-economic factors within any developed society (Cullen, 2001).

Reviewing current global digital and data rules safeguarding intellectual and private property, financial information and company reputation is a crucial part of every organization where data sensitivity and its leakage implications are always seen as a threat to the existence of any organization (North &Kumta, 2014).

The protection of organizations, business entities, and social values comes when intellectual properties are protected along with each and every one's personal privacy, where privacy is could be defined as -- the right to be let alone[1].This term means the right to enjoy life, free from outside disturbance and unobserved by others. This is seen as the basic right of every individual citizen, because people do not want to share their private information. This touches the essence of human identity (i.e. personhood) at the point where individuals can be harmed or debilitated. Unrestricted public access to and use of personal information and cases of breach of privacy can result in surrendering private information into the hands of the wrong people. This could cost some one's life, dignity, respect or even major family issues which then leads to multiple social complications. However the leakage of personal information could lead to a lack of personal expression, creativity and growth, which in turn can affect one's very identity.Furthermore privacy is also about securing the leakage of data contributors to the data analyzers to maintain the ownership of manipulated information (Gehrke& Pass, 2011).

In 2013, upon approval of US legislation to collect phone data on millions of Verison users, in an angry response Al Gore stated; in digital era, privacy must be a priority. Is it just me, or is secret blanket surveillance obscenely outrageous?[2].

---

1 - http://groups.csail.mit.edu/mac/classes/6.805/articles/privacy/Privacy_brand_warr2.html

2 - http://www.nationaljournal.com/nationalsecurity/al-gore-s-pissed-about-the-nsa-spying-and -other-tweets-you-might-have-missed-20130606

There are further fundamental reasons in cases wherevarious commercial and non-commercial (including governmental) organizations hold highly confidential and sophisticated information. Any leakage would have devastating financial and security consequences in which the existence of that particular organization could be under threat. Furthermorethis leakage could lead to the total collapse of every detail of its processes, as well as individual physical and psychological wellbeing (Tahboub&Saleh 2014).

Civilization is the progress toward a society of privacy. The savage's whole existence is public, ruled by the laws of his tribe. Civilization is the process of setting man free from men (1973, Ayn Rand)[3].

Therefore, even if information and intellectual property privacy is not in itself a fundamental right, rather it is necessary to be protectedand secured for the existence of particular organizations and the wellbeing of every single individual on this earth in the current digital society. To fulfill this need for protecting security and privacy new research areas have been introduced which are normally referred to as Cyber Security (Samuelson, 2000).

It is vital to act in the face of privacy violations and safeguarding privacy must be made a priority of continuing research and technical techniques. As Bill Nelson has said; if we don't act now to safeguard our privacy, we could all become victims of identity theft[4].

---

3 - http://freeradical.co.nz/content/58/Privacy.php

4 - http://www.billnelson.senate.gov/newsroom/press-releases/key-privacy-advocates-seek-to-crack-down-on-id-theft

1.2 Cyber Security:

The term Cyber Security could have different meanings for various audiences depending on the nature of the stakeholders. Where normal citizens or individuals mayperhaps feel that it is related to protecting personal information, commercial and governmental organizations may view it as a means for providing business continuity, security and protection of their intellectual information (Von Solms&Van Niekerk, 2013).

Cyber Security has a much broader meaning than just computer and network security, though the core principles behind computer and network security continue to be the main goal. The context is much wider, and typically relates to threats posed to critical infrastructure (Watters, 2012). Thus, the concept of Cyber Security involves explicit acceptance of the fundamentally insecure nature of internet protocols and desktop computers which comprise the bulk of internet traffic and act as potential vectors of attack against critical infrastructure (Watters, 2014).

In a general context, Cyber Security is the collection of all activities using various resources that allow citizens, organizations, business and governments to perform their Information Technology and Internet tasks in a manner in which security, privacy and reliability are seen as major playing factors. However safe use of the internet is a topic which has come into prominence in the past few decades and it has generated the headline - Cyber Security. The term centers on how users (or internet browsers) can safeguard their personal information when it is stored in various forms of media or connected to outside through different communication protocols, or even when stored in any of the cloud storages, where accessibility to hackers from rogue organizations is not impossible (Ralston & Graham & Hieb, 2007)

The concept of Cyber Security involves an explicit acceptance of the fundamentally non-secure nature of internet protocols and desktop computers which comprise the bulk of internet traffic and potential vectors of attack against critical infrastructure (Watters, 2012).

There are many forms of Cyber Security breaches and challenges and various countries face unique combinations of a few types. The factors that contribute to these challenges are critical, especially for developing nations, since, notably, the benefits and risks of technological development are not always shared equally. In a situation cyber community change is occurring at an unbelievable pace, bringing with it equally unbelievable threats, there are several reasons why Cyber Security is needed (Homan, 2010).

Global terrorism has raised the possibility of activities that various notorious organizations could indulge in or their main reason to use internet, the increased threat of terrorism and rogue governments, has also increased other security concerns (Chin & Ham, 2015), these inhuman and un-natural forms of act, creates disorder in the social life of fellow human beings (Janczewski, 2007).

Figure 1: Shows various forms of Cyber Crimes[5]

5 - https://www.enisa.europa.eu/media/news-items/council-of-europe-and-enisa-on-cyber-security

Cyber criminals involved in violating and attacking other sources are always technically advanced people who constantly find ways to penetrate the targeted systems and create the desired mess in a much shorter time than any victim could have expected. Their technical advantage and expertise endangers users through creating more sophisticated attacks. This provokes the users to be always on the alert and to guard their systems through every single means of protecting their personal information (Moore, Clayton & Adnerson, 2009).

The threat to Cyber Security in various parts of the world could differ, just as its influence (or social suffering) may differ in nature. Some countries suffer

Mostly from various forms of virus, malwares and other types of worms, while others suffer more from rogue internet advertising (i.e. fake advertising). This is normally found through rogue websites which could cause tremendous social disorder (Chen, Chen, Huang & Ching, 2006).

Recent studies have highlighted the role that internet advertising plays in supporting the revenue of rogue websites (Taplin, 2013; Watters, 2014).

1.3 Problem Domain:

The raise in internet usage due to its main source of information and entertainment has attracted various rogue advertisers to target main stream audience and in return earn fortune with very little outgoings.

The main cause behind attracting the rogue advertisers are human dependency of the internet and internet availability and on the same time high revues in internet advertising has motivated various marketing companies to use their services due to its immediate and live availability.

The above facts have triggered academic researchers and governmental agencies to understand and track illegitimate online activities, these facts along online advertising (and their types), online spending and advertisement revenues are discussed in Chapter two (i.e. Literature Review) of this thesis.

1.4 Research Outline:

This research has been carried out as part of continuation work by Professor Paul Watters (Massey University, Auckland, New Zealand) where he has done highly effective work to highlight the social and economic role that internet advertising and its effects, along with various analyses extracting and profiling the advertisers from the DMCA complaint samples.

This research started with analyzing, designing and implementing customized software "Auto URL Tracking", which uploads (or inserts) the selected sample of DMCA reported URLs from the raw data (i.e. CSV) data file, which comes from Google Transparency Report, then adding and developing the additional functionality of displaying, advanced search and exporting the selected data to a user defined database name, this is when the user is confirm for automatically tracking each and every single URL links within the sample of selected DMCA complaint pages; this functionality of automatically tracking URL links is the main functionality of the developed application.

Among the various functionalities of the "Auto URL Tracking" software the feature of identifying and extracting the URLs from the DMCA complaint page was developed by Watters, this is when it was inherited, modified and further developed to add a detailed Graphical User Interface and the core of the program which is automation process for tracking URLs within each sample of DMCA complaint pages.

The thesis addresses the following research questions:

- To what extent are automation and auto classification possible through demonstrating an empirical means of accruing an automatic system for recognizing and confirming larger data sets?
- Why is an automated system important for building a strategy to increase accuracy and avoid human dependency?

The main concept and idea always centers on the fact that tracking every website manually is extremely time consuming and in some cases may not be possible where

URLs redirect at a high speed, faster than human eye can detect, not forgetting the fact that the larger data set (or databases) holding thousands of records cannot or maybe should not be manually tracked as this would require very extensive use of time to implement. Accuracy of the testing is another big question.

The process demonstrated in this thesis aims to analyze and answer the two questions of how and to what extent the automation process could be adopted and what would be its accuracy level, and on the other hand, what are the important facts to do with current status and means of enhancement?

The data used for this analysis is collected from a DMCA report which is a database-of-complaints file upheld by Google (through their Transparency Report). This data comes in a CSV (i.e. Comma Delimited Values) file which is downloaded and used for observing and analyzing certain facts and results to enable answering of the above questions.

Other chapters of this thesis are as follows:

This chapter (i.e. Chapter One) gives a general overview of Cyber Security, its challenge and the main conceptual overview of the problem domain, outlining and defining the thesis scope and its range.

Chapter Two defines the methodologies and techniques employed in carrying out the research through data source and collection details.

Chapter Three reiterates the literature review of the entire related work carried out by other academics and researchers.

Chapter Four advances and focuses on the actual research topic, which is automatically tracking the copyright complaints reported to the DMCA database. This talks about the developed application and the technical aspects around its development.

Chapter Five analyses and compares the test results produced by the automated application versus the manual testing at the point at which a defined conclusion could be drawn.

Chapter Six draws the overall conclusions, experiences and patterns which are extracted as a result of carrying out the research, where the thesis questions are answered.

Chapter Seven gives a future overview of the project and relates to question two of the thesis where a discussion is carried out for developing a strategy for further development and maturing the process of automation.

1.5 Thesis Scope:

This thesis starts with a common understanding of Cyber Security, where it covers all the related topics and definition (with explanations) of the concepts used around the topic, then focuses on the automated process of URL tracking where certain data are observed and analyzed to achieve the answers to the questions raised.

The main goal behind this research and thesis presentation is to answer the main questions: to what extent are automation and auto classification possible through demonstrating an empirical means of accruing an automatic system for recognizing and confirming larger data sets? and secondly why is an automated system important for building a strategy to increase accuracy and avoid human dependency?

To achieve the above goal within the defined scope, the following objectives must be achieved:

- Brief overview of the Cyber Security issues,
- Online spending and advertising revenue,
- Introduction to internet advertising and terms falling under its umbrella,
- How is the research is carried out,
- Detailed functionality of “Auto URL Tracking” Software including its features and technical aspects for further development,
- Test results produced by the “Auto URL Tracking” and discussion for drawing conclusions and future work strategies.

1.6 Contribution

In this thesis, an automation process of auto classification is tested through a fully customized and designed application to demonstrate an empirical means of accruing an automated system for recognizing and confirming larger data sets, where manual process is hugely time consuming and error prone. This analysis is shown through the process of tracking URLs from a selected sample of DMCA complaints to identify and track the reported pages until the actual advertiser is found. The automation of URL tracking is possible at this moment in time but its accuracy is dependent on the strategies built around its further development and maturity of the mechanism.

1.7 Summary:

(Cyber Security is one of the great challenges of the digital world. Surely Cyber Insecurity is the challenge and Cyber Security is the answer. The problem comes in large part from internet advertising which now has a twenty year history, during which cyber criminals have attacked normal civilians in various forms, especially through internet advertising. The high risk advertising is placed under the mainstream advertiser's banner, which victimizes audiences through exposure of highly sensitive information. Furthermore, online terrorism and digital marketing, where rogue organizations have taken illegitimate advantage is another challenge for the mainstream public and governmental agencies.

CHAPTER Two

Literature Review

2.1 Background:

This thesis briefly covers the challenges posed by cyber security then it slowly zooms to online advertising where it leads to automatically tracking the rogue websites which is the main questions that call for testing and analysis of automatic URL tracking from the formally reported database of the Google Transparency Report, these reported sites are the copyright violations through various forms of online advertising or internet advertising.

The penetration of modern Information Technology to daily life of every citizen has motivated various marketing companies to advertise online in order to capture the desired audience but this has not come cheap, this has provoked various illegitimate groups (or maybe small organizations) to place ads as “banners” on a host website, as mediated by an advertising network where the heading (and front message) is different than the actual content or final destination upon following or clicking the link, these sites bring reviewed based on their adopted revenue model (e.g. the host website is paid per-view, per-click or per-sale, every time a user views the page and/or engages in one of those activities).

The high dependency of human life on internet where it is replacing the normal cable entertainment, banking and even tradition site shopping has increased the chances for rogue advertisers to appear and do their part of capturing the desired audience. These advertisers are breaching the law to every extend but their end (or stopping them) is fully dependent on their residing sates laws along their actual location for their hosting servers (or website) as various states (or countries) have different laws toward internet advertising which is practiced by these groups or advertisers.

The internet advertising is increasing every day and this progressive increase in stakeholders and revenue has motivated various advertisers to participate but at the cost of multiple social, economic and political right violators who breach every right of normal citizens in order to increase their earnings in any form even if illegitimate ways are adopted (De Mooij, 2013).

2.2 Internet Advertising:

Internet advertising is the act of placing various forms of advertisement on the internet for the purpose of selling any kind of product or service. This is a deviation from cable advertising and the traditional style of advertising through TV, Radio, News Paper or Magazines. In saying that, the internet advertising is wider and more open compared to traditional forms. Internet Advertising is also referred to as Online Advertising or Online Marketing by various stakeholders (Sharma, Chowhan, Singh, Gupta& Srivastava 2015) and it may involve legitimate or illegitimate advertising where various advertising networks play their role as a middle man and connect the advertisers and hosting sites together, as brokers in the middle. Furthermore in most simple words Internet Advertising could be defined as; the most recentmethodof marketing any products or service is marketing it through the internet rather traditional cable or paper advertising (Robinson, Wysocka& Hand, 2007).

The internet advertising is one of the fast growing businesses which have proven its influence on the digital economy and its presence is very important for the media industry where it enjoys a big influence from the fast growing internet technology which also leads to a developed digital economy and social development (Yuan, Abidin, Sloan, & Wang 2012).

Internet Advertising is a fresh concept which is the direct result of emerging Information Technology and the digital societies whose history goes back to last century, have made an enormous impact on the world of internet advertising and marketing, where various vendors use variety of means to sell their products to users by an internet medium which is open and could easily be viewed anywhere anytime at all the times

upon the availability of the internet. This feature of full-time availability has attracted many marketing agencies to concentrate on the internet advertising which has directly resulted in sharp incline to the advertisers' revenues. These internet advertisers earn an extensive amount of money through various means. For example, in 2011 internet advertising exceeded the cable television revenue (IAB Internet Advertising Revenue Report, 2012). This tremendous increase in revenue opens the door to all cyber criminals (IAB Internet Advertising Revenue Report, 2012).

Looking into some revenue data fact sheets and reports we find that in 2013 the revenue for internet advertising in the United States was $42.8 billion, a 17% increase compared to the $36.57 billion in revenues in the year 2012 (IAB Internet Advertising Revenue Report, 2012). Another surprising fact about internet advertising is that United States internet advertising revenue hit a peak at $20.1 billion for the beginning of the year 2013 and again showed increases of 18% over the previous year's numbers (IAB Internet Advertising Revenue Report, 2014). This extensive and progressive increase in internet advertising revenues has some negative consequences, exposing the system to further criminal opportunists. Imagine if there were no security forces or police officers in our streets? What would the life of a normal citizen be like? Similarly, in the case of dishonest internet advertisers, they will not hesitate to use any means, legitimate or illegitimate, to promote their product or service to portray their message across the audience. Online advertising is widely used across virtually all industry sectors (IAB Internet Advertising Revenue Report, 2013).

One result of this massive rise in internet advertising revenues is that it encourages higher budget allocation for online marketing, this suggests a stable and progressive media platform on which advertisers and advertising networks can base additional advertising in different ways to meet various customers' expanding needs and demands. However, once again, due to the nature of its openness and visibility to all kinds of audience, inappropriate aspects of the internet could be displayed to unsuitable audiences due to their age, gender, culture, political or religious beliefs. Furthermore the internet advertisements have opened an unlimited and wide access to all groups without any restrictions, but rather with a high risk of unsuitabilitywhich could lead to various social and psychological disorders as a result (O'Neill, 2014).

Another challenge to Cyber Security is the operation of rogue sites which are seen as direct result of these attractive internet advertising revenues through which various illegal and sensitive products could be commercialized. These products could include copyright violations through illegally selling various copyrighted products, gambling, nudity sites and illegally supply drugs which may be even counterfeited. However most of these sites are hosted in states outside the developed countries (Crosse, US Government Accountability Office & United States of America, 2014).

Many businesses and organizations after being exposed to these eye catching revenues from the internet advertising are switching off from the traditional offline advertising to internet advertising and are using various techniques to drive their customers to websites. This is normally done through using various keywords in search engines when applying SEO (i.e. Search Engine Optimization) techniques and using other means of advertising that includes advertising banners, advertising pop-ups and pay-per-click ads or pay-per-call links (Khraim, &Alkrableih, 2015). These advertisements need to be placed in accordance with similarity of pages, which means they are to be placed where they can be noticed by the targeted customers. This involves selection of the placement based on the target customer's age, interests and gender. This classification of the advertisement and their placement based on their personal information increases the chances of building an audience. Furthermore information about the visitors is also being collected by the internet advertisers through the communicating application (i.e. browser) which supports the advertising network and advertisers in search to target the right audience. However the latest emergence of social networking sites (e.g. Facebook) has also occupied the center of e-commerce which broadly contributes to the raise of internet advertising (Zeng & Dou, 2009).

The internet advertisements are of different types and these are placed on websites in various forms, but once they are clicked, this redirect the user to another site which is different from the first visual appearance and a total deviation from the initial suggestion and intention of the advertisement. This not only reveals the fake side of the advertisement but initiates another form of dis-trust of the internet and advertisers. Furthermore, advertisers and advertising networks have various display patterns, in which advertisements on the same page are randomly changed through various

techniques depending on the type of the plan that an actual advertiser has chosen or subscribed to when the advertisement was requested, not forgetting the fact there are advertisers who manage their own advertisements (i.e. banners) through various third party advertisement networks (McCoy, Everard, Polak, &Galletta, 2007)

Cyber Security is one of the main topics in various governments' agenda and there are various strategies built to tackle cybercrime, the United Kingdom government raised Cyber Security as one of the four risks in relations to United Kingdom's national security risks in 2010 (General, 2013).

Cybercrime effects states economy and its reputation, to encounter this governments and non-governmental originations enhance their security level to encounter the Cyber Security to promote country's white economy (Iles, 2014) as Cyber Security intends to promote mainstream advertising rather high-risk advertising, where high-risk advertising supports grey market and the main cause to white economy destruction (General, 2013).

The growth in internet advertising revenues has seen a remarkable growth every year (Dreze&Zufryden, 1998) while accompanied by the growth of social media and increase in internet use the digital advertising has been even more attractive (Sinclair, 2015) these attractive revenues provide the commercial motivation for criminals to increase their online activities in form operating the rogue websites (Watters, 2014).

2.2.1 Types of Internet Advertising:

Internet advertisement (or simply online advertisements) has penetrated deeply into various aspects of business, organizations and governments. These vendors build their marketing strategies around internet advertising to promote their products and services; this internet advertising has become an important part of the economic life (Goldfarb, 2014).

Many forms of internet advertising are currently available on the web. In order to establish basic understanding of them, the list below[6] shows the most commonly used

internet advertising. These forms are types of intended mainstream advertising; in saying that any violations of their policies might result in removal of the advertisement, these also promote the legitimate products and services which are the main reason behindstates white economy.

6 - http://www.wordstream.com/online-ads

Table 1: List of common form of Internet Advertisements[6]

No	Name
1	Google Search Ads
2	AdWords Ads
3	PPC Ads
4	Bing Ads
5	Facebook Ads
6	Twitter Ads
7	Tumblr Ads
8	Banner Ads
9	Google Display Ads
10	Retargeting Ads
11	Reddit Ads
12	Mobile Ads
13	In-Game Ads
14	AdMob Ads
15	Email Ads
16	Gmail Ads
17	Video Ads
18	YouTube Ads
19	Pinterest Ads
20	Instagram Ads
21	Vine Ads
22	TV Ads
23	Newspaper Ads
24	Radio Ads
25	Urban Ads

The most commonly used method of all is the Google Search Advertisement. According to eMarketerInc research firm the Google Search Advertisement controlled 71.2% of online advertising search market in the year 2007 which is almost the three quarters of the entire market, this rise and market control has also angered its big customers (Steel, 2008).

For an English setting browser these advertisements normally become visible on the right hand side of the Google Search Engine. The nature of revenue for these advertisers is through PPC (i.e. pay-per-click), meaning the advertiser would pay for every click that any particular audience makes or tries to visit in the advertised site.

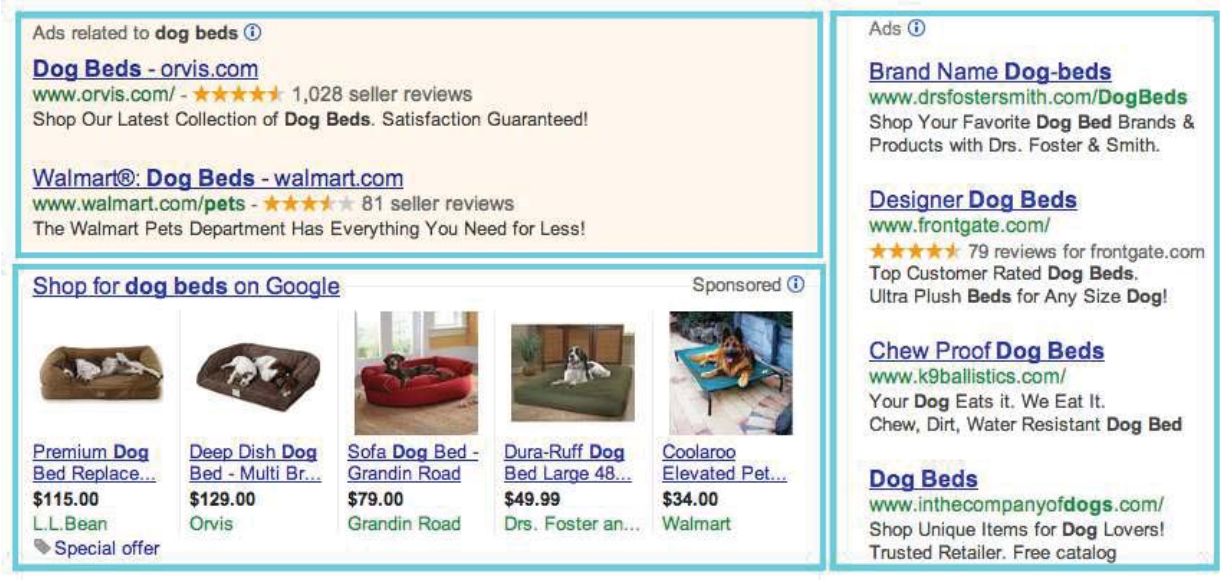


Figure 2: An Example of Google Ads

The AdWord advertisements are identical and appear next to Google's Search advertisements and are created inside Google's AdWords advertising platform (O'Dwyer& Moyle, 2014).

Figure 3: An Example of AdWord Ads

Facebook Advertisements are available in various forms, each having positive and negative sides for each specific type of business, depending on the nature of their produce or services. The progressive social networking site usage has also boosted the Facebook Advertisement, studies show in the past years some Facebook advertisers have increased their expenditure by 10-fold (Womack, 2010). These advertisings appear on the right-hand side of the page with various details for the actual advertisement.

Figure 4: An Example of Facebook Ads

The FBX (i.e. Facebook Exchange) was launched in 2012 and it carries advertisements which appear based on the user's surfing history. These are normally gathered through collecting data from the cookies file created by the browsers (Funk, 2013).

Figure 5: An Example of Facebook Exchange Ads

Twitter advertisements appear on twitter accounts. These are paid advertisements which boost that particular product or service on Twitter to increase the number of tweets. These can be in different forms, such as promoting accounts and following up to extend the attention of the various audiences and followers. (Cheng, Bansal&Koudas, 2013).

Figure 6: An Example of Twitter Ads

The Retargeting ads look into users' previous web history. This information is compiled through collecting various data parameters from the users' enabled cookies (Tran, Acs&Castelluccia, 2014).

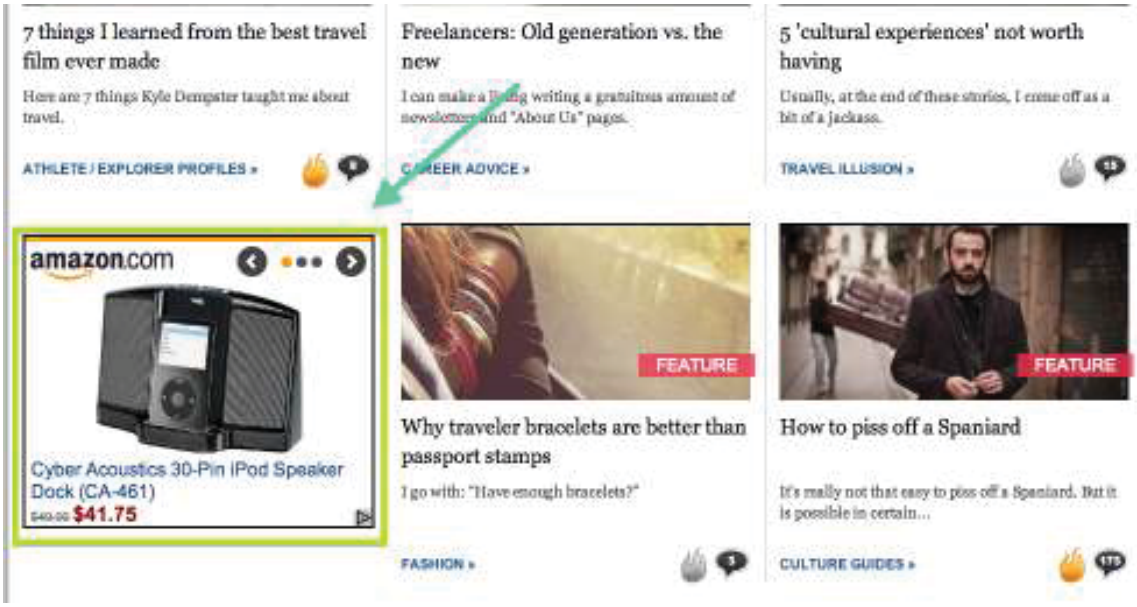


Figure 7: An Example of Retargeting ads

Another most commonly used method for internet advertising, which is the main topic of this thesis, is banner advertisements. These are images that could appear on side, the top, or even the bottom section of the web pages. These bannerscome in various sizes and are commonly found on web based communities, blogs and free online service whose main objective is to drive audiences to their site for free but which make their revenues using advertisement banners. There are various mainstream advertisers who also use this form; examples in New Zealand market would be Trademe and 1-Day sites (Kamen & Shirman, 2014).

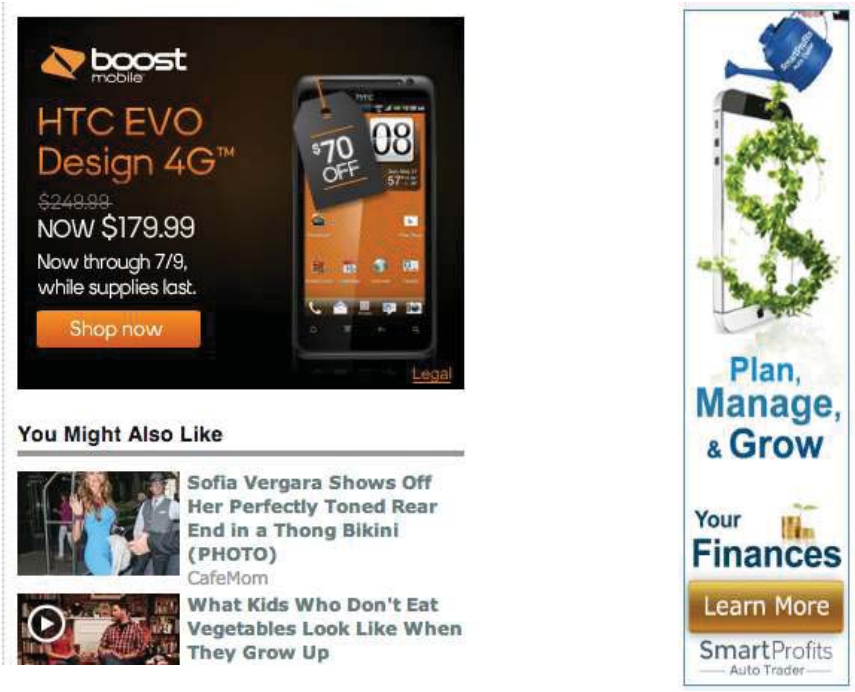


Figure 8: Examples of Banner Ads

Mobile advertisements have emerged extensively recently, mainly through free applications offered through tablets, smartphones and other mobile devices. These advertisements also have security issues and normally on the free downloaded applications (Gao&Zang, 2014).

Figure 9: An Example of Mobile Ads

There are many more examples which have been named in Table 1 of this chapter but the commonest examples have been brought into the discussion to avoid going outside the thesis scope, and to give the reader just a general overview of the internet and online advertising rather than an explicit or detailed explication, while at the same time indicating that all these details are available in various sources on the internet.

There are many more terminologies which are encapsulated in the phrase internet advertising, some of these terms are defined in the upcoming paragraphs.

2.2.2 Mainstream Advertising:

As the word mainstream refers to normality, in its context the compound word stands for a typical advertisement placed by appropriate and genuine commercial entities that play a legitimate part of the legitimate economy (Watters, 2014). Businesses promoting these advertisements would only accept and practice business within the legal constraints and would reject any goods or services whose legitimacy or copyright might be questionable.

Advertisers hosting these advertisements reject anything that falls into the category of the underground economy and legality of the actual advertisement is their priority.

Table 1 of this chapter shows the intended forms of mainstream advertising; meaning upon violation of any of their policies advertisement removal might take place.

2.2.3 High-Risk Advertising:

High-Risk advertising is placed by rogue advertisers whose main intention is to avoid normal, legal or legitimate advertising. These promote illegitimate goods and services which cannot be hosted within any white market (Watters, 2014).

The products mainly promoted by these advertisers maybe illegal or restricted according to law. They could be counterfeit products or any other product which may have an age restriction on it before its exposure. Examples of these advertisements may include nudity and sexual sites, gambling or types of rogue software which downloads viruses or malware to any particular computer machine (Yu, 2015).

The high-risk advertisement could be categorized in various forms; some of these are as followings:

- Copyright violated products (i.e. Pirated Movies and Audio CDs)
- Online gambling sites
- Fake software (i.e. applications), normally downloads viruses and malwares to the browsing machine
- Scams in various forms (i.e. Fake Prize Competition, Fake Employment Opportunities, Investment Frauds)
- Sexual contents in various kinds and forms

2.2.4 Advertising Networks:

These are business and commercial entities which connect the advertisers to the sites where advertisements can be placed or hosted. Such networks analyses customer needs and based on the nature of their business they suggest and promote a particular host or form of advertisement. Furthermore these networks can act like broker agencies which connect the advertisers to the actual hosting sites (Schaeffler, 2014).

Some common examples of advertising networks are; Google AdSense, Google AdWords, Yahoo! Advertising Network, Microsoft Advertising adCentre and Microsoft Ad Network but Google AdSense and Google AdWords are the most widely used advertising networks or platforms (Shah, Banga, Sampat& Patel, 2015).

2.2.5 Internet Advertisers:

These are commercial or even governmental entities or associations which desire to sell their products or services or even merely spread information to the targeted audience through the means of internet advertising.

There is a tremendous shift from traditional to internet advertising by various advertisers, a phenomenon which plays a major role in boosting internet advertising revenue.

2.2.6 Rogue websites:

As the name suggests, these websites provided to the user (or internet browser) with dishonest material. They host inappropriate or illegitimate products on their websites which is uploaded either through intention browsing or through a link which has been clicked to be at their website. In normal scenarios an appropriate advertisement (through link or banner) is clicked or followed those visual suggestions differs than from the final destinations, this is when it leads to these rogue websites.

The main ambition and motivation for browsers is to gain access to the illegitimate content which is not included in the mainstream sites. Furthermore the rogue websites host various forms of stolen streaming and other fake products and prescriptions which at the peak of copyright violations do huge harm to the legitimate economy. These sites are hosted in various parts of the world and follow various techniques to deceive the internet users through exposing the legitimate face of well-known businesses to achieve their credit card or bank details, and therefore their funds (Poulter, 2012).

The rogue websites pose potential harm to business, societies and the states white economy, where businesses are dependent on intellectual property ownership in which enormous amount of money is spent on research and development of any product. However, these rogue sites sell what appears to be the same product at much lower cost, reducing business trust and revenue. Furthermore they play a big part in wearying the revenues of streaming products such as Movies, TV shows and other forms of legitimate song albums by promoting their availability on these illegitimate sites (Watters, 2014).

The real threat of rogue sites has brought together academic researchers, commercial entities, organizations, law enforcement units along with various kinds of stakeholders, seeking to build between them the technology and strategies through which the identification, and tracking of these rogue websites becomes possible (Katin-Borland, 2012).

Figure 10: Example of rogue site

2.3 Google Transparency Report:

This regularly issued report by Google, includes raw data requested by various brand owners, agencies and even governments. It can relate to users' data, records, or even contents.

This report contains raw data for three CSV (i.e. comma separate values) files and each file contains various levels of details needed to meet the basic requests, as well as links to reported sites which are considered as violations and must be removed.

Google receives requests either directly from copyright owners or on their behalf through an acting agency to remove various search results that are links to the illegitimate sites or violations of the copyrights to products. These requests contain details about the copyright owner and the specific URLs which need to be removed. Google also receives requests from governmental agencies including requests to remove information from products, but each of these requests is reviewed and analyzed before any action is taken to ensure that the content deserves to be removed[7].

This transparency report is used widely by various researchers, agencies and governmental institutions for acquiring various analytical and statistical values from which strategic solutions could be found for dealing with these infringing rogue websites.

This research and thesis also takes DMCA samples from Google Transparency Reports and runs them through the customized in-house built software.

---

7 - http://www.google.com/transparencyreport/removals/copyright/?hl=en

2.3.1 Digital Millennium Copyright Act (DMCA):

The DMCA is an act related to United Sates copyright law which provides assurance against any liability for copyright breaches, when there exist an agreement to cooperate in taking down all alleged infringing materials, especially once it is confirmed and notified by the copyright holders. The Google Transparency Report confirms these requests which have been auctioned on behalf of copyright owners to provide transparency and assurance to their users (Nimmer, 2013).

2.3.2 Chilling Effects Database

Chilling Effects is the project of the Berkman Center for Internet & Society and it is an independent third party research project studying cease and desist letters concerning online content; they analyze complaints about online activity, especially requests to remove content from online, and their main objectives are to provide the raw data to all researchers, business, organizations or even agencies. This data can then be analyzed in order to achieve better online transparency and to ensure that copyright violations are kept to minimum, so that the legitimate economy is supported and the grey market is kept to its minimum[8].

8 - https://www.chillingeffects.org/pages/about

2.4 Increasing Online Spending:

The internet advertising has been the matter of focus and interest due to its high revenue and low operations cost. This has initiated various studies to be carried out by various researchers to highlight the actual role of internet advertising and their generated revenues and no doubt such advertising generates an enormous amount of pure profit as their running (or operating) cost is much lower than a tradition retail or any online business.

Recent investigated study suggests there is harm to users from viewing the increasingly "high risk" nature of advertising being hosted on these sites (Watters, 2013).

There is a continues increase in number of internet users, these numbers are increasing every day which directly effecting the worldwide digital spending because internet has made distanced shorter and it is not only used for reviewing information rather is also hugely used for communication and social networking, in saying that the arrival of smart phone has further increased the use of internet through various means of internet.

According to published numbers by eMarketer's estimations there is a continues increase in digital advertising and in four years the digital advertising could account for 38% of the global spending which is the same as TV advertising which is enjoying the biggest share for more than a decade.

The growth in online advertisement (and spending) is the direct result of increased internet users connecting to the internet through various unrestricted devices around the world using various kinds of internets available to them. Furthermore the arrival of smart phones, data internet and social networking sites has pushed the numbers even higher. According to Statista in 2015 there is a 52.7 percent mobile phone usage while in 2016 it could increase to 56.1 percent these increases show users are switching and starting to use the handheld smart devices including smart phones and tablets which are more portable and their accessibility to data internet through SIM (i.e. Subscriber Identity Module or Subscriber Identification Module) cards.

2.5 Online Advertisement Revenue:

The internet advertising has been the matter of interest due to its high revenue and low operations cost. This has initiated various studies to be carried out by various researchers to highlight the actual role of internet advertising and their generated revenues and no doubt such advertising generates an enormous amount of pure profit as their running (or operating) cost is much lower than a tradition retail or any online business.

The secrete behind success of these online advertisings is the large revenue generated, alone in USA the sum of US$42b is annually spent on online advertising which exceeds the tradition TV advertising. These numbers attract more agencies to shift their focus to online advertising. Furthermore this enormous increase in the focus has pushed the rogue advertising to its peak where advertisers see the high profit margins which are not possible through transparent advertising.

The top three rogue sites in 2014 are The Pirate Bay (Alexa rank 79), Kickass Torrents (Alexa rank 103) and Torrentz (Alexa rank 153).(Watters, 2014).

The enormous profit margins generated from these websites has attracted and motivated various groups or organizations to work in the areas of internet advertising and concentrate on particular content which are forwarded to upon clicking a specific link where needs of their products and services are met.

The main challenge (or issue) is not behind progressing internet usage, internet advertising or spending in the current age of digitalization and globalization rather the main challenging risk is the rogue advertising which targets various audiences with age, cultural and religious sensitive contents.

2.6 Dangers of Rogue Sites:

The increase in internet usage has increased online advertisement and spending but as a result an intolerable risk of rogue advertising has emerged and targeting the undelibrate viewers.

Rogue websites has deviates from mainstream advertising and intentionally displays the inappropriate or even illegitimate content which could also bear high risk to various viewers and certain age group. Furthermore these rogue sites affected the overall response toward online advertising along more worries to parents where certain ways for preventing them has been adopted to minimize the risk.

The rogue sites could engage individuals to high-risk activities such as gambling, scams and even pornography which are highly effective on various ages' behavior especially to younger aged where innocence is at peak, all these could result in social disorders.

The rogue sites also pose threat to branded products, copyright materials and even academic research and publications which indirectly transfers the trauma to state economy.

Rogue sites and their inappropriate contents are also affecting New Zealand market and its economy, a related study carried out by Professor Paul Watters from Massey University suggests;

New Zealand has a vibrant digital economy, including film, television, music, games, books and authors that depend on respect for intellectual property to generate sustainable revenues (Watters, 2014).

The effect and risk of rogue sites and illegitimate advertising is real and requires tracking for further action and prevention. The research carried out by Professor Paul Watters has always been a manual tracking and site follow-up which limits the tracking due to time constraints and human eye accuracy but this thesis follows the automatic URL tracking.

2.7 Tracking Rogue Sites:

Tracking and finding the pattern of rogue sites has been a matter of focus for various researchers including academics which always been welcomed by litigate enterprises and governmental agencies.

The rogue site complains are normally submitted to Google Transparency Report which holds a database of all other reported sites; this thesis uses these reported links to track and find the source of these links until their final destination is found.

In the past the process of tracking these links has been carried in manual form but this thesis follows a technique where automation processes is carried out through fully customized application. The automation process of tracking rogues sites goes through a specific methodology of data selection which is found in the next chapter (i.e. Research Methodology).

Little work has been done on the process of automation but on the other hand there are enormous studies carried out by Professor Paul Watters and colleagues verifying certain advertisers and higher risk advertisers versus mainstream advertisers it also analyzes their effects on social and economic factors while considering various states based on their geographical locations or politically recognized boundaries.

All data and studies conclude there is a serious need of tracking these rogue sites for creating a safer online contents and communities. Furthermore commercial enterprises are interested in finding these rogue advertisers to minimize their negative effect on their litigate produce and services while governments agencies are series in tracking and locating these rogue advertisers for creating a fare society where rights and sensitivities of all citizens are protected.

2.8 Automatic URL Tracking:

In general terms the word automatic refers to a device or process, working by itself, with little or no direct human input or control. In this research context it is the process of

automatically tracking all reported URLS for the purpose of tracking and identifying the rogue advertisers from the randomly selected sample of DMCA complaints. These samples are taken from the raw CSV data files provided by Google Transparency Report.

The Automatic URL Tracking is normally done through software or an application which is designed to take sample data from a specified source (could be a database or any data file) for the purpose of identifying, following and tracking each and every individual record, after which the required and desired information is extracted and stored in a defined (or preferred) media destination.

In this research specific software has been jointly developed for the purpose of automatically inserting data samples from raw CSV files provided by the Google Transparency Report and then tracking each and every single record from the list of selected sample DMCA complaints.

Chapter Four describes this topic in detail along with all its technical parameters.

2.9 Summary:

Increased internet usage has increased online spending which has directly affected the increase in the online advertising versus traditional cable advertising. This increase in online spending has resulted in high revenues which have encouraged rogue organizations to perform their activities through various means of inappropriate advertisements and this has been verified by Watters through various case studies which he has investigated. Tracking these rogue advertisers remains an active concern, where manual processes might have the most reliable face, this when the automation process becomes useful.

CHAPTER THREE

Research Methodology

3.1 Background:

The key methodology used and followed in this research is; first running the Automatic URL Tracking software to automatically track and identify the randomly selected sample of DMCA reported complaints which are forwarded by the copyright owners and are included in the Google Transparency Report and then manually observing the pattern and advertisers identification where a comparison analysis could be built to see the pattern.

The complainants could be individual copyright owners or any party that claims the right of ownership of the content, whose property copyright is violated. They report (i.e. submit) their complaints to Google Transparency within their defined place to report content that is required to be removed from Google's services, under defined applicable laws, when providing the appropriate and complete information to help investigate the inquiry.

These complaints are held in the Chilling Effect database which is originally shared by Google Transparency in the form of raw CSV data files that could be used to gather and analyze information.

The methodology used in this research operates by downloading the Google Transparency Report, uploading the raw data file (i.e. requests.csv) which is a comma delimited values text file, after it is uploaded (i.e. using insert functionality) to “Auto URL Tracking Software” then any number of random data sets are selected to be finally

exported to a user named database (i.e. schema) where these will be read for further tracking and identifications.

3.2 Data Selection:

The total “50” complaints have been considered for this research, and from each complaints page extracting the first “50” infringement URL. Hence the total number of tested records will be “2500” URLs to follow and extract for further tracking and redirections. The feature of controlling the number links within each request page is controlled by a variable within the program providing the tracking functionality; more details are available Chapter Five of this thesis.

The reason behind the selected number of records and links is the fact that each DMCA complaint or notice could contain thousands of individual links (or URLs). This is why only the first 50 links or URLs can be used as a randomly selected sample as the main concept behind taking a subset of any set of partially identical records. Furthermore, selection of larger data ranges would mean the produced results would have been extremely large, rendering a manual comparison of the data highly questionable.

The application which has been developed to create the results from the raw CSV file provided by the Google Transparency Report uploads the data to a local default database (i.e. schema), allowing the user to select the data range and export the selected range to a user defined (i.e. named) database (i.e. schema) under three tables (i.e. their names are notices, data and redirects) after tracking and identifying each and every single URL within the DMCA complaint page.

3.3 Algorithm:

The main algorithm used to prepare test data for analysis of this research is applied through the following steps:

1. The Auto URL Tracking software for the required database system is installed and readied for desired operation.
2. The latest raw data file from the Google Transparency Report is downloaded; this lists all DMCA complaints requesting page removals for the previous month.
3. The value 50 is assigned to a variable “pageLimit” to restrict the number of links within each DMCA complaint page.
4. The DMCA entire data or a sample of it is inserted in the local default database through the “Insert” function within the Auto URL Tracking application.
5. The entire database is searched and a random sample of 50 DMCA requests is selected, for the purpose of achieving the best and most unbiased rest results.
6. Any random sample was selected and nothing was excluded, to avoid any inaccuracy in the results.
7. The selection is then exported through “Export” functionality to a local table in a user-defined database name, here the maximum number of records can also be defined and constrained.
8. Once the export to a local database is completed then the application gives the option for auto URL tracking; upon the user’s positive responses the program automatically tracks all the requests and desired links within each request page, for this example 50 request pages and 50 links within each page gives the total of 2500 data sets.
9. The main tracking function starts from collecting the information from the “adblock.txt” file to inform of an arraylist which is the key source and each word within this file is used to identify the URL advertiser, if the URL advertiser contains any word from the file or combination of the words, then upon this information the decision is made that the URL is an advertiser.

10. The Auto URL Track will execute repeatedly through the entire selection all records (i.e. DMCA complaints) simultaneously (i.e. one by one) from the "notices" table of the current schema (i.e. database) and investigates every 50 links for each page. This process will continue until all records are reached at the end.
11. As a result of this process, among the three tables within a user defined database (i.e. schema), the first one will be a "notices" table which holds the entire selected records for tracking in sequence; the second one is the "data" table which holds all the links and final advertiser and the third one is the "redirect" table which holds the details of all the redirected links until the final URL is reached.

This thesis draws together data from the Google Transparency Website, where a fully customized (i.e. in-house written) software program uploads the data for various automation processes. It focuses on the method of following the URLs upon their uploading to a local database then follows their appropriate links or redirected links until the final advertiser is detected or found.

This methodology and data source is used instead of any other methods and data samples because this thesis compares the manual process versus automation process where choosing the same sample data is necessary for achieving better results and drawing conclusion toward answering the questions.

The findings as a result of this automated process are used to draw a conclusion where this newly developed automated system could be used, trusted or even funded for future projects and this can be carried forward to build strategies around further development of the automation processes.

3.4Summary:

This chapter demonstrates the data samples taken to carry out this research, where it states that 50 samples of DMCA complaints have been taken and from each of the complaints the first 50 links have been considered for further tracking and identification through the automation process which is carried out by a customized software especially designed and developed to carry out this research.

CHAPTER FOUR

Automatic URL Tracking

4.1 Background:

Automatic URL Tracking is the concept of identifying, following and tracking the sample of reported links (provided by the Google Transparency Report) until all the rogue advertisers are found, where a comprehensive solution could be found for the inability to analyze a solutions which could bring to an end rogue and high-risk advertising.

The basis of this research was initiated by Professor Paul Watters from Massey University where various samples of reported URLs ranking their number of reported URLs were taken for further investigation; a specific program was also written by him for extracting the records to a local database, downloading the pages and then taking their screenshots for further investigation but using a manual method for identifying and analyzing the advertiser.

This research and application has been taken a step forward and an entire and comprehensive application has been developed to insert, display, search, export and track the DMCA reported URLs. At this point it stores the results in a user named local database.

The specific program written by Watters has also be been inherited in which a wide range of modification, enhancement and further development is added to encounter the functionality of tracking every single URL for the purpose of identifying the rogue advertisers from selected sample data for further investigation.

The idea of Auto URL Tracking and identifying the advertisers has emerged and developed after using and following various manual methods in various case studies. The automation process is a replacement for the manual process, not forgetting the fact that it also faces certain challenges that almost every automatic online application would face, but again it must be initiated and worked through until reliable results are achieved. This chapter focuses on the automation process and the specific application written to achieve the expected results.

4.2 Auto URL Track:

The word 'automatic' refers to a device or process, working by itself, with little or no direct human input or control. In the context of this research it refers to the process of software automation where the sample of selected DMCA complaints is considered for tracking and identifying a defined number of URLs (i.e. links) within each and every reported DMCA complaints page which have come from the Google Transparency Report in the form of a raw CSV type data file.

The Automatic URL Tracking and identification of rogue advertisers is done through a fully in-house designed and developed software which is designed to collect specific (or identified) data from a specified source (could be a database or any data file) for the purpose of identifying, following and tracking each individual record.

4.3 Auto URL Track Software:

The Auto URL Track software is totally in-house designed and developed customized software, built using Java and MySQL technologies for the purpose of tracking and identifying advertisers from a selected sample of reported DMCA complaints, which

have originally come from the Google Transparency Report and is hosted by Chilling Effect's database.

This application is developed for the requirement and fulfillment of this research (and thesis paper) where various raw data samples (through their Transparency Report) are used to produce the results for analysis and comparison with the manual method to enable the researcher to answer the main question; can the automation process of URL tracking be made reliable or not?

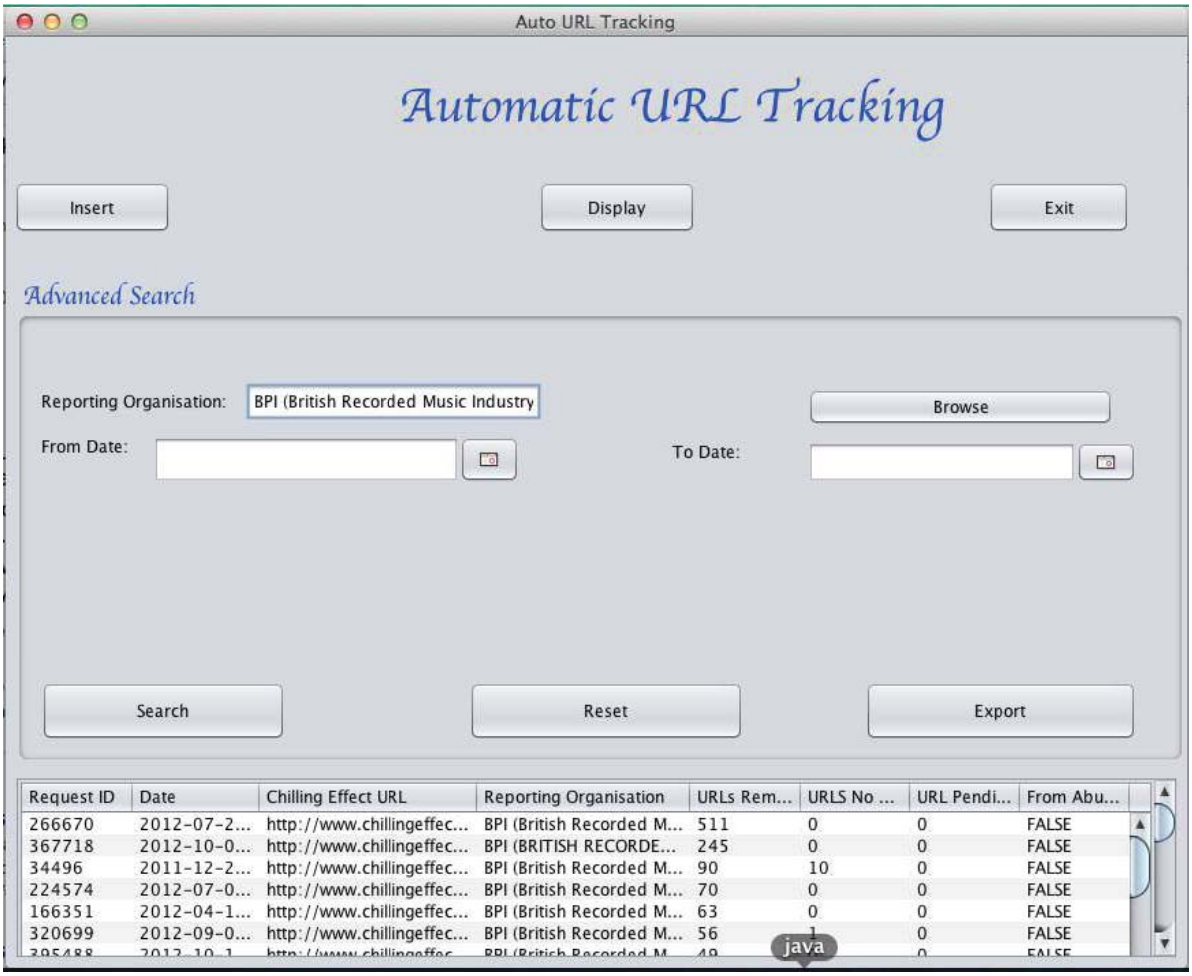


Figure 11: Screenshot of Auto URL Tracking software

This extremely user friendly software allows users to employ GUI to interact with the database values and to retrieve specific records for further tracking until the rogue advertisers are found or located.

The main features of this software, as can be seen from section 4.3.3 , are that it allows users to insert raw data into the local database, displays the data from a local database, provides an advanced search facility for a specific record or group of records relating to any particular record and then exports them to a user defined database name; this is where the user will have the opportunity to track every record for the purpose of identifying rogue advertisers.

The Auto URL Tracking application has been entirely developed for the purpose of this research and thesis. Its technical specifications are as follows:

4.3.1 Technical Specifications:

The following development environment and tools are used for the development of the Auto URL Tracking.

Platform:

This application is fully portable and platform independent. It can be migrated to any platform providing the right development tools are set for its modification. Given that, this application is fully developed in a Mac environment (i.e. operations system), the Mac OS version used is OS X Version 10.9.5 (i.e. latest version of Mac).

Front End:

The frontend is developed using Java Swing programming language in Netbeans IDE Version 8.0 GUI interfaced development tool. The choice of Netbeans Swing IDE is a good choice because of its portability through a massive inbuilt library of various functions.

Back End:

The backend has been developed using MySQL database, and a MySQL Workbenach Version 6.1 is used as GUI interfaced developed tool. The MySAQL, along with its free-of-cost feature, also brings reliability and speed among its list of features where its reliability is seen to its next stage. Furthermore it is highly compatible with the Netbeans IDE development tool.

4.3.2 Data Type and Source:

The raw data used (i.e. consumed) by Auto URL Tracking Software comes from a monthly updated Google Transparency Report, then a sample of that is selected for further tracking and identification of rogue advertisers.

The Google Transparency Report is a monthly issued raw data files containing various statements, complaints and details put forward by various commercial and non-commercial (including governments) organizations. This DMCA complaints report comes in zip format and contains three CSV (Comma Separated Values) text files.Their names are; “requests.csv”, “domains.csv” and “urls-no-action-taken.csv”, but this application only used the “requests.csv” file for inserting the records.

The “requests.csv” file holds all the core and necessary data which is required for any identification of basic DMCA complaints, or in other words, this file is actually the output of a table called “requests” from the Google database which has been exported and made available to the public for further investigation and analysis.

The “domains.csv” file is an export from the “domains” table from the Google’s database and it holds some basic information centered around the domain. This can be used when all other detailed information or fields are not needed.

The “urls-no-action-taken.csv” file is also an export from the “urls-no-action-taken” table from the Google database; this is an even shorter file which carries some basic information and an extra field that carries the information about the URLs which are not considered for removal, or which Google took no action to remove.

Raw Data Structure:

The raw data uploaded in this application is only the "requests.csv" file, which is a comma delimited text file where each data field value is separated with a comma.

Table 2: Data Structure for requests.csv file

Name	Description	Required?
Request ID	An ID unique to each copyright removal request	Yes
Date	The date and time (in UTC)	Yes
Chilling Effects URL	URL to the Chilling Effects page documenting the request	No
Copyright owner ID	The ID number of a unique copyright owner	Yes
Copyright owner name	The name of the owner associated with the request	No
Reporting organization ID	The ID number of a unique reporting organization	Yes
Reporting organization name	The name of the reporting organization associated with the request	No
URLs removed	The number of URLs removed.	Yes
URLs for which we took no action	The number of URLs for which we took no action.	Yes
URLs pending review	The number of URLs that are still pending review.	Yes
From Abuser	If the request was submitted by someone we believe to be abusing the process	Yes

 Appendix D describes full data structure of all tables in CSV file.

Auto URL Tracking Technical Features:

Some basic features of Auto URL Tracking application are as follows:

- This application is fully in-house developed software and is specifically developed for this research;
- This application has an (attractive OR pleasant) and user friendly GUI (i.e. Graphical User Interface) for various types of audiences;
- This application takes raw data from the Google Transparency Report which contains all DMCA complaints;
- This application gives the opportunity to the user to search for a specific record, organization or even within specific dates;
- This application exports the raw data to a user defined database (i.e. schema) and then it enables the user to decide if automatic URL tracking is required or not;
- This application then scans all the selected samples of DMCA complaints and extracts all reported links for further tracking, following and identification of rogue advertisers;
- This application's main functionality within theautomation mechanism (which tracks URLs automatically) is based on adblock.txt file;this file is used for identification of every URL to determine whether it is the high-risk advertiser or not;
- The main algorithm within the automation program is recursive, it will only be executed according to requirements (flags), which means no memory or processing overhead will be a problem in future;
- There is one helper library (JSoup library) automation program which is used in this application.The function of this is to scan the advertiser page and find the final URL link of the advertiser; this provides the same functionality when the user clicks on final advertiser, which takes the user to the final page upon identifying the page,
- The entire code within this program has an advantage for the OOP future programmer to manage future development, analysis and improvement,
- The database implementation in the code is very robust and time saving; all database queries implemented in this program are very short and work without any error in future,

- This application uses the “PUT” query in a very efficient manner for exporting the CSV, importing the raw data to the local database (i.e. schema),
- This program uses specific functions, loops and termination statements in a very efficient manner, which enables it to avoid breakdown situations in any place,
- This program uses Swing components as a user interface, which is the robust Java technology for representing any procedure as a visual theme.
- This application uses ajdbc type-IV driver to establish the secure database connection and processing the transaction done,

Appendix A describes the user’s manual on how to use this application.

4.3.4 Auto URL Tracking Functionalities:

The functionalities, technical specifications and pseudo-codes for each of the programs within Automatic URL Tracking application are described in this section under their appropriate headings.

Appendix D contains all code for each of below programs and functionalities.

4.3.4.1 Main Window:

This window appears as the program executes and it is the first GUI screen which appears upon running the program.

This window (or screen) hosts all buttons (i.e. triggers) for all other functionalities available on the application, through this main window users can insert, display, search and export the records.

Program File: Program file containing the code for this functionality is named as "ONMain.java".

4.3.4.2Insert Function:

This is the process of uploading the raw data (i.e. DMCA complaints) file and then inserting it into the default database. The DMCA complaint file by default is in CSV format and this application has been designed to handle the data and insert it into the specifically designed table within applications recognized (or set) as the default database schema.

Program File: Program file containing the code for this functionality is named as "ONInsert.java".

The process is as followings:

- This process is triggered upon clicking the "Insert" button,
- It will initialize the action event and "btn_insertActionPerformed" method will be executed then,
- It will then initialize the insert box and ask user input for selecting (or browsing) the CSV data file from your local machine or any attached media, this data file is the DMCA complaints, which is originally has come from Google Transparency Report,
- Once user selects the data file then clicks the upload button, the "btn_uploadActionPerformed" of "ONInsert.java" will execute which triggers the followings:
    - All data from the "requests" table will deleted,
    - All data from CSV file will be inserted to the "requests" table.
- The above two ends the Insert (or upload) functionality.
-

Pseudo Code:

```
BEGIN
    Declare file path
    Declare variables
    Try
        Create sql statement
        Truncate the requests table
        Execute the statement
    Catch
    Sql Statement: Load comma separate data to the table
    Give message as "Task performed successfully"
END
```

4.3.4.3 Display Function:

This process is triggered once the "Display" button is pressed and simply it reads through (i.e. runs a MySQL Query) for all records within the "requests" table which were originally inserted from the DMCA complaints CSV file and then displays them on jTable in a defined jForm window.

Program File: Program file containing the code for this functionality is named as "ONDisplay.java".

Pseudo Code:

```
BEGIN
        Connect to database
        Call Update_Table function
        Calls: Update_Table:
        try
                Prepares sql query (each table field attached to column heading)
                Execute the query
                Display table using inbuilt Table Model
        Catch
        Sql Statement: Load comma separate data to the table
        Give message as "Task performed successfully"
END
```

4.3.4.4 Exit Function:

This button will close (i.e. terminate) the entire program from running and takes the user to the development environment.

Program File:This functionality exists within the Main Window (i.e. screen) which is inside the "ONDisplay.java" program file.

4.3.4.5 Search Function:

This feature will allow users to search for a specific record based on the defined values in specific screen fields under the “Advanced Search” heading and upon definition of the desired values the search button should search the record from the “requests” table of a default schema and display it in the screen below.

It is worth mentioning thatusers can enter the reporting organization’s name directly or could search all the report organizations upon clicking the “Browse” button, then from the displayed window a desired organization or organizations could be selected and results will be brought back to the main windows, while saying that users can define specific data of the complaints which could further constrain the search.

Program File: This functionality exists within the Main Window (i.e. screen) which is inside the “ONDisplay.java” program file.

4.3.4.6 Reset Function:

This button will reset all values within the screens (or fields) under the “Advanced Search”.

Program File: This functionality exists within the Main Window (i.e. screen) which is inside the “ONDisplay.java” program file.

4.3.4.7 Export Function:

This functionality allows the core functionality of the Automatic URL Tracking where this button will work and will only function once there are records displayed in

the jTable display screen below, then a new window will appear where the schema name (i.e. database name) and maximum number of records are entered by the user, upon entering these results and pressing the "Export" button then these records will be inserted to a new table called "notices" within a user defined named schema (or database).

Once all the records are inserted tothe "notice" table on the user defined schema then the program automatically confirms the insertion and then it will ask the user whether they want to automatically track all the inserted records to the "notices" table, upon user's confirmation the program will automatically track all the inserted records.

Program File: Program file containing the code for this functionality is named as "ONExport.java".

Pseudo Code:

```
BEGIN

    Declare required variables

    Get user defined schema name

    Get user defined maximum number of records

    Get all organization names, if more than one split them by comma

    If (Organization name is empty) and (To and from dates are empty)

    {

        Try
```

Create sql query

Connect to database

Create notices table within user defined schema name

Insert all records related to this particular organization(s) to notices table within the same user named schema (i.e. database).

Catch

}

Else if (Org name is empty) and (To and from dates are not empty)

{

Try

Create sql query

Connect to database

Create notices table within user defined schema name

Insert all records related to any organization but between these two date ranges to notices table within the same user named schema (i.e. database).

Catch

} Else if (Org name is not empty) and (To and from dates are not empty)

{

```
        Try
                Create sql query
                Connect to database
                Create notices table within user defined schema name
                Insert all records related to this particular organization(s) and
                between these two date ranges to notices table within the same
                user named schema (i.e. database).
        Catch
    }
END
```

4.3.4.8 Auto URL Track Function:

This functionality is the core of the entire program, even though it triggers within the Export Function and called from the "ONExport.java" program but still this functionality is entirely within a new program.

Program File: Program file containing the code for this functionality is named as "ONByURL.java".

History:

This program was originally inherited from Prof. Paul Watters, where the code was extracting the records and their possible links to a local database, but extensive modification, enhancement and further development was added to encounter the functionality of tracking every single URL which has been saved as a record within the notices table.

This program has gone through further development where it stores extra information to “data” table, creates new “redirect” table and stores all re-direction to it, at the end once the rogue advertiser is found then it will be saved along with other information within “data” table.

The program has been developed in Java and ithandles all the core functionality of automatically tracking the URLs and identifying their rogue advertisers, and this program is named as “ONByURL”.Some of its specifications are as followings:

- ONByURL function is used to identify the advertiser from the particular given link within the DMCA complaints sample and then it checks final advertiser;
- The program goes through every link within the DMCA complaints page and stores all re-directions and final advertiser;
- At the end of the program, the user will have all advertisers from the page, and (gets OR obtains) information about advertiser URL redirection.

Program Functionality:

As discussed earlier this program is the heart (i.e. core) of the functionality within this application, the entire functionality of automatic URL tracking is within a program named as “ONByURL” within the Auto URL Tracking application.

The structure of the program is as follows:

- The OnByUrl function starts from collecting all information in the form of arraylist from “adblock.txt” file.

- The “pageLimit” variable can limit or constrain the number of links within any DMCA complaint to be tracked or followed.
- The Adblock.txt file is the key source and each word within this file is used to identify the URL advertiser, if the URL advertiser contains any word from the file or combination of the words, then upon this information the decision is made that the URL is an advertiser.
- The function “moreURLs” will be executed then repeatedly, the moreURLsfunction reads all records (i.e. DMCA complaints) simultaneously (i.e. one by one)from “notices” table of the current schema (i.e. database). The condition for collecting/reading the URL(by moreUrls functionwill be based on the “done” field of the “notices” table, if its value is “0” this is read as false or else when it is “1” it means the operation is on this record and the logic decides the record has been done which will be skipped and next record will be considered for reading or processing.
- The logic within “lookupLineNumber” function will execute and identify the line number/ sequence number of the selected URLfrom the notices table.
- The logic within “rollbackIfNecessary” function will execute and remove the urls from “data” table if the url has been already processed before on the base of sequence number.
- The “readInDMCA” function will be executed and it will process the URL which scans DMCA (list of copyright complains) then this will create the list of websites which have violated the internet rules.
- This list of websites will be scanned simultaneously (i.e. one by one) and validated (valid html URL) by “processURL” function, upon completing the validation it passes the URL to the “processFile”.
- The “processFile” function will execute the each URL from list, scan the whole html page(URL points from list) for advertiser for each URL. Advertiser will be selected on the basis of list adblock.txt.
- If any advertiser is found on this page by “checkForAds” function then advertiser will be checked for redirection through “identifyAdvertiser” function.
- Then “checkForAds_h” function will check the final advertiser URLwhich contains any advertiser through scanning the final advertiser URL html page.

- If "checkForAds_h" function returns positive, then it will check again for the final advertiser, in this process in case of any further redirection detection then this will be stored in "redirects" table.
- The function "Original_url" and "identifyAdvertiser" are helper (or assisting) functions to identify the redirection of advertiser URL, then the "urlComplete" function will be called for updating the rows with "1" (i.e. true) for the "done" variable, to safeguard or make sure a single URL is not processed more than once.
- The MoreUrls function will execute recursively until it finds the (last OR final)URL (DMCA complaint record) from "notices" table to process it simultaneously in a recursive manner.

Features:

- The main purpose of this program is to find the advertiser in a sample ofDMCA complaintsthrough each reported link that isallegedly complained of in a Google Transparency Report.
- It is possible that the advertiser URL on the site is fake (a rogue advertiser) and it redirects to some other website which is not related to advertiser, so the current program will identify final URL (final advertiser) and track it into database.
- Multiple redirection of the advertiser is also going to be tracked by this software by storing it into database.
- The database transaction has been done with secure connection with credential each time. So after once making the binary package, it will bea much more secure database transaction.
- The program uses different available java libraries which you can identify through the Import statements at the top of the file.These libraries are a very efficient part of the java and made up of more robust codes.
- Some of the necessary libraries applied as "javax.swing", "java.awt", "java.sql" and "java.text" are initially part of the java libraries.

- The program is fully object-oriented as java programming language has been used to develop it.
- The prime condition to run this project success fully is a fully working database with required schema structure with MySQL.

Benefits:

- The ONByURL program is very useful when users have some URL which contains different advertisers.
- Ifuser wants to check whether ot not all advertisers are genuine or are just redirecting to another location on the web.
- It is possible that the advertiser will redirect to another advertiser, when using this program user can track this strategy and collect all results.
- This algorithm works very fast and it is recursive;it keeps recurring until it finds final result it will be continuously working.
- Before processing each URL the logic within the algorithm checks to see whether the URL is valid or not, because with the invalid URLs, the current working program will never stop at any particular moment.
- The benefits of uploading theCSV function are that each CSV file will directly uploaded to a given table of the database with the help of a single database query.

Page Limit Controller:

The “pageLimit” variable can limit or constrain the number of links within any DMCA complaint to be tracked or followed; this feature is very important as some DMCA complaints have almost a thousand reported links which would rendering it maybe become an impossible process in case of not setting this limitation or in case this feature lacked.

Pseudo Code:

The Pseudo-code of this program is as followings:

BEGIN

Declare all variable (execRoot, root, pageLimit, dmcaTable) before.

Calls: main function

Calls: loadAdBlock

Input Parameters: adblock.txt

Description: The list of adblock from text file(adblock.txt) supplied as input.

Returns: List of adblock.

Define variable line and initialize it to null.

Initialize line variable through function.

Calls: moreURLs

Description: Get the single URL record each time from notices table if the flag is false.

Returns: The URLscanned from notices table.

---

while the moreURLs function returns value other than "null".

Declare and initialize int variable "num".

Calls: lookupLineNumber

Input Parameters: line

Description: Get the sequence no from notices table of the given URLas input.

Returns: sequence number for given url as input. And assign it to "num".

Calls: rollbackIfNecessary

Input Parameters: "num"

Description: Clear data table entries with "num" as "page" from data table.

Calls: readInDMCA

Input Parameters:: "line","adblock list", "num"

Description: Read the given DMCA URLwebsite, and scan all infringing url. It makes list of infringing url on base of limit variable.

While line has value other than null

Add each infringing url in the list

Read new line with the help of readLine function.

Endwhile

Calls: processURL

Input Parameters: infringing url list, adblock list, num, dmcaurl

Description: Scan the infringing url for redirection and validation of url.

For 0 to number of lines in infringing URL

Scans each line of the page for valid URL.

If: URLis valid then

Calls:processFile

Input Parameters: valid URLand other necessary element.

Description: Check each valid URLlist for advertiser URLwith the help of checkforads function.

Calls: checkForAds

Input Parameters: necessary parameters

Description: Helper function for the processFile. Each URL line will be checked here for any advertiser.

If advertiser found:

Calls:checkForAds_h

Input Parameters: advertiser and adblock list

Description: Checks that final advertiser is again advertiser URLof working URL.

If checkForAds_h returns true

Scan whole advertiser document and find final advertiser URLfrom that page.

Endif

Function identifyAdvertiser will help to identify any redirection of the URLand log it into database.

End if

Endif:

Endfor

Calls: urlComplete

Input Parameters: "line"

Description: Mark the flag variable true after scanning the url from notices table.

Call function moreURLs again and initialize return value to "line" variable.

Endwhile

END

Function Descriptions:

The brief function descriptions used within the ONByURL.java program code are as follows:

- Function: (rollbackIfNecessary)
  - It clears data table entries, if the URL from notices table has already been processed. It creates a new connection through "ONconn.java" and updates the database through the update query in mysql.

- Function: (urlComplete)
  - It marks the flag variable true after scanning the URLfrom notices table. It will make the program more efficient so that no single URLwill be executed and scanned in a single process.

- Function: (moreURLs)
  - Get the single url record each time from notices table if the flag is false. Returns each time new URL entries from notices table which has flag marked as true.

- Function: (lookupLineNumber)
    - Get the sequence no from notices table of the given URL as input; it takes URL string as input.

- Function: (loadAdBlock)
    - Get the list of adblock from text file(adblock.txt) supplied as input, it makes "arraylist" object and returns it to parent calling function for further execution of the program. It uses recursive procedure while looping to extract data from text file and make a list.

- Function: (readInDMCA)
    - Read the given DMCA URL(link from notices table) website, and scan all infringing URLs. It makes list of infringing URLson basis of (limit OR limited) variables.
    - It calls "Original_url" function to check whether the URLis valid URL. It also checks for the redirection URL.
    - It reads the html page line by line and identifies, then it makes the list of infringing URLsand adds it into ArrayList object.
    - This function is called function "processURL" and passes the infringing URLlist for further processing.

- Function: (processURL)
    - Scan each infringing url for redirection and validation of URL
    - It will run in recursive processes while looping for each infringing URLand the function "processFile" will be called for each infringing URL.

- Function: (processFile)
    - Scan the infringing html page through each line, and check whether or not the line contains the advertiser url according to adblock.txt file.

- Function: (insertIntoDB)
    - Insert the scanned advertiser with all details in data table of the database.

- Function: (extractHostname)
    - Extract Hostname from given URLand return the hostname to the caller function.

- Function: (checkForAds)
    - Check for validation and redirection of advertiser. If final advertiser URLis again advertiser according to adblock list then scan the final advertiser html page and get the real advertiser URL.
    - This function uses adblock list to identify the advertiser URLfrom the given html URL.
    - After identifying the advertiser url, the “identifyAdvertiser” functions to check the redirection or final advertiser page.
    - Then the final advertiser again will be checkedfor whether the URLis normal URLor advertiser URLthrough “checkForAds_h” function.
    - If “checkForAds_h” returns positive, then “checkForAds” will find valid final URLthrough scanning the html page for valid html links.
    - This validation will be done on the basis of the iFrame element or anchor tag element inside html. If it finds any valid URL, then it will be considered as the final advertiser valid URLand logged inside database table.

- Function: (identifyAdvertiser)
    - This function will call through the “checkForAds” function. It will take the advertiser URLas the input parameter and scan URLfor redirection check. If URLis redirected to another URL, then it returns the final URLotherwise returning the same URL.
    - After tracking the redirection, it will log that redirection in redirect table in database.
    - Function tracks the redirection through accessing the “HttpURLConnection” Object and “Location” attribute of this object.

- Function: (Original_url)
    - This function willcall in the “readInDMCA” function to check whether the allegedly URLredirects to anywhere else or not.
    - It will not log the redirection in the database, because it is not necessary. It will also use “HttpURLConnection” object of “java.net” package to ascertain the redirection path.

- Function: (checkForAds_h)
    - The requirement for our application is that the final advertiser should be a normal working URL, rather than an ad URLagain. This function checks that the final advertiser URLis valid URLby returning the flag to its parent calling function.
    - Again this function uses the adblock list for filtering the advertiser URL in either the normal URLor ad-URL.
    - It uses for looping a recursive iterator which executes on the required condition to be true at a particular time.

- Function: (exists)
    - Check whether the given URL exists on the web or not. It will return an appropriate result to the parent function.
    - Again it uses “HttpURLConnection” object from “java.net” package for identification purpose.

Algorithm:

1. Set the database name in variable DMCA,
2. Call the main method,
3. Take adblock.txt as input and make list of all words,
4. Take single URLfrom “notices” table,
5. Get the line number (seq) from “notices” table as input,
6. Delete the records from “data” table if it contains same sequence number,

7. If URLis valid, then scan whole html page and find allegedly infringing URL, and make list of bad URLs,
8. Process infringing URL (html page) line by line, and find the advertiser URLbased on adblock list,
9. Advertiser URL will be check for redirection loop. At the end it will find final URL,
10. If final URLis not advertiser then store it into database,
11. If final URLis again advertiser, then scan the final advertiser page and get the URLlink for the real advertiser,
12. Then store the real advertiser to the database as final advertiser,
13. Repeat 8-12 for each infringing URLwith given limit in the program.
14. Repeat 4-12 for each single URLfrom “notices” table,
15. Update the flag(done) of each URLfrom “notices” table after processing it,
16. During the whole process if program finds any redirection then store in database redirect table.

4.4 Summary:

The Automatic URL Tracking software was initiated by Professor Paul Waters and then has been developed further to encounter the functionality of automatic URL tracking; since then it has been integrated within a user friendly in-house built GUI application which provides other functions such as insert, display and export data which are the selected samples of DMCA complains reported to Google Transparency and distributed as a report.

CHAPTER FIVE

Experiments and Test Results

The total of “50” complaints have been considered for this research where the first “50” URL links are extracted for each complaint page, which makes the entire total of “2500” URLs to follow and track.

The reason behind the selected number of records and links is the fact that each DMCA complaint or request for removal could contain over a thousand individual URLs, for this reason only first 50 URLs are selected as an identifying sample,which are the representative subset of all other URLs. Furthermore in the case of selecting larger data ranges the produced results would have been extremely large where the accuracy and comparison of the manual testing would have been a questionable.

The automatic URL tracking results are shown alongside the manual tracking agreements, the software test results and manualagreements are demonstrated inthe table below, meaning the program has run and produced these outputs which have been compared with the manual checks.

Appendix G details the entire output produced by Automatic URL Tracking Software

To draw a comparison analysis, there has been assumption behind the manual tracking where a human judgments are considered to be always correct based on this consideration the table below draws a distinct comparison with the automation process.

Table 3: Test results produced by automation Vs Manual Tracking

URL	Complainant	Auto Track	Manual Track	Seq
http://www.chillingeffects.org/notice.cgi?sID=869990	Froytal Services Ltd	0	3	1
http://www.chillingeffects.org/notice.cgi?sID=1210329	Froytal Services Ltd	0	1	2
http://www.chillingeffects.org/notice.cgi?sID=1306104	Froytal Services Ltd	0	2	3
http://www.chillingeffects.org/notice.cgi?sID=822711	Paramount	1	1	4
http://www.chillingeffects.org/notice.cgi?sID=790950	Froytal Services Ltd	7	7	5
http://www.chillingeffects.org/notice.cgi?sID=818739	Froytal Services Ltd	50	50	6
http://www.chillingeffects.org/notice.cgi?sID=918623	Top Media Distribution Ltd.	0	2	7
http://www.chillingeffects.org/notice.cgi?sID=633579	Froytal Services Ltd	0	3	8
http://www.chillingeffects.org/notice.cgi?sID=612164	Froytal Services Ltd	24	24	9
http://www.chillingeffects.org/notice.cgi?sID=105485	Froytal Services Ltd	0	2	10
http://www.chillingeffects.org/notice.cgi?sID=1258622	Top Media Distribution Ltd.	0	5	11
http://www.chillingeffects.org/notice.cgi?sID=555039	Froytal Services Ltd	0	0	12
http://www.chillingeffects.org/notice.cgi?sID=583736	Froytal Services Ltd	0	0	13
http://www.chillingeffects.org/notice.cgi?sID=1777888	Sony Music Entertainment	0	0	14
http://www.chillingeffects.org/notice.cgi?sID=1463110	Sony Music Entertainment	2	2	15
http://www.chillingeffects.org/notice.cgi?sID=519469	Froytal Services Ltd	0	0	16
http://www.chillingeffects.org/notice.cgi?sID=1732880	Paramount	0	0	17
http://www.chillingeffects.org/notice.cgi?sID=572179	Froytal Services Ltd	0	1	18
http://www.chillingeffects.org/notice.cgi?sID=525390	Froytal Services Ltd	0	3	19
http://www.chillingeffects.org/notice.cgi?sID=1790628	Musical Freedom	0	0	20
http://www.chillingeffects.org/notice.cgi?sID=621368	Froytal Services Ltd	0	0	21
http://www.chillingeffects.org/notice.cgi?sID=617838	Froytal Services Ltd	0	0	22
http://www.chillingeffects.org/notice.cgi?sID=555410	Sony Music Entertainment	53	50	23
http://www.chillingeffects.org/notice.cgi?sID=488209	Sony Music Entertainment	0	0	24
http://www.chillingeffects.org/notice.cgi?sID=1902970	Musical Freedom	8	6	25
http://www.chillingeffects.org/notice.cgi?sID=1166042	Froytal Services Ltd	0	0	26
http://www.chillingeffects.org/notice.cgi?sID=124937	Froytal Services Ltd	0	0	27
http://www.chillingeffects.org/notice.cgi?sID=159114	Sony Music Entertainment	0	0	28
http://www.chillingeffects.org/notice.cgi?sID=538630	Froytal Services Ltd	0	2	29
http://www.chillingeffects.org/notice.cgi?sID=469650	Sony Music Entertainment	0	0	30
http://www.chillingeffects.org/notice.cgi?sID=254193	Bubblegum Films INC	0	4	31
http://www.chillingeffects.org/notice.cgi?sID=448587	Froytal Services Ltd	0	0	32
http://www.chillingeffects.org/notice.cgi?sID=524736	Froytal Services Ltd	0	0	33
http://www.chillingeffects.org/notice.cgi?sID=223810	Froytal Services Ltd	0	0	34
http://www.chillingeffects.org/notice.cgi?sID=506312	Sony Music Entertainment	0	0	35
http://www.chillingeffects.org/notice.cgi?sID=215915	Froytal Services Ltd	0	0	36

http://www.chillingeffects.org/notice.cgi?sID=1453800	Froytal Services Ltd	0	0	37
http://www.chillingeffects.org/notice.cgi?sID=685478	Bubblegum Films INC	2	2	38
http://www.chillingeffects.org/notice.cgi?sID=829667	Bubblegum Films INC	0	0	39
http://www.chillingeffects.org/notice.cgi?sID=1896262	Musical Freedom	6	4	40
http://www.chillingeffects.org/notice.cgi?sID=183180	Froytal Services Ltd	0	0	41
http://www.chillingeffects.org/notice.cgi?sID=507429	Sony Music Entertainment	1	2	42
http://www.chillingeffects.org/notice.cgi?sID=1912638	Black Hole	0	2	43
http://www.chillingeffects.org/notice.cgi?sID=487190	Sony Music Entertainment	0	0	44
http://www.chillingeffects.org/notice.cgi?sID=555921	Sony Music Entertainment	0	0	45
http://www.chillingeffects.org/notice.cgi?sID=1746821	Musical Freedom	0	0	46
http://www.chillingeffects.org/notice.cgi?sID=399809	Froytal Services Ltd	0	0	47
http://www.chillingeffects.org/notice.cgi?sID=177477	Sony Music Entertainment	0	0	48
http://www.chillingeffects.org/notice.cgi?sID=1304605	Froytal Services Ltd	0	0	49
http://www.chillingeffects.org/notice.cgi?sID=1733540	Paramount	4	2	50
Totals		**158**	**180**	

The agreement percentage between the automation process and manual check is 87.7%

The agreement percentage between the software automation and manual has not reached 100% and it may not be possible at stage of time due to various techniques used by the advertiser hosts to overcome and block the automation process for various reasons.

There are certain patterns have been recognized during the manual detections and tracking as various factors influence the precision results where the most common one is the website's techniques to block the automation process through adding interactive pages, in this process browsers are asked to enter a randomly generated number or their agreement to terms and conditions and even entering an "Enter" button to enter the site, which was the most common one in the identified samples

Summary:

The total 2500 URLs have been taken for the sampling, which are the first 50 URLs from the randomly selected DMCA complaints. After running the application the results when compared to the manual analysis are lower than full accuracy, but again not very bad considering the obstacles in the way of software automation security concerns.

CHAPTER Six

Conclusion and Discussion

The main goal behind this thesis was to test the automatic URL tracking and showing its results alongside its comparison with the manual process.

The two main questions behind the entire research direction and scope are answered and discussed as:

- To what extent are automation and auto classification possible through demonstrating an empirical means of accruing an automatic system for recognizing and confirming larger data sets?
- Why is an automated system important for building a strategy to increase accuracy and avoid human dependency?

It is important to be noted here that the application was run with the sample of 50 DMCA complains in which first 50 URLs are extracted from each page. The results were produced and added to an excel sheet, thus each one of these URLs was taken and manually checked for results to see the similarity of the pattern or accuracy of the software automation.

When the URLs within DMCA complaints were searched or entered in the browsers, there were many broken links. This indicates the links have been removed, rendering their tracking not possible, but again it has not affected the overall result as even the application did not see it as an active link, while manual comparison would indicate the same.

Answer the first question then, the automation process (i.e. software) has not produced 100% agreement with the manual process, while in saying that these ideal agreements may not be easily achieved, due to the availability of multiple online technologies and the cleverness of the rogue advertiser hosting agents. As their skills increase every day they always find new ways to overcome the software automation process through various skillful techniques. Furthermore these techniques of blocking the software automation or simply of page access area also practiced by the mainstream advertisers due to security breaches and site hackers. This practice could also be categorized as a form of security precaution by the mainstream advertisers versus the smart technique of blocking the automated access to the site.

The results achieved in the previous chapter (i.e. Experiments and Test Results) are the best possible results which is indicates scope or to what extend an automatic application can track any particular URL from the reported database provided by Google Transparency Report.

There are various influencing factors, some are discussed in below paragraphs, after following and tracking every URL and then comparing it to the software automation results, the following influencing factors are seen which are limiting the empirical outcomes.

The biggest and most influential factor influencing the results where the automation processes seem to fail is the interactive entry on visited page once the advertiser is identified. Normally once an advertising banner is clicked it takes you to a new page and asks you if you are over 18 to click the enter button, type random numbers or simply if you want to continue click the enter button. None of these manual interactive hosting sites are detected by the automation software.

Another important area which has affected the low results is shown in pages where the content was in a different language. The program seemed to get confused picking the advertisers automatically.

The automation software does not seem to respond very well on the advertisements on the pop-up pages as it fails to load the page in its memory or to save its catch.

The software automation program also fails on the links where the rogue advertiser, through the auto download process, automatically downloads a file to the local machine upon loading the page; this auto download is not captured through this automation program.

All these results show software automation is possible and its further development comes with a challenge. It is not impossible given further development and enhancement of coding and technique. Furthermore the software automation cannot be fully relied on but with the addition of further enhanced techniques of overcoming the manual entry pages, various language sites and other faced issues along with the introduction of artificial intelligence to the code would yield better results.

Answering the second question of the thesis, all the above influencing factors which are limiting the accuracy of results emphasis the need for building a strategy to increase accuracy in the automation process through developing further techniques to overcome the difficulty of advertiser limiting the automation processes. Furthermore a real strategy is required to overcome the limiting factors on the automation process because this indicates the advertisers are trying every single technique to bypass the automation through introducing all available technologies but this has to be opposed through introducing new techniques and technologies within the automation processes, these are discussed in the next chapter (i.e. Future Work).

Summary:

The main factors behind differences in results to draw agreements are the manual page entries, foreign languages, popup pages and downloaded files, and the final conclusion behind software automation is that while it is not impossible the achieving of better results depends on further development and enhancement of the application.

CHAPTER SEVEN

Future Work

The automation process can be further developed through introducing the new techniques and technologies to bypass the all the limitations and features added by the advertisers.

Introducing an artificial intelligence could be introduced in the application which uses neuron learning to learn the pattern as the application runs could produce much better results. The smart features like functionality that could bypass the interactive pages could also be added that could bypass the manual entry, language difficulties and other issues, if these smart features could find the pattern from page contents the bypassing these interactive pages could be achieved.

A reporting function could be added to the application where it would report on every single URL at which it faced redirection difficulty, and could therefore analyze time or page content variables. This kind of reporting would help developers to understand the facts behind various difficulties the application is facing when tracking every single URL.

The program could be made more robust and user-friendly where various users are allowed to use the program to achieve their feedback comments, this could add to extension and enhancement of the programs functionality.

The “pageLimit” variable controls the number of links to be read within each DMCA complaint. While currently this variable is set through the program code and has no GUI screen entry, functionality of changing this variable could be made available to users through the screen.

References

- Allen, B. (2014). Booz Allen Hamilton. Retrieved 2015, from Booz Allen: http://www.boozallen.com/insights/2012/03/cyber-threats-innovative-cybersecurity

- Cavelty, M. D. (2007). Cyber-security and threat politics: US efforts to secure the information age. Routledge.

- Chen, Y. N., Chen, H. M., Huang, W., & Ching, R. K. (2006). E-government strategies in developed and developing countries: An implementation framework and case study. Journal of Global Information Management (JGIM),14(1), 23-46.

- Cheng, A., Bansal, N., & Koudas, N. (2013, June). Peckalytics: Analyzing experts and interests on Twitter. In Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data (pp. 973-976). ACM.

- Chin, R. C., & Ham, R. E. (2015). U.S. Patent No. 8,927,260. Washington, DC: U.S. Patent and Trademark Office.

- Choi, L. V. (2005). Cybersecurity and homeland security . Nova Science Publishers, Inc.

- Crosse, M., US Government Accountability Office, & United States of America. (2014). Internet Pharmacies: Most Rogue Sites Operate from Abroad, and Many Sell Counterfeit Drugs.

- Cullen, R. (2001). Addressing the digital divide. Online information review,25(5), 311-320.

- De Mooij, M. (2013). Global marketing and advertising: Understanding cultural paradoxes. Sage Publications.
- Dreze, X., & Zufryden, F. (1998). Is Internet advertising ready for prime time?.Journal of Advertising Research, 38, 7-18.

- Felici, M. (2013). Cyber security and privacy : trust in the digital world and cyber security and privacy EU Forum 2013, Brussels, Belgium, April 2013, Revised selected papers . Heidelberg : Springer.

- Felt, A. P., Finifter, M., Chin, E., Hanna, S., & Wagner, D. (2011, October). A survey of mobile malware in the wild. In Proceedings of the 1st ACM workshop on Security and privacy in smartphones and mobile devices (pp. 3-14). ACM.

- Funk, T. (2013). Facebook Advertising. In Advanced Social Media Marketing(pp. 75-101). Apress.

- Gao, S., & Zang, Z. (2014). An empirical examination of users' adoption of mobile advertising in China. Information Development, 0266666914550113.

- Gehrke, J., Lui, E., & Pass, R. (2011). Towards privacy for social networks: A zero-knowledge based definition of privacy. In Theory of Cryptography (pp. 432-449). Springer Berlin Heidelberg.

- General, A. (2013). The UK cyber security strategy: Landscape review.

- GENERAL, A. (2013). The UK cyber security strategy: Landscape review.

- Gerrard, P., Barton Cunningham, J., & Devlin, J. F. (2006). Why consumers are not using internet banking: a qualitative study. Journal of Services Marketing,20(3), 160-168.

- Goldfarb, A. (2014). What is different about online advertising?. Review of Industrial Organization, 44(2), 115-129.

- Groenfedlt, T. (2014). Cybersecurity Threats Are Rising -- EY. Retrieved 2015, from Forbes: http://www.forbes.com/sites/tomgroenfeldt/2013/11/11/cybersecurity-threats-are-rising-ey/

- Homan, M. (2010). Promoting community change: Making it happen in the real world. Cengage Learning.

- IAB, P. (2012). IAB Internet advertising revenue report 2011 full-year results. Market research report, Interactive Advertising Bureau (IAB) and PricewaterhouseCoopers (PwC). http://www. iab. net/media/file/IAB_Internet_Advertising _Revenue_Report_FY_2011. pdf.

- Iles, J. (2014). Cyber Security in the UK: A critical analysis of the threat to the UK and the Government Response (National Cyber Security Strategy 2011).

- James Graham, R. H. (2011). Cyber security essentials . Boca Raton: Auerbach Publications.

- Janczewski, L. (Ed.). (2007). Cyber warfare and cyber terrorism. IGI Global.

- Kamen, Y., & Shirman, L. (2014). U.S. Patent No. 8,683,512. Washington, DC: U.S. Patent and Trademark Office.

- Katin-Borland, N. (2012). Cyberwar: A real and growing threat. Cyberspaces and Global Affairs, 1.

- Khraim, H. S., & Alkrableih, A. A. (2015). The Effect of Using Pay Per Click Advertisement on Online Advertisement Effectiveness and Attracting Customers in E-marketing Companies in Jordan. International Journal of Marketing Studies, 7(1), p180.

- Mairs, B. (2014, 11 4). Thomson Reuters Survey Reveals Increased Cybersecurity Risk to Boardroom Communications. Retrieved 1 3, 2015, from Thomson Reuters : http://thomsonreuters.com/press-releases/112014/increased-cybersecurity-risk-to-boardroom-communications

- McCoy, S., Everard, A., Polak, P., & Galletta, D. F. (2007). The effects of online advertising. Communications of the ACM, 50(3), 84-88.

- Mo, Y., Kim, T. H., Brancik, K., Dickinson, D., Lee, H., Perrig, A., & Sinopoli, B. (2012). Cyber–physical security of a smart grid infrastructure. Proceedings of the IEEE, 100(1), 195-209.

- Moore, T., Clayton, R., & Anderson, R. (2009). The economics of online crime.The Journal of Economic Perspectives, 23(3), 3-20.

- Nimmer, D. (2013). Nimmer on copyright. LexisNexis.

- North, K., & Kumta, G. (2014). Measuring and Safeguarding Intellectual Capital. In Knowledge Management (pp. 251-277). Springer International Publishing.

- O'Dwyer, S. T., & Moyle, W. (2014). Using Google Adwords to recruit family carers of people with dementia. Australasian journal on ageing, 33(2), 128-131.

- O'Neill, B. (2014). Report of the Internet Content Governance Advisory Group.

- Online Ads: A Guide to Online Ad Types and Formats. (2014). Retrieved 2015, from Wordstream: http://www.wordstream.com/online-ads

- Paris, R. (2001). Human security: Paradigm shift or hot air?. International security, 26(2), 87-102.

- Paul A. Watters, M. F. (2014). Maximising Eyeballs but Facilitating Cybercrime? Ethical Challenges forOnline Advertising in New Zealand. (pp. 1-6). Auckland: IEEE.
- Paul A. Watters, M. W. (2014). Malicious Advertising and Music Piracy: A New Zealand Case Study. The Cybercrime and Trustworthy Computing Conference (pp. 1-6). Auckland: Social Science Research Network.

- Poulter, S. (2012). eBook pirates cash in on Kindle sales boom as thousands turn to rogue sites for cheap downloads. Daily Mail. http://www. dailymail. co. uk/sciencetech/article-2081072/Onlinepirates-threaten-Kindle-profits-thousands-turn-sites-download-freeeBooks. html# ixzz1iIg2cDGE (20.03. 2014).

- PricewaterhouseCoopers, I. A. B. (2013). Iab internet advertising revenue report, 2012 full year results.

- Ralston, P. A. S., Graham, J. H., & Hieb, J. L. (2007). Cyber security risk assessment for SCADA and DCS networks. ISA transactions, 46(4), 583-594.

- Ramzan, M. J. (2008). Crimeware : understanding new attacks and defenses. Addison-Wesley.

- Robinson, H., Wysocka, A., & Hand, C. (2007). Internet advertising effectiveness: the effect of design on click-through rates for banner ads.International Journal of Advertising, 26(4), 527-541.

- Samuelson, P. (2000). Privacy as intellectual property?. Stanford Law Review, 1125-1173.

- Schaeffler, J. (2014). Digital signage: software, networks, advertising, and displays: a primer for understanding the business. CRC Press.

- Shah, N. J., Banga, J. S., Sampat, M., & Patel, A. (2015). U.S. Patent No. 20,150,032,550. Washington, DC: U.S. Patent and Trademark Office.

- Sharma, D., Chowhan, D., Singh, S., Gupta, D., & Srivastava, M. (2015). Consumer Perception on Online-Business: A Marketing Strategy for New Entrepreneur. Sudhinder Singh and Gupta, Dr. Devesh and Srivastava, Mr. Vishal, Consumer Perception on Online-Business: A Marketing Strategy for New Entrepreneur (January 15, 2015).

- Sinclair, J. (2015). Advertising, the Media, and Globalization. Media Industries,1(3).

- Steel, E. (2008). Google search ads rile its big customers. The Wall Street Journal, B1.

- Symantic. (2015). symantic.com. Retrieved 2015, from Symantic: http://www.symantec.com/en/au/security-

intelligence/?&om_sem_cid=biz_sem_s138855558423834%7cpcrid%7c65436602128%7cpmt%7cb%7cplc%7c%7cpdv%7cc

- Tahboub, R., & Saleh, Y. (2014, January). Data Leakage/Loss Prevention Systems (DLP). In Computer Applications and Information Systems (WCCAIS), 2014 World Congress on (pp. 1-6). IEEE.

- Tahboub, R., & Saleh, Y. (2014, January). Data Leakage/Loss Prevention Systems (DLP). In Computer Applications and Information Systems (WCCAIS), 2014 World Congress on (pp. 1-6). IEEE.

- Thuraisingham, E. F. (2006). Web and Information Security. Pennsylvania: IGI Global.

- Tran, M. D., Acs, G., & Castelluccia, C. (2014). Retargeting Without Tracking.arXiv preprint arXiv:1404.4533.

- Usage and Population Stats. (2014). Retrieved 2015, from Internet World Stats: http://www.internetworldstats.com/stats.htm

- Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. computers & security, 38, 97-102.
- Watters, P. (2012). Cyber Security: Concepts and Cases. Government of Victoria: British Scientific Publishing.

- Watters, P. (2014). Do Rogue Websites and High Risk Advertising Undermine Social Policy in Vietnam?. Available at SSRN 2505552.

- Watters, P. (2014). Mainstream Advertising on Rogue Websites in Malaysia: A Comparison of Local and Foreign Content. Available at SSRN 2469606.

- Watters, P. (2014). Mainstream Advertising Support for Online Piracy in Taiwan. Available at SSRN 2405281.

- Watters, P. A. (2014). A Systematic Approach to Measuring Advertising Transparency Online: An Australian Case Study . Proceedings of the Second Australasian Web Conference (AWC 2014) (pp. 1-4). Auckland: IEEE.

- Watters, P. A. (2014). Sweet As? Advertising on Rogue Websites in New Zealand. (pp. 1-6). Auckland: Social Science Research Network.

- Watters, P. A. (2014, January). A systematic approach to measuring advertising transparency online: An Australian case study. In Proceedings of the Second Australasian Web Conference-Volume 155 (pp. 59-67). Australian Computer Society, Inc..

- Watters, P. A. (2014, January). A systematic approach to measuring advertising transparency online: An Australian case study. In Proceedings of the Second Australasian Web Conference-Volume 155 (pp. 59-67). Australian Computer Society, Inc..

- Williams, G. B. (2007). Online business security systems [electronic resource.] New York: Springer.

- Womack, B. (2010). Facebook advertisers boost spending 10-fold, COO says.Bloomberg, August, 3.

- Yu, P. K. (2015). Digital Copyright Enforcement Measures and Their Human Rights Threats. RESEARCH HANDBOOK ON HUMAN RIGHTS AND INTELLECTUAL PROPERTY, Christophe Geiger, ed., Edward Elgar Publishing.

- Yuan, S., Abidin, A. Z., Sloan, M., & Wang, J. (2012). Internet Advertising: An Interplay among Advertisers, Online Publishers, Ad Exchanges and Web Users. arXiv preprint arXiv:1206.1754.

- Zeng, F., Huang, L., & Dou, W. (2009). Social factors in user perceptions and responses to advertising in online social networking communities. Journal of Interactive Advertising, 10(1), 1-13.

- Watters, P., Watters, M. & Ziegler, J. (2015). Maximising Eyeballs but Facilitating Cybercrime? Ethical Challenges for Online Advertising in New Zealand, 48th Hawaii International Conference on System Sciences

- eMarkter. (n.d.). 2 Billion Consumers Worldwide to Get Smart(phones) by 2016. Retrieved November 27, 2015, from eMarketer: http://www.emarketer.com/Article/2-Billion-Consumers-Worldwide-Smartphones-by-2016/1011694

- Inc, S. (n.d.). The Statiscs Portal. Retrieved November 27, 2015, from Statista: Mobile phone internet user penetration worldwide from 2014 to 2019

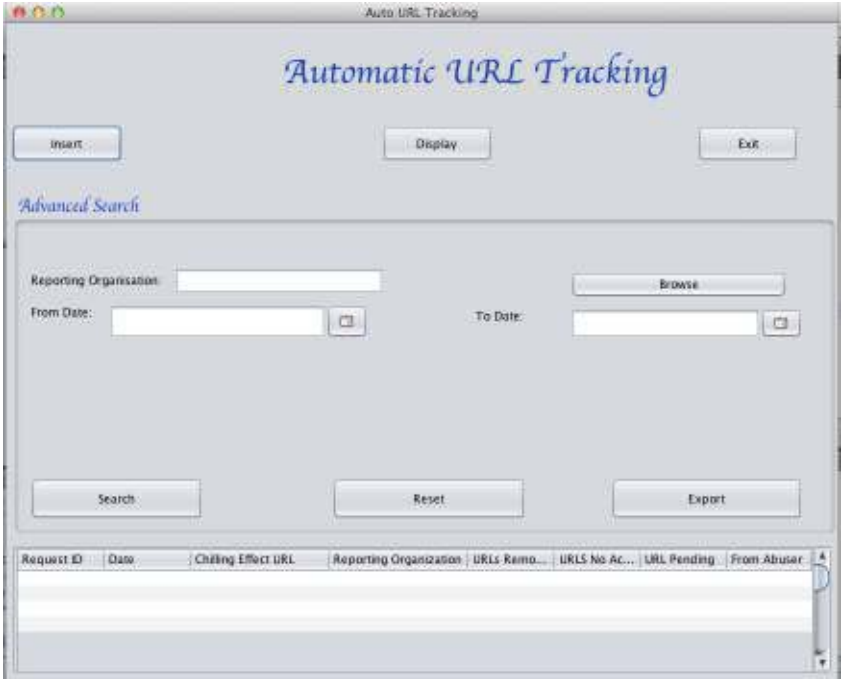
Appendix A
Automatic URL Tracking Software

Below screenshots describe how Automatic URL Tracking software could (or should) be used.

A.1 Loading Application

To load the development environment and run the program, double click both “Netbeans IDE” and “MySQL Workbench” launching shortcuts.

Then the main program should look like this:

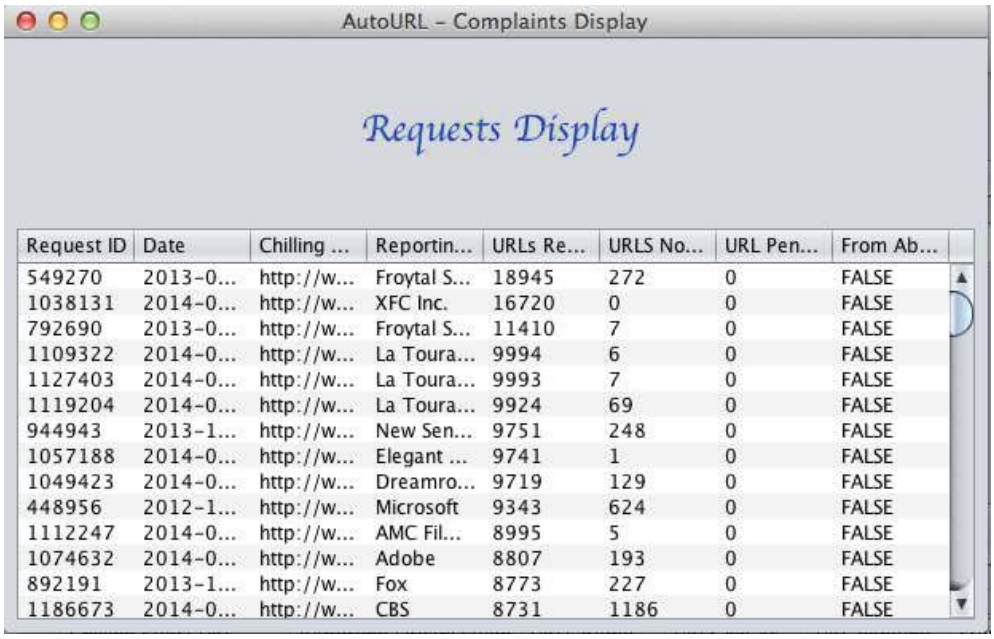


A.2 Insert Function

Click “Insert” button from the main window, either browse for thefile in the any of the attached media sources or type the file path and name then click “Upload” button to upload the raw DMCA complaints data file to a default database.

A.3 Display Function

Click “Display” button from the main window, then the program simply reads through all records within the “requests” table for displaying the in window.



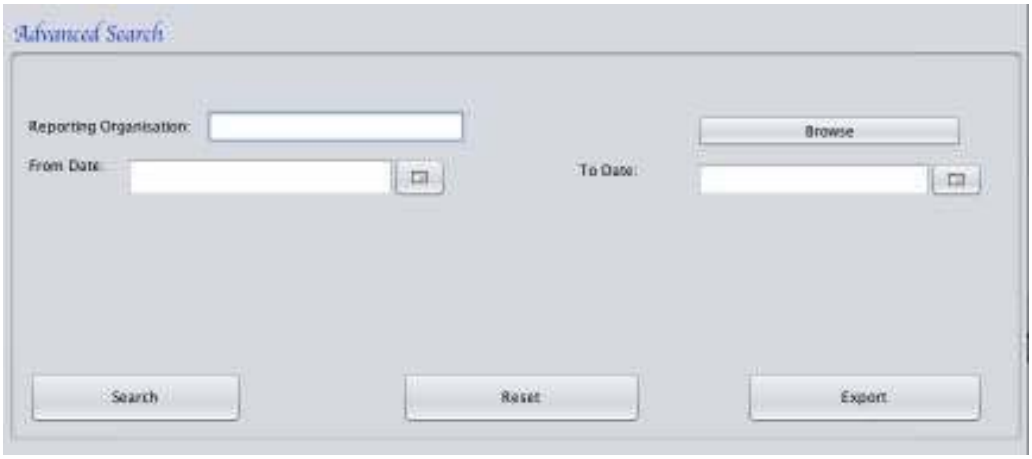
Request ID	Date	Chilling ...	Reportin...	URLs Re...	URLS No...	URL Pen...	From Ab...
549270	2013-0...	http://w...	Froytal S...	18945	272	0	FALSE
1038131	2014-0...	http://w...	XFC Inc.	16720	0	0	FALSE
792690	2013-0...	http://w...	Froytal S...	11410	7	0	FALSE
1109322	2014-0...	http://w...	La Toura...	9994	6	0	FALSE
1127403	2014-0...	http://w...	La Toura...	9993	7	0	FALSE
1119204	2014-0...	http://w...	La Toura...	9924	69	0	FALSE
944943	2013-1...	http://w...	New Sen...	9751	248	0	FALSE
1057188	2014-0...	http://w...	Elegant ...	9741	1	0	FALSE
1049423	2014-0...	http://w...	Dreamro...	9719	129	0	FALSE
448956	2012-1...	http://w...	Microsoft	9343	624	0	FALSE
1112247	2014-0...	http://w...	AMC Fil...	8995	5	0	FALSE
1074632	2014-0...	http://w...	Adobe	8807	193	0	FALSE
892191	2013-1...	http://w...	Fox	8773	227	0	FALSE
1186673	2014-0...	http://w...	CBS	8731	1186	0	FALSE

A.4Exit Function

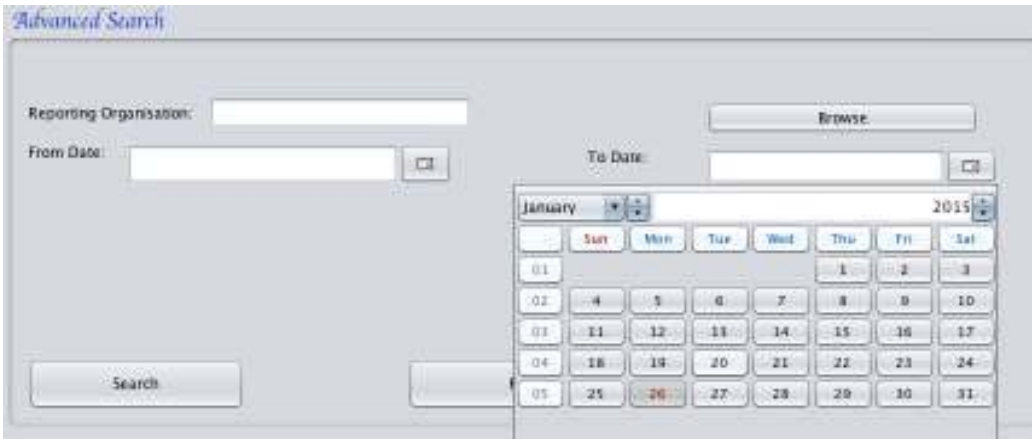
Click “Exit” button from the main window. This button will close (i.e. terminate) the entire program from running and take the user back to the development environment.

A.5 Advanced Search Function:

Click “Search” button from the main window to search for the specific records based on defined values in the specific screen fields under the “Advanced Search”. This allows users to search for a specific organization, date ranges and multiple organizations.

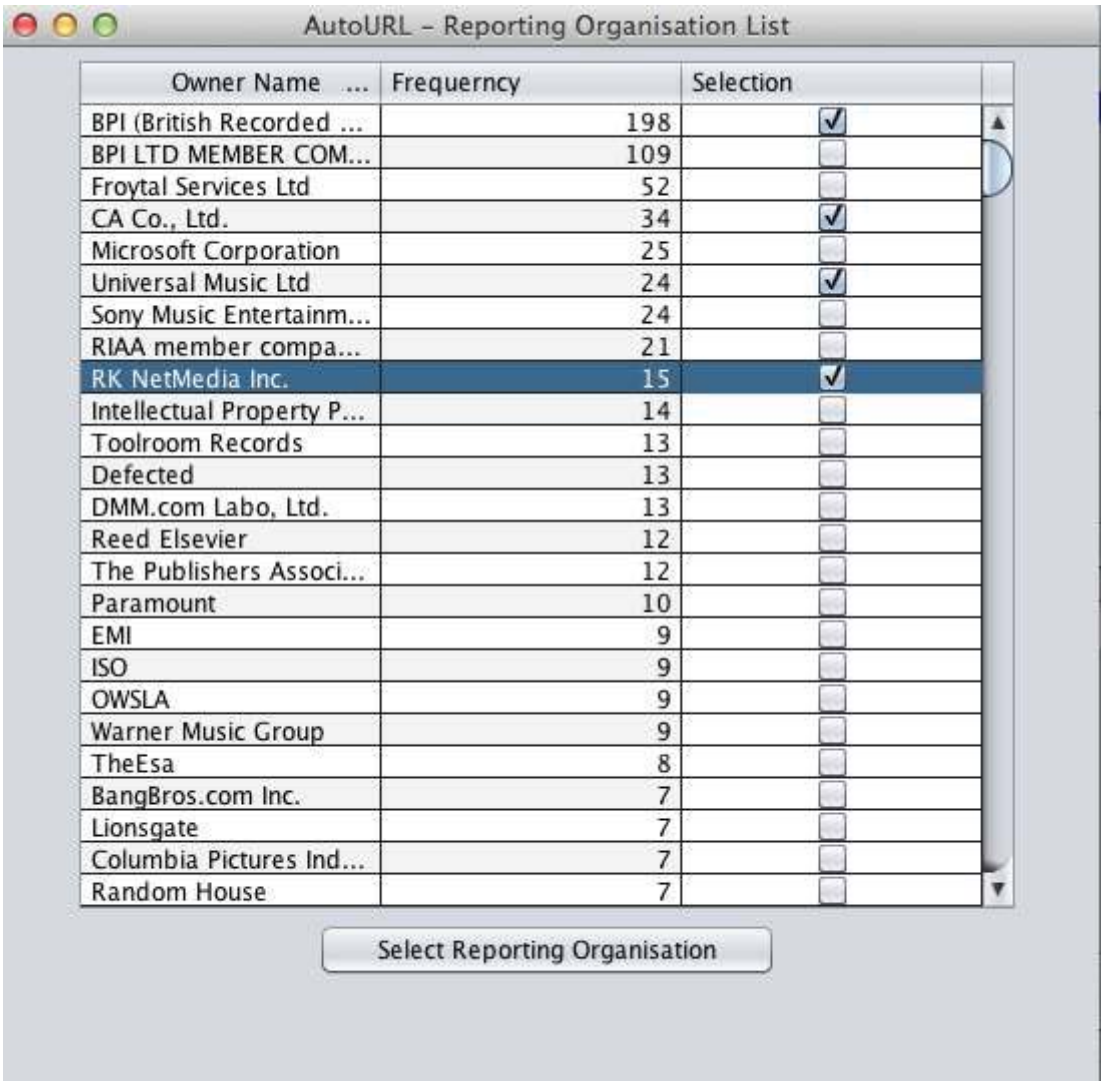


Date ranges can also be specified through a user friendly date picker, as shown in below screenshot.



A.6 Selecting Multiple Organizations

Click “Browse” button from the main window to search for all reporting organizations in order to number of reporting. This allows selecting as many organizations as required to be part of the search.



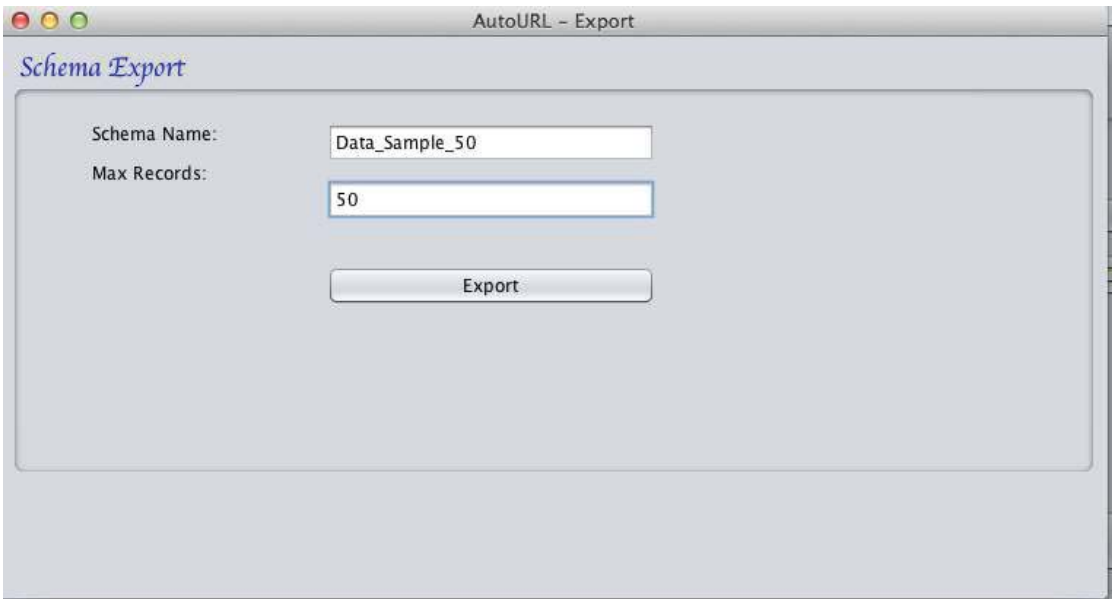
The above screenshot shows all organizations are listed in order of their frequency of reporting complaints.

A.7 Reset Function:

Click “Reset” button from the main window to reset all the fields to null where new data can be entered.

A.7 Export Function:

Click “Export” button from the main window to enter the export function. This is the start of the most important (rather heart of the entire program of functions) where it also leads to Automatic URL Tracking.

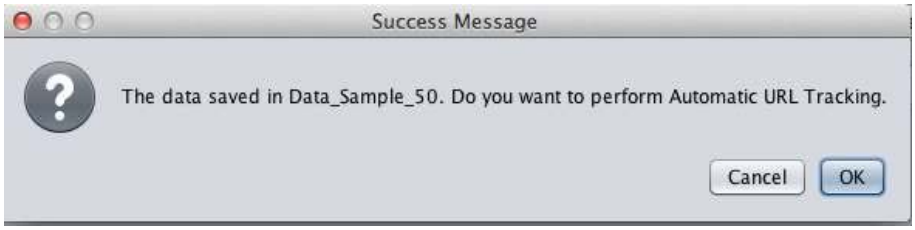


Upon clicking the “Export” button from the main window a new export windows appears. This window allows users to define the database (i.e. schema) name, maximum records to be exported and by clicking the button “Export” button.

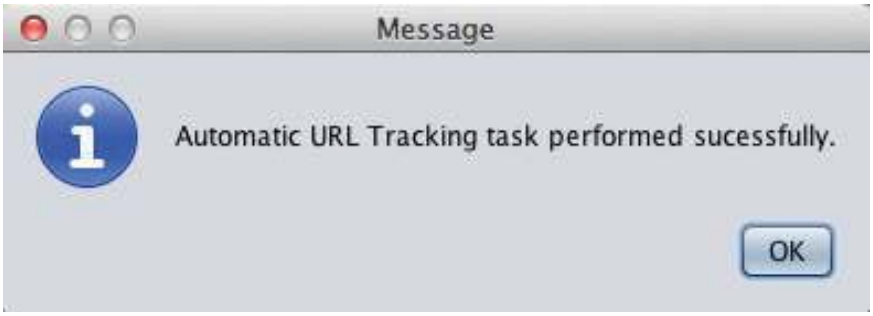
A.8 Automatic URL Tracking Function:

This process is automatic and once the process of exporting completes the program prompts user that the data is stored in their defined database and is ready for automatic tracking. Do they wish to proceed with automatic tracking; if users response is yes then it will automatically load the automatic URL function to perform its task.

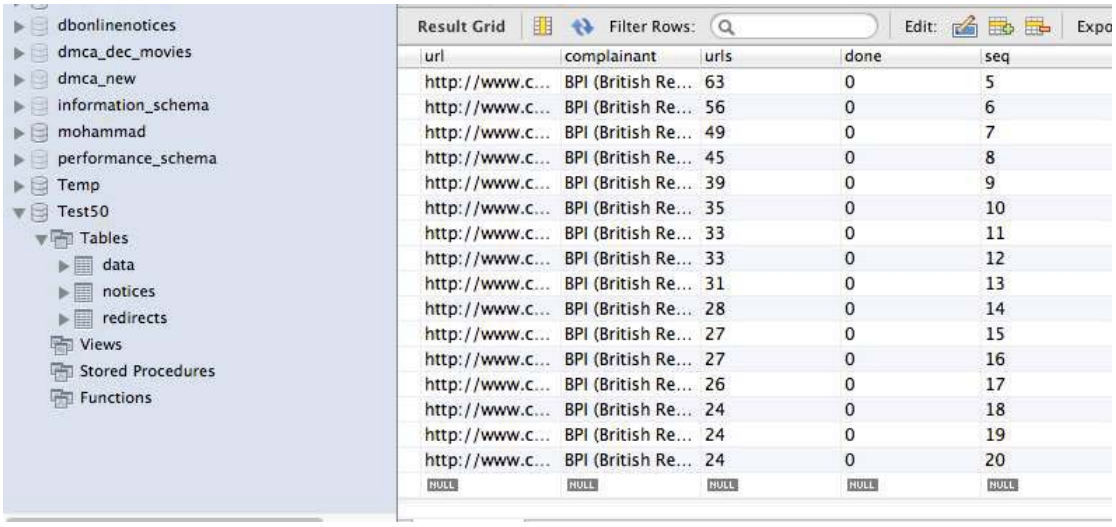
This process happens in the background and it only prompts the user to confirm the start and prompts again to start the automation.



Once the automation process completes, it confirms to the user its task completion.



Then the MySQL Workbench can be checked to view the created tables and their data records.



This is the end of the user manual for Automatic URL Tracking software.

Appendix B
Automatic URL Tracking Software Database Tables Structures

The tables below describe the database table structures used or created by the Automatic URL Tracking Software. The Automatic URL Tracking software connects to MySQL database system using the Java Driver Manager connection. This code is found within "ONJavaConnect.java" program file.

B.1 Used Tables

This table exists within the default database schema.

requests table

Name	Description
id	An ID unique to each copyright removal request
date	The date and time (in UTC) that the request was received, in ISO 8601 format
ceurl	URL to the Chilling Effects page documenting the request
owner_id	The ID number of a unique copyright owner
owner_name	The name of the copyright owner associated with the request
org_id	The ID number of a unique reporting organization
org_name	The name of the reporting organization associated with the request

url_removed	The number of URLs removed.
url_no_action	The number of URLs for which we took no action
url_pending	The number of URLs that are still pending review
from_abuser	If the request was submitted by someone we believe to be abusing the process
seq	Record sequence number, automatically generated.

B.2Created Tables

These tables are created within the user named database (i.e. schema) during execution of the application inside export functionality, when it is triggered through "Export" button.

notices table:

Name	Description
url	URL to the Chilling Effects page documenting the request
complaint	The name of the copyright owner associated with the request
urls	The number of URLs removed.
done	"0" for un-completed and "1" for completed
seq	Record sequence number, automatically generated.

data table:

Name	Description

page	Page number (i.e. request number) from selected samples
link	Identified link within the above page
hostname	Hostname for the link
Domain	Domain name for the link
advertiser	Found advertiser
category	Single character category

redirect table:

Name	Description
page	Page number (i.e. request number) from selected samples
link	Link number (from the page) of the selected samples
url	URL being checked for redirection
domain	Domain name for the url
seq	Record sequence number, automatically generated.

Appendix C
Auto URL Tracking Environment

Setup Guide

To create the development environment for Automatic URL Tracking Software, the following steps must be followed.

The Auto URL Tracking application has been developed entirely in Java environment using Netbeans IDE Swing for the front end and for the back end MySQL Workbench is being used.

C.1Download Tools

The following tools are required to be downloaded.

- Latest version of Netbeans IDE from the website.
  Link: https://netbeans.org/downloads

- Latest version of MySQL Workbench from the website
  Link: http://dev.mysql.com/downloads/workbench

NOTE:

In case of Windows machine, the downloaded executable package will automatically download.NET Plateform, MySQL server and other required patches but in case of Mac platform MySQL server must be downloaded separately.

C.2Installation Tools

Once the above files are downloaded then they must be installed by double clicking their downloaded executable packages.

Appendix D
Google Transparency Report

Table Structures

The following table structures are for the file containing raw data of copyright removal requests described in the Transparency Report. The data is stored in three comma-separated-value (CSV) files and is updated with the same frequency as the Copyright section of the Transparency Report. The columns in each CSV file are described below.

Requests (requests.csv)

Name	Description	Req?
Request ID	An ID unique to each copyright removal request	Yes
Date	The date and time (in UTC)	Yes
Chilling Effects URL	URL to the Chilling Effects page documenting the request	No
Copyright owner ID	The ID number of a unique copyright owner	Yes
Copyright owner name	The name of the copyright owner associated with the request	No
Reporting organization ID	The ID number of a unique reporting organization	Yes
Reporting organization name	The name of the reporting organization associated with the request	No
URLs removed	The number of URLs removed.	Yes
URLs for which we took no action	The number of URLs for which we took no action.	Yes

URLs pending review	The number of URLs that are still pending review.	Yes
From Abuser	If the request was submitted by someone we believe to be abusing the process	Yes

Domains (domains.csv)

Name	Description	Req?
Request ID	An ID unique to each copyright removal request	Yes
Domain	A normalized domain specified within the request	Yes
URLs removed	The number of URLs removed.	Yes
URLs for which we took no action	The number of URLs for which we took no action.	Yes
URLs pending review	The number of URLs that are still pending review.	Yes
From Abuser	If the request was submitted by someone we believe to be abusing the process	Yes

URLs for which we took no action (urls-no-action-taken.csv)

Name	Description	Req?
Request ID	An ID unique to each copyright removal request	Yes
Domain	A normalized domain specified within the request	Yes
URL	A URL that was specified in the request but not removed	Yes
From Abuser	If the request was submitted by someone we believe to be abusing the process	Yes

Appendix E
Automatic URL Tracking Software

Program Code

Software codes in this section are the programming codes for each individual program within the package together with the main function associated with them.

Each program name starts with the prefix of “ON” which stands for “Online Notices” which indicates all these programs are within the “OnlineNotices” package.

Due huge code size the files are added to the drop box and the folder link is as followings:

Dropbox Link:

https://www.dropbox.com/sh/6d7xhqn51ph6hqw/AABE_0WyP6QyoDdnaUkgXuN6a?dl=0

Appendix F
Automatic URL Tracking Output

Result Tables

Tables and data in this section are produced as a result of research carried out for this thesis. These are exact data outputs produced by the application.

Due huge table size the tables are added to the drop box and the links is as followings:

Dropbox Link:

https://www.dropbox.com/s/e5o8ewnxe493ewh/Appendix%20F%20-%20Automatic%20URL%20Table%20Output.docx?dl=0

Appendix G
Automatic URL Tracking Software Program Execution Output

The output in this section is produced by the Automatic URL Tracking Software, which is produced during the execution of the application for producing the results, used for this research. These results (i.e. tables) are also included in Appendix F of this documentation.

Due huge file size the output file is added to the drop box and the links is as followings:

Dropbox Link:

https://www.dropbox.com/s/5b6m10u49nxf1t0/Appendix%20G%20-%20Automatic%20URL%20Tracking%20Software%20Program%20Execution%20Output.docx