MASSEY UNIVERSITY

MAXIMISING THE EFFECTIVENESS OF THREAT RESPONSES USING DATA MINING: A PIRACY CASE STUDY

Seung Jun Lee

Master of Information Sciences

Information Technology

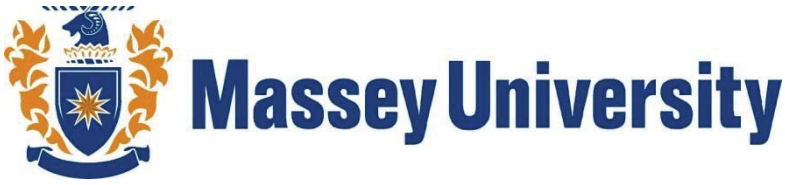
To my

Beloved grandmother and uncle who left me during the writing of this thesis,

I dedicate this thesis to them and may their souls rest in peace…

Dedication also goes to my beloved parents and my wife who have always been very supportive with much love and providing their sincere support to me during the entire research of this thesis.

I also dedicate this thesis to my beloved grandfather who left us a few years ago who always gave lots of support and love to my beloved family…



MAXIMISING THE EFFECTIVENESS OF THREAT RESPONSES USING DATA MINING: A PIRACY CASE STUDY

This thesis is presented in partial fulfillment of the requirements for the degree of

Master of Information Sciences in
Information Technology

School of Engineering and Advanced Technology
At Massey University Albany, Auckland, New Zealand

By

Seung Jun Lee, PG Diploma in Information Sciences and
B Information Sciences (Massey University)

August 2015

Acknowledgements

I would like to thank my supervisor, Professor Paul Watters for his sincere support, clear guidance, encouragement and strong belief in me during the entire process and research of this thesis. It was much privileged to have an invaluable opportunity working on this thesis project with his continuous support and helpful insight throughout the past year.

I am also very grateful to my dearest family, especially my parents who have always been very supportive to me with much love, endless support and patience. I also would like to thank my wife, Sily for providing me with much support, dedication and love to our family. My gratitude also goes to all the staff and faculty members of the School of Engineering and Advanced Technology at Massey University, Albany for their positive support and assistance over the past few years. I also would like to thank all my friends at Massey University who were shared lots of invaluable moments and great ideas that helped me to enhance and build more motivation during the past few years of studies.

Abstract

Companies with limited budgets must decide how best to defend against threats. This thesis presents and develops a robust approach to grouping together threats which present the highest (and lowest) risk, using film piracy as a case study. Techniques like cluster analysis can be used effectively to group together sites based on a wide range of attributes, such as income earned per day and estimated worth. The attributes of high earning and low earning websites could also give some useful insight into policy options which might be effective in reducing earnings by pirate websites. For instance, are all low value sites based in a country with effective internet controls? One of the practical data mining techniques such as a decision tree or classification tree could help rightsholders to interpret these attributes.

The purpose of analysing the data in this thesis was to answer three main research questions in this thesis. It was found that, as predicted, there were two natural clusters of the most complained about sites (high income and low income). This means that rightsholders should focus their efforts and resources on only high income sites, and ignore the others.

It was also found that the main significant factors or key critical variables for separating high-income vs low-income rogue websites included daily page-views, number of internal and external links, social media shares (i.e. social network engagement) and element of the page structure, including HTML page and JavaScript sizes. Further research should investigate why these factors were important in driving website revenue higher. For instance, why is high revenue associated with smaller HTML pages and less JavaScript? Is it because the pages are simply faster to load? A similar pattern is observed with the number of links. These results could form a study looking into what attributes make e-commerce successful more broadly.

It is important to note that this was a preliminary study only looking at the Top 20 rogue websites basically suggested by Google Transparency Report (2015). Whilst these account for the majority of complaints, a different picture may emerge if we analysed more sites, and/or selected them based on different sets of criteria, such the time period, geographic location, content category (software versus movies, for example), and so on. Future research should also extend the clustering technique to other security domains.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ARFF	Attribute-Relation File Format
ASCII	American Standard Code for Information Interchange
BT	Behavioral Targeting
CDA	Communications Decency Act
COPA	Children Online Protection Act
COPPA	Children Online Privacy Protection Act
CPC	Cost-Per-Click
CPM	Cost-Per-Thousand-Impressions
CSS	Cascading Style Sheet
CSV	Comma Separated Values
CTR	Click-Through-Rate
DMCA	Digital Millennium Copyright Act
DMOZ	Directory Mozilla
EU	European Union
FTC	Federal Trade Commission
GUI	Graphical User Interface
HIC	High Income Cluster
HTML	Hypertext Markup Language
IP	Internet Protocol
IQR	Interquartile Range
IT	Information Technology
LIC	Low Income Cluster
LVF	Lower Visual Field
NA	Not-Applicable
OVA	Online Video Advertising

P/E	Price-To-Earnings ratio
PPA	Pay-Per-Auction
SAS	Statistical Analysis System
SEO	Search Engine Optimisation
TRA	Theory of Reasoned Action
URL	Uniform Resource Locator
US	United States
USC	University of Southern California
UVF	Upper Visual Field
WEKA	Waikato Environment for Knowledge Analysis
WOT	Web of Trust

CHAPTER ONE

Introduction

1.1 The Cost of Online Piracy and Cyber Security

In recent years, cyber security threats have increase exponentially, putting corporate Information Technology (IT) budgets under severe strain, as managers try and work out the best way to organise and fund defences against potential cyber-attacks.

The research literature suggests that the problem will endure: Shane & Hunker (2013) implied from wide perspective that the overall size and elaboration of cybercrime will continue to increase constantly. Furthermore, Shane & Hunker (2013) indicated the fact that every few months or even every couple of weeks appears to continually produce up-to-date news of the worst ever cybercrimes or cyber theft activities that may be still occurring as huge concerns around the globe. As a typical example of such serious cyber theft incidents from Arthur (2011), (as cited in Shane & Hunker (2013)) also implied that the seriousness of the cyber theft or cybercrime activity can certainly be realised through the consecutive cyber thefts incident of more than 100,000,000 PlayStation user confidential accounts from Sony which has significantly occurred in April 2011. Therefore, this particular incident provides a meaningful implication to both the public and industries that it is important for them to consistently maintain the robust cybersecurity infrastructure and ensure the effective protection countermeasures for sustaining particularly their intellectual properties and copyright contents against various types of potential cybercrime activities or threats such as illegal film and music contents piracy, high-risk online advertising and links to piracy and banking malware etc.

However, it is a current fact that the sustainability of ensuring or protecting their valuable intellectual properties/contents against potential cyber threats or privacy is primarily associated with the effective establishment of the optimised security countermeasures or solutions which may often lead the relevant organisations or rightsholders into generating a large amount of budgets for constructing the optimised security countermeasures efficiently.

In addition, people often ask, why is going to the cinema so expensive? Or why do bank charge such expensive bank fees for their accounts? The answer, partly, is that the security measures which are put in place are very expensive.

Copyright removal requests received for Search in the past month

54,810,885	URLs Requested to be Removed
81,161	Specified domains
5,987	Copyright Owners
2,671	Reporting Organisations

Table 1: A numerical information about the most recent copyright removal requests received for search in the past month as of 15/08/2015

(Source from: Google Transparency Report (2015)).

As illustrated in Table 1 above, this particular numerical data shows that Google in the past month (July 2015) alone received very surprisingly 54,810,885 notifications to remove or eliminate items from the search index that infringe copyright. Assuming that the cost of sending a single notification in regard to copyright violated contents is approximately between $10 and $100, this means that many hundreds of millions or even billions of dollars are immensely spent on security measures. Imagine if this money could be spent elsewhere in the industry. For instance, this kind of possible scenarios can be appeared by giving monetary grants to young film makers or funding emerging artists to record their first CD etc.

Figure 1: A trend graph showing URLs requested to be removed from search per week as of 15/08/2015 (Source from: Google Transparency Report (2015)).

As illustrated in Figure 1 above, this particular trend graph from the recent Google Transparency Report (2015) shows about the total number of URLs requested to be removed from search per week. According to Google Transparency Report (2015), this trend graph indicated that the following six enormous measurements of URLs requested by copyright owners and reporting organisations were observed as below based on the several timeframes:

(1) On week of 18 July in 2011, it was measured as 129,822 URLs in total.

(2) On week of 23 July in 2012, it was measured as 1,669,841 URLs in total.

(3) On week of 29 July in 2013, it was measured as 4,536,644 URLs in total.

(4) On week of 04 August in 2014, it was measured as 6,957,143 URLs in total.

(5) On week of 27 July in 2015, it was measured as 12,773,487 URLs in total.

(6) On week of 03 August in 2015, it was measured as 12,241,970 URLs in total.

As illustrated in Figure 1, one of the most significant findings from this trend graph clearly implied that there is an enormously increasing trend in regard to URLs requested by both copyright owners and reporting organisations to be removed from week of 18 July 2011 to week of 03 August 2015. Moreover, Figure 1 showed that the highest value of URLs requested to be removed was measured particularly on week of 27 July in 2015 and very surprisingly 12,773,487 URLs in total were appeared to be removed during this particular timeframe (i.e. week of 27 July in 2015). Most recently, Figure 1 showed that 12,241,970 URLs in total was measured during the week of 03 August in 2015.

The question of whether security budgets are being effective is therefore critical to the future success of creative industries, as is the case for justifying expenditure on countermeasures in any security environment. To make budgets effective, security managers need to assess the risk posed by different threats. In the case of film piracy, this can be done by looking at the value of the sites which are responsible for enabling piracy. We can use a standard business valuation methodology – such as the price/earnings (P/E) ratio – to do this.

According to Russell Indexes (2015), it is primarily defined that “The Russell 2000 Index is designed to calculate the performance of the selected small-cap stock market index or segment in accordance with the United States (US) equity universe. The Russell 2000 Index can be described as a subset of the Russell 3000 Index, which accounts for about 10% of the market capitalization of the index. It is primarily consisted of about 2000 shares of the small securities on the basis of a mixture of both present index membership and their stock market cap. The Russell 2000 is established to offer a broad, unbiased and impartial small-cap barometer and reconstitution is entirely continued annually to prevent the potential distortion from larger stocks in terms of both the features and performance of the true small-cap opportunity set”.

We can work out the value of the Top 20 piracy sites using a simple formula, as shown below:

Valuation in Russell 2000 Index = P/E Ratio * Total annual amount of income from Top 20 rogue sites

E.g. Valuation in Russell 2000 Index = 78.97 * Total annual amount of income from Top 20 rogue sites

= 78.97 * $63,409,908.24

= $5,007,480,453.71

Thus, the size of the threat is significant – film piracy is more than a $5 billion enormous industry!

In this thesis, the question of whether countermeasures against piracy are effective, is essentially addressed. In particular, the assumption that the risk posed by piracy or rogue websites is uniform, is questioned. Indeed, while the evidence suggests that the number of piracy websites is continually growing, and that rightsholders continue to issue numerous complaints, no-one has examined before whether these websites in fact attract any users at all. In this study, the link between being complained about and revenue is directly examined.

The basic hypothesis in this study is that there are actually two distinct groups of piracy or rogue websites: (1) the high income website and (2) the low income website. Furthermore, the hypothesis will be tested in this study by examining the top twenty most complained about the rogue websites, using cluster analysis through the implementation of a simple K-means clustering algorithm.

The main purpose of testing this particular hypothesis by examining the top twenty most complained about websites based on using a data mining technique (cluster analysis) is to see if there are two natural groupings of these top twenty rogue websites (commercially successful or not) as follows: the high income website (i.e. those sites that generate revenue) and the low income website (those sites that do not generate much revenue).

If it is found that a large proportion of the most complained about websites in fact make or generate no revenue, these could be excluded from future notice generation campaign, saving a significant amount of money or budget from the perspectives of copyright owners and reporting organisations.

In addition, if a profile can be developed by identifying the key attributes of commercially successful piracy or rogue websites, this could be used in the future to classify new piracy websites into the categories of successful versus unsuccessful sites, and informed decisions about budget expenditure could be made on a rational basis.

1.2 Advertising and Risk

This particular section introduces about the following five related-contents comprehensively in regard to this study:

(1) The Digital Millennium Copyright Act (DMCA)

(2) Chilling Effects Database

(3) Google Transparency Report (i.e. Copyright removal requests from Google index)

(4) Mainstream advertising will be basically discussed and how piracy is funded or supported.

(5) High-risk advertising will be basically discussed and their links to piracy websites will also be introduced briefly in this particular section (i.e. section 1.2.5).

1.2.1 The Digital Millennium Copyright Act (DMCA)

This particular section describes briefly about the DMCA as below.

Google Transparency Report (2015) highlighted that the DMCA is a copyright law from the United States (US) that is primarily designed to deviate from monetary liability for copyright infringement or violation in the online service or business provider such as Google. According to Wikipedia (2015), the beginning of the DMCA has been officially enacted since on October 28, 1998 in the US. In the DMCA, Google Transparency Report (2015) also indicated that relevant online operators or service providers have the responsibilities to remove material that is allegedly a copyright infringing claims or violated contents immediately. Therefore, Google Transparency Report (2015) emphasised that one of the core requirements in the DMCA is that online service providers like Google should respond immediately by removing these copyright violated materials or claims (i.e. or by disabling access to these copyright violated contents) in relation to safe harbour provisions if certain requests are received or reported that essentially satisfies the requirements of the DMCA. Moreover, Google Transparency Report (2015) confirmed that Google complies faithfully with respect to the requirements of the DMCA in regard to responding the removal requests of copyright for the purpose of providing more assurance and transparency particularly for supporting their users.

1.2.2 Chilling Effects Database

This particular section describes briefly about chilling effects database as below.

Chilling Effects (2015) indicated that Chilling Effects is a distinct collaboration from the Electronic Frontier Foundation and the following various law school clinics based in the US such as George Washington School of Law, Berkeley, Stanford and Harvard. According to Chilling Effects (2015), it is primarily a research-based project from the foundation of the Berkman Center for Internet and Society in regard to stopping and desisting about online contents. Chilling Effects (2015) also indicated that the main feature of Chilling Effects is to collect complaints and also analyse about various types of online activities such as removing or deleting content online. According to Chilling Effects (2015), the aim of this particular effect is largely classified into the following three goals as below:

(1) To educate the communities or public, (2) To promote effective research about various types of complaints received particularly on the deletion request from service providers or online publishers and (3) To provide the highest transparency as possible.

Furthermore, Google Transparency Report (2015) highlighted that in place of removed content or material, Google appeared to connect their distinct search results with the requests suggested and posted by Chilling Effects when Google is able to implement it within legitimate scope.

1.2.3 Google Transparency Report

This particular section describes about Google transparency report as below.

Google Transparency Report (2015) suggested that Google currently provides the following seven extensive transparency reports in terms of these main aspects below:

(1) Google provides a detailed transparency report about various requests for removing content from government.

(2) Google provides detailed transparency report about government various requests to hand over some information such as user data and account information about Google users.

(3) Google also provides detailed transparency information on either demands or requests by copyright holders upon their request to remove search results.

(4) Google provides detailed transparency report about overall information about Google product traffic such as traffic patterns since 2008 and the availability regarding Google products around the globe in the real-time based etc.

(5) Google provides detailed transparency report about comprehensive statistics based on weekly detection from the number of malware websites including phishing websites. Detailed Google transparency report about which networks basically host or contain malware websites is also provided in terms of overall safe browsing aspects from Google.

(6) Detailed Google transparency report on the overall content/traffic volume of email exchange between other service providers and Gmail is provided by Google to ensure the protection against snooping over the Internet. According to Google Transparency Report (2015), this particular Google transparency report also provides detailed information through the optimised encryption method of email in transit.

(7) Google also provides a detailed transparency report about statistics on European privacy requests for removing content from search results through the implementation of Google's distinct data protection removal process in Europe.

In addition, Google Transparency Report (2015) showed that the following main significant contents can be described as below in regard to requests to remove content perspectives:

- Copyright owners and their representative organisations regularly send requests to Google to remove content from search results that may contain a high possibility of connecting particular link to material or contents that is allegedly violated copyrights. Furthermore, Google Transparency Report (2015) indicated that each request basically names particular URLs to be deleted and then Google provides a list of the domain parts of the requested URLs for processing content removal under a specifically designated domain.
- Google Transparency Report (2015) also indicated that around the globe, there are currently a large number of requests received from various government agencies to Google in regard to requesting content or information removal from various Google-related products. This particular transparency report from Google Transparency Report (2015) indicated that the basic reviews from these various government agencies to remove content is first determined by Google carefully to decide and ensure whether the corresponding contents should be deserved to remove because of the violation of a law/copyrights or relevant policies from Google before any action is undertaken.

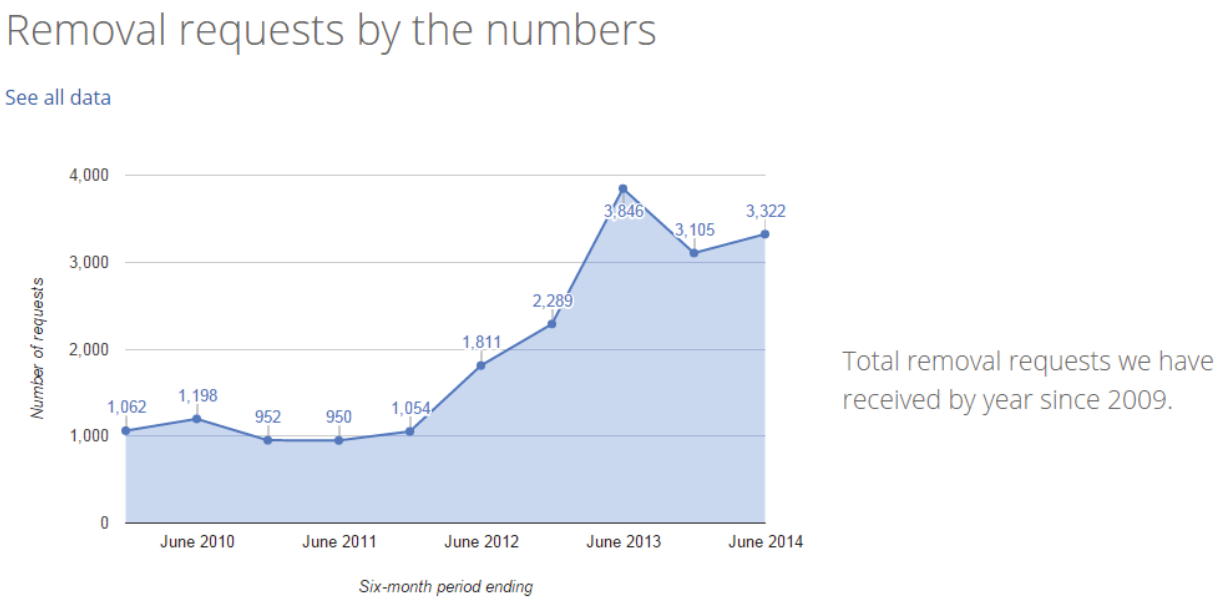


Figure 2: A trend graph showing the total removal requests received from Google in regard to government requests around the world to remove content since 2009.

(Source from: Google Transparency Report (2015)).

As illustrated in Figure 2 above, one of the most significant findings from this trend graph clearly implied that there is an enormously increasing trend between December 2011 and December 2012 in relation to the requests from government agencies around the world to remove content. During the timeframe between December 2012 and June 2013, Figure 2 also shows that an immensely increasing trend was observed in regard to government requests around the globe to remove content due to the violation of a relevant copyright law or product policies as determined by Google.

The significant research process of this thesis is particularly focused on the essential data collection from Top 20 Specified Domains within the following main section – Due to copyright, Requests to remove content in Google Transparency Report (2015). Based on data from this Top 20 Specified Domains, the research methodology has been examined and implemented to generate the significant results by utilising k-means cluster analysis, to develop approach to grouping together which presents the highest revenue risk to rightsholders or copyright owners based on Top 20 most complained rogue websites (i.e. Specified Domains from Google Transparency Report (2015)). For more detailed information, this particular process to obtain the significant findings generated in the research will be discussed in chapter three – Research Methodology section.

1.2.4 Mainstream Advertising and How Piracy is funded

This particular section describes about mainstream advertising and how piracy is funded.

An advertising transparency report from University of Southern California (USC) Annenberg Innovation Lab (2013) in January 2013 indicated the fact that the new advertising networks currently seem to show enormous growth of advertising market and inventory over the past five years in the wideband time. However, this particular report from USC Annenberg Innovation Lab (2013) showed that many parts of these advertising inventories are appeared to exist on over 150,000 copyright infringed entertainment websites which are also generally known as pirate websites. This particular report from USC Annenberg Innovation Lab (2013) indicated that it is basically designed to provide a comprehensive monthly overview about the Top Ten advertising networks which allow placing the most ads and then leading into many illegal file sharing websites in connection with these ads. According to USC Annenberg Innovation Lab (2013), the following significant findings were observed about the relationship between mainstream advertising and how piracy is basically funded as below:

- A related-report from PRS for Music & Google (2012), (as cited in USC Annenberg Innovation Lab (2013)) investigated that a large number of current advertising networks can be clearly affected to support the various activities of pirate websites which are mainly based on the category of movie and music. This report from PRS for Music & Google (2012) indicated that 86% of the peer-to-peer (P2P) search sites with illicit file sharing content or distributed material are appeared to be funded financially by advertising method. PRS for Music & Google (2012) implied from this significant finding that a variety of main brands are not really realised about the fact that advertising is primarily an important source of funding to support the key activities of piracy-based illicit industries such piracy movie or music sites.
- A related-information from Google Transparency Report (2013), (as cited in USC Annenberg Innovation Lab (2013)) indicated that over 2,300,000 particular URLs have been observed particularly from Filestube.com in terms of copyright violation. Hence, this enormous result from Google Transparency Report (2013) implied that these serious activities from illegal file sharing or piracy websites are affected very negatively indeed to the creative industries or community around the globe by maintaining to steal the important intellectual properties or assets such as copyrighted material or unique trademark.
- Furthermore, USC Annenberg Innovation Lab (2013) in February 2013 suggested that it can be clearly seen from the corresponding list based on the discovered infringing websites that particularly many young adults appeared to be seen as having a strong attraction to the following categories such as mobile phone, car, car insurance and credit rating agencies on the piracy sites. Hence, USC Annenberg Innovation Lab (2013) implied a possible reason behind this meaningful finding that the frequency of advertising occurred in the case of American Express can be seen often on rogue or piracy sites as an example.

1.2.5 High-Risk Advertising and Their Links to Piracy Websites

This particular section describes about high-risk advertising and their links to piracy websites as below. This is more important that just looking at whether companies are being wayward, because high-risk advertising exposures to children such as online gambling, scams, pornography and banking malware etc. in the real life can be very harmful through a number of various real studies or cases that have already been widely examined and implemented in Australia, New Zealand and Canada as follows:

1.2.5.1 High-Risk Advertising: Case Studies in Canada

This particular study from Watters (2015) indicated the fact that although most nations around the globe are currently appeared to maintain an imperfect regulation in terms of censorship, at the same time, the sovereign rights to defend themselves should be recognised. This study also suggests that Watters (2015) examined and implemented an effective approach about how the unregulated Internet regulation can be seriously influenced to produce significant harms to the users. This study showed that in the sample based on Canadian users, the overall 5,000 pages of rogue websites have been considered and analysed on the basis of particularly the most complained TV shows and movies. In this study, Watters (2015) also identified that 12,190 advertising items in total were found and 3,025 ads in overall were appeared as visible ads category. As a result, this study from Watters (2015) highlighted that the following significant findings were found: (1) 89% of these ads above delivered to Canadian users were appeared as High-Risk ads and (2) 11% of these ads above delivered to Canadian users were appeared as mainstream ads. As illustrated in Table 2 below, this particular table from Watters (2015) shows the overall frequency distribution for the case of high-risk ads only in terms of ads category such as malware, gambling, scams and download etc. Table 2 showed that the highest risk category from advertising was appeared as malware (i.e. 43.6%) which allows the banner ads to prevalently lead into the other potential high-risk links including malwares. In Table 2, the next highest risk ads were appeared from ads based on sex industry (i.e. 30.0%) and also from ads based on scams (i.e. 18.2%) respectively.

	Sex Industry	Malware	Download	Gambling	Scams
N	805	1,172	106	113	489
Percentage	30.0%	43.6%	3.9%	4.2%	18.2%

Table 2: Frequency by ad category – High-Risk ads (Source from: Watters (2015)).

1.2.5.2 High-Risk Advertising: Case Studies in Australia

This particular section describes about high-risk advertising and their links to piracy websites in Australia. As an objective of this particular study from Watters (2014, January), a systematic approach has been evolved to analyse and investigate about online advertising to target on Australians. Watters (2014, January) suggested that this study is particularly concentrated on sites for the Top 500 DMCA complaints in regard to mainly TV and movie content which were supported by Google. As a result, this systematic approach from Watters (2014, January) highlighted that the following significant findings were found: (1) 99% of these ads were appeared as High-Risk ads and only 1% of these ads were appeared as mainstream ads. (2) This study showed that in the sample, only one website was found in regard to displaying mainstream ad only; this study also showed that the other remaining sites appeared to contain no ads or only ads were displayed by sources generated from High-Risk, or mainstream ads were observed as a small number.

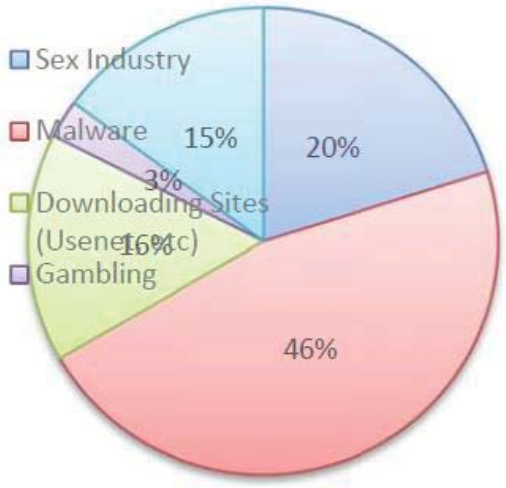


Figure 3: A pie-chart distribution of High-Risk Advertisings

(Source from: Watters (2014, January)).

As illustrated in Figure 3 above, this particular table from Watters (2014, January) shows the overall distribution for the case of high-risk ads only in terms of ads category such as malware, gambling, sex industry and downloading sites. Figure 3 showed that the highest risk category from advertising was appeared as malware (i.e. 46.49%) which allows the banner ads to prevalently lead into the other high-risk links including malwares. In Figure 3, Watters (2014, January) indicated that the second highest risk ad was appeared from ads based on sex industry (i.e. 20.18%). In addition, Watters (2014, January) showed that 14.91% category was appeared on the basis of scams which consist of various types. (e.g. investment scams). In Figure 3, this particular category for scams (i.e. 14.91%) is basically referred to the corresponding light blue allocation where it displays a distribution: 15% from the pie-chart distribution. Moreover, Figure 3 showed that in comparison, the prevalence of High-Risk advertising from both malware (i.e. 46.49%) and downloading sites (i.e. approximately 16%)

in Australia is higher than the prevalence of High-Risk ad from both malware (i.e. 43.6%) and downloading (i.e. 3.9%) in Canada as shown in Table 2. On the other hand, Table 2 showed that in comparison, the prevalence of High-Risk advertising from both sex industry (i.e. 30.00%) and scams (i.e. 18.2%) in Canada is higher than the prevalence of High-Risk advertising from both sex industry (i.e. 20.18%) and scams (i.e. 14.91%) in Australia as shown in Figure 3.

1.2.5.3 High-Risk Advertising: Case Studies in New Zealand

This particular section describes about high-risk advertising and their links to piracy websites in New Zealand. According to the previous related-study from Watters (2014, January), (as cited in Watters, Watters & Ziegler (2015, January)) this study used an approach as suggested by Watters (2014, January) that in the sample based on New Zealand users, the overall 5,000 webpages have been basically considered, captured and analysed on the basis of particularly the most complained TV shows and movies. In this study, Watters, Watters & Ziegler (2015, January) also identified that 5,547 advertising items in total were found and 2,268 ads in overall were appeared as visible ads category which are primarily based on these 5,000 webpages collected. As a result, Watters, Watters & Ziegler (2015, January) highlighted that the following significant findings were found in this study: (1) Between 93 and 96% of these ads as indicated above delivered to New Zealand users were appeared as High-Risk ads for TV, movies and music only and (2) Between 3% and 7% of these ads delivered to New Zealand users were appeared as mainstream ads.

	Sex Industry	Malware	Download	Gambling	Scams
N	317	1,257	257	140	207
Percentage	14.55%	57.71%	11.80%	6.43%	9.50%

Table 3: Frequency by ad category – High-Risk ads

(Source from: Watters, Watters & Ziegler (2015, January)).

As illustrated in Table 3 above, this particular table from Watters, Watters & Ziegler (2015, January) shows the overall frequency distribution for the case of high-risk ads only in terms of overall ads category such as malware, gambling, scams, sex industry and download. Table 3 showed that the highest risk category from advertising was appeared as malware (i.e. 57.71%) which allows the banner ads to prevalently lead into the other potential high-risk links including malwares or malicious code. In Table 3, Watters, Watters & Ziegler (2015, January) also showed that the next highest risk ads were appeared from ads based on sex industry (i.e. 14.55%) and also from ads based on download (i.e. 11.80%) and followed by ads on scams (i.e. 9.50%) respectively. Moreover, Table 3 showed that in comparison, the

prevalence of High-Risk advertising from both malware (i.e. 57.71%) and gambling (i.e. 6.43%) in New Zealand is higher than the prevalence of High-Risk ad from both malware (i.e. 46.49%) and gambling (i.e. 3.00%) in Australia as suggested by Watters (2014, January) and Figure 3. On the other hand, Watters (2014, January) showed that in comparison, the prevalence of High-Risk advertising from sex industry (i.e. 20.18%), download (i.e. approximately 16%) and scams (i.e. 14.91%) in Australia is higher than the prevalence of High-Risk advertising from sex industry (i.e. 14.55%), download (i.e. 11.80%) and scams (i.e. 9.50%) respectively in New Zealand as shown in Table 3.

According to Watters, Watters & Ziegler (2015, January), this particular study already showed that between 93 and 96% of these ads delivered to New Zealand users were appeared or classified as High-Risk ads for TV, movies and music only. Between 3% and 7% of these ads were appeared as mainstream ads in the case of New Zealand. Therefore, this result implies that the allocation of High-Risk ads (i.e. between 93 and 96%) delivered to New Zealand users certainly have less High-Risk ads compared to the allocation of High-Risk ads (i.e. 99%) delivered to Australians as identified by Watters (2014, January). A recent study from Watters (2015) also confirmed that 89% ads primarily delivered to Canadian users were appeared as High-Risk ads for the most complained TV shows and movies. This particular study from Watters (2015) showed that 11% ads delivered to Canadian users were appeared as mainstream ads. Hence, this result clearly implies that the allocation of High-Risk ads (i.e. between 93 and 96%) delivered to New Zealand users as identified by Watters, Watters & Ziegler (2015, January) certainly have more High-Risk ads compared to the allocation of High-Risk ads (i.e. 89%) delivered to Canadians as identified by Watters (2015).

Watters, Watters & Ziegler (2015, January) also provided an important implication that High-Risk ads on malware were appeared as 57.71% across TV and movie sites only. However, Watters, Watters & Ziegler (2015, January) provided a significant finding in this study that enormously 96.34% ads for music category only were classified as malware.

1.3 Research Scope and Research Objectives

The **objective** of this thesis is:

To develop a robust approach to grouping together which present the highest revenue risk to rightsholders or copyright owners. Techniques like cluster analysis can be used effectively to group together sites based on a wide range of attributes, such as income earned per day and estimated worth. The attributes of high earning and low earning websites could also give some useful insight into policy options which might be effective in reducing earnings by pirate websites. For example, are all low value sites based in a country with effective internet controls? One of the practical data mining techniques such as a decision tree or classification tree could help rightsholders to interpret these attributes.

1.4 Research Questions and Hypotheses

The main research questions of this thesis are described as follows:

Research Question 1 – What are the main significant factors or key critical variables which can be determined to influence proportionally more into either the revenue results or estimated worth between high-income rogue website vs low-income rogue website from the URLs of Top Twenty rogue websites initially obtained by Google Transparent Report (which of them are most complained about Top Twenty rogue websites (or pirate websites)).

Research Question 2 – What are the main significant indications or prediction that can be able to discover or derive from the results or patterns of certain variables used in entire analysis? (e.g. results from location longitude, location latitude, Google pagerank, ALEXA rank – cluster analysis)

Research Question 3 – What are the main significant indications or prediction that could find out or derive from the analysis of clustering? (a simple K-means clustering algorithm etc.)

Two types of sites were hypothesised, and cluster analysis was used to confirm this.

The hypothesis of this thesis is described as follows:

Hypothesis – there are two groups of rogue sites which are most complained about – high value and low value – if hypothesis is true, then law enforcement and rightsholders should only be targeting high value sites, because these represent the greatest risk. The null hypothesis is that the most complained about sites are all high value.

1.5 Structure of the Thesis

The basic structure of this thesis is described as follows:

This chapter (i.e. Chapter One) provided an overview of the thesis by setting the cost of online piracy and cyber security, advertising and risk, introducing the research scope and research objectives and also describing the research questions and hypotheses. In addition, this chapter provided the flow and brief description of research methodology in the thesis. This chapter also provided information about business valuation based on Russell 2000 Indexes (2015) in relation to the relevant research process for the thesis. The chapter one now ends with a structure of the thesis and the following remaining chapters of this thesis are as below:

Chapter Two – This chapter presents and provides detailed reviews of the entire related works primarily done, investigated and conducted by various academics and researchers in relation to this thesis. This chapter provides a comprehensive detailed literature review on the basis of the following four main topics such as online advertising and all the related contents within online advertising, High-risk advertising and Piracy Websites.

Chapter Three – This chapter presents and describes in detail regarding the research methodology and statistical data mining technique used and implemented in this study. This chapter also provides a detailed description of the entire research methodology through the following main processes such as data collection and project implementation based on data source by using k-means clustering analysis or technique. Furthermore, this chapter provides a brief background in regard to research methodology and the chapter also includes theoretical information about cluster analysis and descriptive statistics.

Chapter Four – This chapter comprehensively analyses, examines and interprets the final significant outcomes or data generated from the implementation of the project in this study.

Chapter Five – This chapter provides the conclusion and discussion with a summary of the significant research findings in relation to this study.

Chapter Six – This chapter concludes the thesis with an overview of future work and this chapter also introduces about possible future research in terms of the following further components: (a) Utilisation of applying different timeframes in this study, (b) Utilisation of applying and implementing through various factors such as geographic location and various types of content category and (c) Utilisation of applying other data mining techniques.

1.6 Summary

This chapter demonstrates that we can use a standard business valuation methodology such as the price/earnings (P/E) ratio to enhance the effectiveness of security budgets by assessing the risk posed by different threats or potential cyber-attacks from the perspective of security management. This chapter also demonstrates that online advertising is a critical source of supporting the various illicit activities of piracy music or piracy movie sites. In addition, this chapter implies that high-risk advertising exposures to people especially children such as online gambling, scams, sex industry and banking malware can be very harmful through a number of various real cases in countries such as Canada, Australia and New Zealand. This chapter also demonstrates that their links to piracy websites may lead into the associated links containing various types of malwares which can be further threats or social issues to both the users and many nations around the globe. Furthermore, this chapter describes about mainstream advertising and also demonstrates regarding how piracy is funded or supported in regard to mainstream advertising environment.

CHAPTER TWO

Literature Review

2.1 Background of Online Advertising

This particular section (i.e. 2.1 Background) will be discussed and reviewed about a general background of Internet and online advertising identified by the various related-literature reviews of researchers and other academics. This section will also be designed to provide a comprehensive typology of the various types of online advertising that were basically classified and investigated mainly in terms of the academic literature reviews from them.

2.1.1 Online Advertising and Behavioral Targeting (BT)

In this section, the state of the art in Behavioral Targeting (BT) in online advertising is examined. A key empirical study from Yan et al (2009) suggested a number of main properties and influences between BT and online advertising:

- Recently, this study indicated that the effective application of BT is an innovative technique which is used to expand the effectiveness and usefulness of their online and Internet advertising campaigns by the corresponding advertisers and it is certainly the current increasing trend which is taking a more significant role in the online advertising market these days.
- However, it is a current phenomenon that the method regarding how effective the BT that can be directly influenced and targeted to online advertising on search engines is currently being studied in the institute or academia. In this particular empirical study, the impact of BT has been revealed based on experiments in which the click-through record or log of the advertisements through the data collection process of a commercial search engine.
- In this study, the following three significant findings as shown below have been discovered through the samples of more than six million users and 17,901 experiments based on the click-through log of real world advertisements for the duration of over seven days timeframe:
  (1) It has been observed that users who made the selection of clicking or choosing the same advertisement will clearly generate the similar patterns or behaviours on the web.

(2) Click-Through Rate (CTR) of an advertisement can be grown up to 670% on average over the entire ads collected in this particular study by appropriately dividing or distributing the targeted users for BT-based advertising on the sponsored search ads. In this outcome, one core condition has given that it is also important to adopt the most essential clustering algorithms for the user in terms of BT perspectives.
(3) In this study, it has also been observed that it is more preferable and much effective to use the short-term user behaviours to basically stand for the targeted users on BT which are compared to using or applying the long-term user behaviours on BT. Therefore, it indicates that tracing the short-term user search behaviour (i.e. various search activities) may certainly generate much better result or performance than tracing the long-term user browsing behaviour (or activities) in terms of BT.

Furthermore, a recent economic study from Chen & Stallaert (2014) highlighted the following significant features and current trends of online advertising based on BT broadly from the standpoint of an economic analysis:

- In recent years, it has been shown that both online publishing companies and advertisers are constantly seen to generate the expanding interest and higher attention in the use of targeted advertising based on online environment. On the basis of the users' previous search and browsing behaviour/habits and also their additional information (e.g. various types of activities or interests registered on a website) which are currently available, this particular online targeting as suggested from the study of Chen & Stallaert (2014) basically allows the online publishing companies and advertisers to present users with ads that are very appropriate and a well-fitted effectively. Hence, it indicates from Chen & Stallaert (2014) that BT is a technique which has a very close relationship with online advertising in terms of its prospective effectiveness or usefulness.
- This study has described the related-economic effects or implications in the cases where online publishers (or advertisers) are undertaken or implemented in BT.
- In this study, it has discovered the critical factors that can be directly influenced to the revenue of publishers or publishing companies, social welfare and the remunerations of the advertisers by utilizing a modelling technique called "Horizontal Differentiation Model" to obtain the optimised fit between an user and an advertisement shown.
- When using BT of online publishers, Chen & Stallaert (2014) has also suggested that it may possibly increase the profit doubled in some situations or cases, however, it is not fully ensured or guaranteed for the online publisher to receive the benefits of increased revenue/profit.
- In this study, the following two significant effects in regards to BT have been identified: (1) a Competitive Effect and (2) a Propensity Effect. In addition, as a relative strength of the two main effects, these particular effects can be a core decision point to determine whether the revenue generated has a relative effect that could bring the significant impact for the publisher.

- This study has also revealed that small advertisers are preferable to maintain an online advertising infrastructure based on BT, although social welfare is expanded. On the other hand, the leading or dominant advertiser in the market might be reluctant to adopt it under BT by shifting from conventional advertising.

2.1.2 Online Advertising and Intrusiveness

This particular section introduces about the reviews of the related-literatures in regards to online advertising and intrusiveness.

According to the study from McCoy et al (2007), the following significant findings and relationships were observed comprehensively between online advertising and intrusiveness of various forms of web advertisements (e.g. In-Line ads, Pop-Up ads and Pop-Under ads) as shown below:

- This study has primarily highlighted that negative characteristics and components of the advisements have significant relative effects in retaining both the site maintenance and ad contents. Furthermore, this research is basically designed to deliver a support for a contention that users will likely have an overall negative view or intentions of a website being displayed with advertisements rather than the site which does not display the ads.
- In this study, one of the most interesting findings is that In-Line advertisements allow both the site and ad contents to be recalled more distinctly compared to Pop-Up ads and Pop-Under ads. The reason for being this effect is that the performance/act of closing the ad window can be interfered the users in using or utilizing the website. In addition, therefore, it is basically seen to be visible to the user's point of view for a shorter time or moment indeed when such distraction affects the users by closing the ad window.
- The previous related-research primarily from Denes (2001), (as cited in McCoy et al (2007)) has identified the specific experiment result of implementing the website actions between the banner ad and Pop-Up ads in terms of interference analysis. In this particular study, Denes (2001) has proved that 84% of respondents indicated that Pop-Up ads can certainly be a significant distraction factor to the users when using, viewing or accessing the contents within a web page. On the other hand, Denes (2001) has demonstrated that 54% of respondents only indicated that banner ads can be a main factor to influence the interference to the users when viewing or reading the contents within a web page. (as cited in McCoy et al (2007))
- McCoy et al (2007) suggested that the following significant findings were observed in this study as below:

(1) This research provides an implication that intrusiveness is basically a very critical factor to all the corresponding users such as advertisers and web designers within an online advertising environment.
(2) This study suggests that viewers perceive certain web advertisements (i.e. Pop-Up ads and Pop-Under ads) to be more intrusive than In-Line ads.

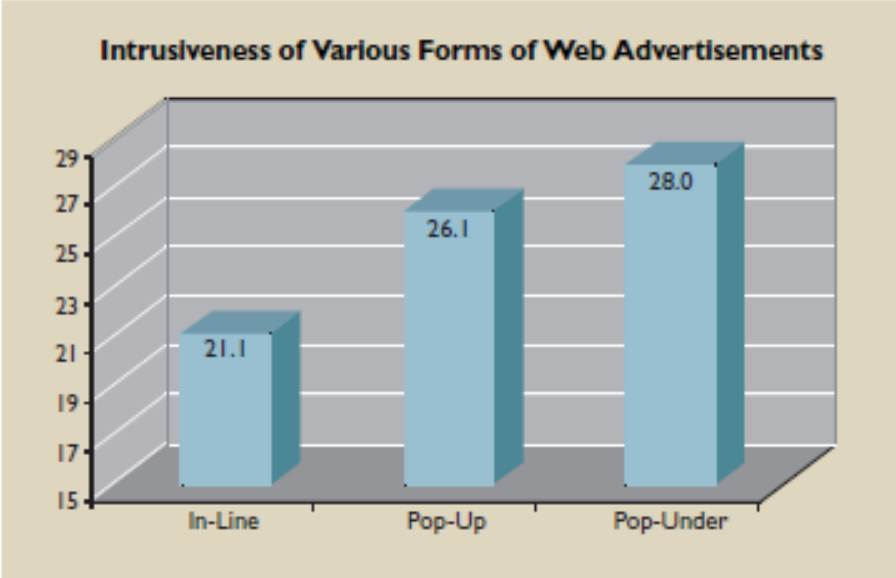


Figure 4: Perceived intrusiveness depending on various forms of web advertisements (i.e. In-Line, Pop-Up and Pop-Under) – (Source from: McCoy et al (2007))

As illustrated in Figure 4 above, it implies that the bar-graph result of Pop-Up advertisements were observed as 24% more intrusive compared to the bar-graph result of In-Line advertisements. (i.e. p-value is less than 0.001). Furthermore, the bar-graph result of Pop-Under advertisements was appeared as 33.1% more intrusive significantly compared to the result of In-Line advertisements. (i.e. p-value is less than 0.001).

(3) This study from McCoy et al (2007) has also provided a significant finding additionally that the various forms of web advertisements such as In-Line ads, Pop-Up ads, Pop-Under ad and No Ad etc. can be implied to influence the website user's decision negatively in regards to intentions of returning or revisiting the main site. As illustrated in Figure 5 below, it shows that the intentions to returning to the site with Pop-Up ads and Pop-Under ads is indicated clearly less than revisiting the site involved with In-Line ads. In addition, the bar-graph result from Figure 5 also indicates that the intention to returning or revisiting to the site involved with No Advertisement is indicated as the highest response compared to other three advertisement types (i.e. In-Line ad, Pop-Up ad and Pop-Under ad). Moreover, Figure 5 as illustrated below shows that the value of intention to returning or revisiting to the site involved with In-Line ads is also relatively higher than Pop-Up ads or Pop-Under ads.

(4) According to McCoy et al (2007), this study has also provided an implication that In-Line advertisement and Pop-Up advertisement are largely different in terms of intrusiveness's level. A host may be available to operated it safely and implement the effective use of utilizing the In-Line ads.

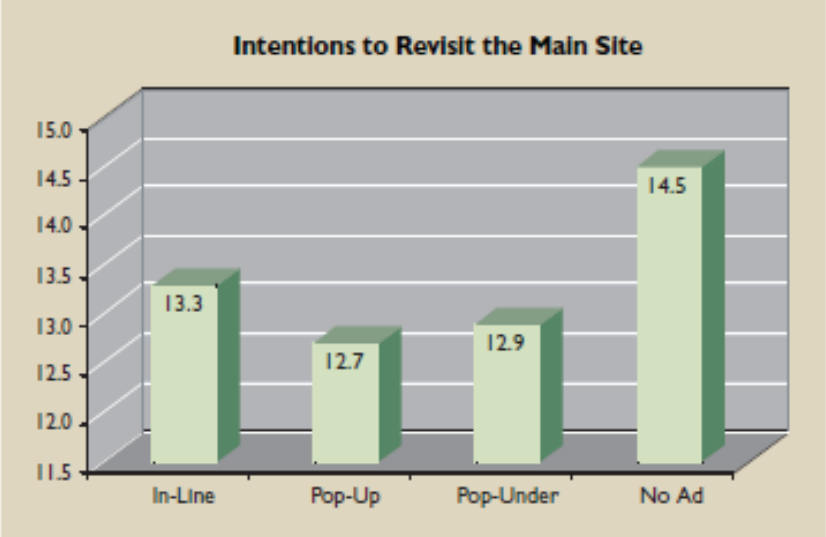


Figure 5: Intentions to return or revisit the website containing the various forms of web advertisements (Source from: McCoy et al (2007)).

2.1.3 Online Advertising and Privacy

This particular section introduces about the reviews of the-related literatures in regards to online advertising and privacy.

A study from Goldfarb & Tucker (2011) highlighted that the following significant properties and relationships between online advertising and privacy regulations were comprehensively described as below:

- In recent years, it has been shown that online advertisers are utilizing their online customer information/data in order to aim their marketing being more appeals. Therefore, this study from Goldfarb & Tucker (2011) suggests that this has basically raised or caused the privacy issues for customers. As an establishment to solve these concerns for the customers, the governments have constantly enforced and updated the related-laws and regulations to protect customer privacy and it may lead the governments to be passed the laws through. For instance, the protection of consumer privacy can be implemented by utilizing the various types of methods or solutions (e.g. restriction in regards to the use of data or restriction of implementing online tracking technologies which is normally used by sites).

- In this study, Goldfarb & Tucker (2011) suggests an experiment that the total number of responses from 3.3 million survey-participants has been used and these survey-participants during this particular experiment had been basically exposed to 9,596 online advertising campaigns randomly. Furthermore, in this study, the purpose of this particular random survey is to investigate or discover how the privacy regulation (or laws) has affected the online advertising effectiveness in the European Union (EU).
- This study from Goldfarb & Tucker (2011) observes that as the privacy laws came into effect in Europe, banner advertisements (i.e. displaying advertising) have generated an experiment outcome of 65% on average about the reduction based on effectiveness in Europe in terms of status change for a purchase intention. However, this study shows that in the case of non-European countries during the same period, similar changes in regards to advertising effectiveness were not found or discovered. Therefore, on the basis of this significant evidence as suggested from Goldfarb & Tucker (2011) above, this study proves the fact that the effectiveness of advertising can be reduced by establishing or applying appropriate privacy regulation.
- In this study, it suggests that reduction in the effectiveness of online advertising was more appeared specifically in websites that typically contain more general information or content (e.g. news and web services sites). In addition, this study suggests that the reduction in the effectiveness of online advertisements was also observed to be appearing more in advertisements that basically consist of the following features: no further interactive, visual or audio features. In addition, this study demonstrates that there is a direct correlation between privacy regulation and reduction of effectiveness for these advertisements.

According to the study of Evans (2009) from economic perspectives, the following significant features between online advertising and privacy were also suggested as below:

- This study has highlighted that online advertising is generally known as providing a series of two-sided markets or platforms that can be enabled to perform the matching between advertisers and consumers. This study from Evans (2009) shows that the key activities of these main mediators are progressively improving to build the matching of advertisers and consumers effectively. Evans (2009) also suggests that these particular activities between advertisers and consumers can be used by utilizing or implementing the following methods: (a) Use of Matching Algorithms, (b) Use of Predictive Methods and (c) Use of detailed customer or individual data.
- However, this study describes that some of these methods as mentioned above may possibly cause the public policy problems. Therefore, this study suggests that the appropriate balancing process is required by providing valuable online advertising services which will be offered to the consumers or end-users. Moreover, Evans (2009) suggests that the use of detailed consumer or individual data from intermediaries may lead into the possibility of losing privacy from the standpoint of consumers. Hence, this study also implies that the consideration of privacy in regards to online

advertising should be carefully taken into account when advertisers normally handle/use their valuable customer data or important consumer information online.

A related-study from Guha, Cheng & Francis (2010) suggested that the following significant properties and relationship between online advertising systems and privacy were also highlighted as below:

- In recent years, this study indicates that online advertising is designed to provide a wide range of various Internet or online services including online search engines, social media services or e-mail applications etc. However, Guha, Cheng & Francis (2010) suggests that the problems of valuable privacy loss in regards to user targeting perspectives are known as becoming very prevalent and huge concerns for many users and customers.
- This study from Guha, Cheng & Francis (2010) suggests the fact that nowadays there is a limited knowledge regarding the operation of ad networks in the public. Therefore, this study implies that the primary operation and method of advertisement networks can become a predominant concern especially when they typically access user data in order to target or aim the users.
- This study introduces about a specific measure methodologies in regards to advertisement network (i.e. known as ad network) which is developed by the technical theory and enhancement from Guha, Cheng & Francis (2010). Furthermore, this study suggests that there is an analysis regarding how the following three different types of ad networks can be commonly used to utilize or access user profile data: (a) Online Social Network Ads, (b) Search Ads and (c) Contextual Ads.

2.1.4 Online Advertising and Obtrusiveness

This particular section introduces about the reviews of the-related literatures in regards to online advertising and obtrusiveness.

According to the study of Goldfarb & Tucker (2011), the following significant findings and main features were observed between online advertising and obtrusiveness:

- This study suggests an investigation about what fundamental factors can be directly affected the effectiveness of advertising. This study from Goldfarb & Tucker (2011) primarily uses a large scale experiment data from the total number of 2,892 online display advertising campaign which is based on various types of key groups or categories used in this particular experiment.
- Firstly, this study suggests that the performance of following online advertising strategies are very effective and work well when they only operate independently: (a)

Highly visible online advertising (i.e. increasing ad's obtrusiveness) and (b) Context-based online advertising (i.e. an ad which is focused on website content). However, when these two strategies are combined then this study from Goldfarb & Tucker (2011) indicates that they basically appear not to perform or operate well when combined. In addition, this study shows that this particular outcome can often be appeared more in classes of certain products which are likely preferable by people or users being more private as well as protecting their valuable privacy at the same time. Therefore, this study from Goldfarb & Tucker (2011) suggests that customer's basic perceptions in regards to privacy aspect can be induced from the ineffectiveness between the combination of context-based online advertising and highly visible (i.e. obtrusiveness) online advertising.

- Secondly, this study suggests that online advertising has been increasingly growing and it is also continually maintaining which is largely divided between highly contextual targeting ads and more obtrusive online ads. For instance, this study suggests that the success of a certain product (e.g. AdSense from Google) is a good example in regards to this particular finding. However, this study highlights that the utilization of relatively less targeted ads are used in the current online advertising environment.
- Finally, this study also suggests that policy implications are being derived on the basis of both privacy and intrusiveness in order to ultimately follow the same route of government policy or regulations. The previous related-research primarily from Shatz (2009), (as cited in Goldfarb & Tucker (2011)) has identified that there is a constantly increasing movement and ongoing pressure in Europe and the United States in order to control the use of data in regards to browsing various forms of behaviours or activities to target ad. Furthermore, this study from Goldfarb & Tucker (2011) suggests that the appropriate consideration of using a potential solution such as trade-off can be applied or needed by regulators in regards to such regulation as suggested by Shatz (2009), (as cited in Goldfarb & Tucker (2011)).

2.1.5 Online Advertising and Economic Factors

This particular section introduces about the reviews of the-related literatures in regards to online advertising and economic factors.

According to the study of Asdemir, Kumar & Jacob (2012), the following economic factors and main features in regards to online advertising were suggested based on the two pricing models were (i.e. cost per thousand impressions (CPM) and cost per click (CPC)) as below:

- This study emphasises that the continuous success of online advertising mainly from the components such as software and services is demonstrating a current fact nowadays as seen through the tremendous evolvement from Google and the related-firms. Hence, this study from Asdemir, Kumar & Jacob (2012) implies that the optimised selection or effective decision of a pricing model is consistently a critical component in online advertising from the standpoint of these related-industries or firms.
- This study suggests that the following two significantly most popular pricing models in online advertising have been used comprehensively to design a suitable model in terms of pricing models perspective by utilizing a game-theoretic framework of the principal-agent based:
  (1) Cost per thousand impressions (CPM) this is primarily input-based.
  (2) Cost per click-through (CPC) which is primarily performance-based.

- This study from Asdemir, Kumar & Jacob (2012) indicates that the following four significant factors are also suggested in regards to influencing the selection of a pricing model between CPM and CPC as shown below:

  (1) Uncertainty Effect: this study suggests that this particular factor can be described as an uncertainty over the boundaries of target segment. In addition, Asdemir, Kumar & Jacob (2012) implies that the feature of this uncertainty is designed to favour the CPC pricing model for both the publisher and the online advertiser.
  (2) Exposure-value Effect: this study suggests that the main feature of this factor is designed to affect the value of online advertising to the target segment. This study also indicates that uncertainty effect can be moderated by using or applying exposure-value effect for the online advertiser. Furthermore, this study suggests that for the online advertiser, exposure-value effect can be used to favour the CPM pricing model. For the publisher, Asdemir, Kumar & Jacob (2012) suggests that this particular effect can also be used to favour the CPC pricing model.
  (3) Mistargeting Effect: this study from Asdemir, Kumar & Jacob (2012) suggests that the basic property of mistargeting effect is normally known as the cost in association with mistargeting ads to customers or users in the case of non-target segment.

(4) Alignment Effect: this study from Asdemir, Kumar & Jacob (2012) suggests that alignment effect is an influential factor or effect which stands for the difference for both online advertiser and the publisher in terms of the incentives-based alignment.

Moreover, this study from Asdemir, Kumar & Jacob (2012) suggests that these four factors as indicated above can be affected to generate impact or conflicts between online advertisers and publishers in terms of the pricing model preference (i.e. CPM or CPC).

Thomes (2013) indicated in this study from the standpoint of an economic analysis that the following main features are suggested in terms of online streaming music services as below:

- This study suggests that a theoretical analysis of an online streaming music-based business model is largely classified into the following two forms of services as below. In addition, this study indicates that these two distinctive forms of services are offered by a sole owner or monopoliser.
  (1) Thomes (2013), this study also suggests that the main significant features of this first online streaming music service are consisted of the following components: (a) low quality of online streaming music services is basically provided, (b) the process is entirely costless and (c) this particular service (i.e. first service) is primarily financed by the method of advertising.
  (2) This study also suggests that the main significant features of this second online streaming music service are consisted of the following components: (a) high quality of online streaming music services is primarily offered and (b) this second service is fully operated by charging their online streaming music users or customers.
  Therefore, this study from Thomes (2013) implies that these two distinct forms of online streaming music services are largely distinguished by both the quality and the cost.
- Furthermore, this study demonstrates that the monopoliser or sole owner can be able obtain the benefits by using the method of an online adverting funding if users for online streaming music services are very lenient in regards to online ads or commercials. Therefore, this study from Thomes (2013) suggests that users of the charged service to support demand for online advertising-based service may be charged by a large portion of the costs or a high price. Moreover, the result from this study suggests that there are a number of effective and influential policies such as music provision supported by advertising and the welfare consequences of such business model as mentioned in order to solve the issues against digital piracy nowadays.

According to the economic study of Evans (2008), this study suggests that the same economic considerations of the search advertising platforms is widely used and applied in

traditional media platforms at the same time in terms of various aspects. This study also suggests that the fundamental correlation between these two types of platforms (i.e. search ad platforms and traditional media platforms) is basically performed as providing two-sided businesses or two-sided market structure which is based on using business strategies to attract the customer's interest or attention and selling access to these strategies to advertisers. However, this study from Evans (2008) suggests that the search advertising platforms have a various number of distinctive features of the technologies compared to traditional media platforms.

This particular study of Evans (2008) from economic perspectives suggests that the following two significant economic factors and properties in regards to online advertising system were comprehensively highlighted as below in terms of affecting two-sided market structure as below:

2.1.5.1 Economic Factor of Online Ad: Pricing of Keywords

According to Evans (2008), this study suggests that the following significant economic factors in regards to online advertising were appeared comprehensively in terms of the pricing of keywords as below:

- Evans (2008) highlights that on each platform, the cost-per-click (CPC) is eventually decided by the activity of the keyword bid auction. This study from Evans (2008) shows that the outcome of those auctions may generate an implication in terms of similar CPCs for the given query terms if all of the following conditions are satisfied as similar status: (a) the status of the auction rules, (b) the same bidders and (c) the leads value for different platforms. Therefore, this study suggests as identified in the outcome above that auction bidder of the search advertising platform, the efficiency of auctions and the value of leads are proportional to the CPCs for particular or specific keywords.

2.1.5.2 Economic Factor of Online Ad: The Role of Indirect Network Effects

This study from Evans (2008) suggests that the following economic factors in regards to online advertising were significant in terms of the role of indirect network effects as below:

- The previous related-study primarily from Eisenmann (2007), (as cited in Evans (2008)) has identified that indirect network effects with respect to search advertising platforms are likely to be insignificant when these two components are present at all. This study from Evans (2008) suggests that the feature of search advertising platforms are alike to other forms of transaction or business platforms that are essentially designed to correspond to buyers with sellers and competent trades. Furthermore, Evans (2008) highlights that as more buyers are involved, the seller could obtain a higher possibility of an appropriate match that will affect to provide more revenue or profits for sellers. On the other hand, this study suggests that as more sellers are involved, the buyer could obtain a higher possibility of an appropriate match that will influence to provide more effective purchase for buyers. In addition, the previous related-researches primarily from Garbade & Silber (1979) and Economides (1993), (as cited in Evans (2008)) have identified that the fundamental significance of a theoretical concept called "Liquidity" is necessary component for having the mutually beneficial trades or transaction between the sellers and the buyers. Hence, this study from Evans (2008) suggests that the structure of two-sided markets are narrow and market sustainability will be difficult to maintain without having a sufficient liquidity between the mutually profitable trades of sellers and buyers.
- In this study, Evans (2008) also suggests that the basic presence of fixed costs in association with platform value difference are able to generate a significant economic effect on the search advertising platforms, providing the given fact that pricing structure of the CPC is prevalently applied in the case of smaller platforms. Moreover, this study from Evans (2008) based on economic perspective suggests that the following two main types of costs are prevalently experienced and involved by the online advertisers in order to operate the campaigns or activities in regards to search advertising platform as below:

  (1) This study indicates that these operating costs are experienced or incurred by the online advertisers such as platform set up, software installation and learning process about how to use them etc.
  (2) This study indicates that the costs of operating activities on keywords are also incurred by the online advertisers. Therefore, Evans (2008) suggests that the most appropriate decisions should be made by the advertisers on the bids in order to observe or follow the performance of the activities in regards to search advertising platform.

2.1.6 Online Advertising and Social Factors

This particular section introduces about the reviews of the-related literatures in regards to online advertising and social Factors.

According to the study from Zeng, Huang & Dou (2009), the following significant features and correlations between online advertising and social factors were highlighted comprehensively as below:

- With the advent and a bigger growth of online social networks, such as Facebook, Twitter, LinkedIn, Google Plus and MySpace, this study shows that the current trend of online social networking communities are definitely situated in the core centre of e-commerce environment nowadays. This study from Zeng, Huang & Dou (2009) basically suggests that the effective online trade-off or business transaction between user experience and advertising revenue should be balanced by the core activities and efforts of online social networking communities.
- On the basis of related-literature in both sociology and advertising aspects, this study from Zeng, Huang & Dou (2009) examines the following two key effects of (a) social identity and (b) group norms primarily about the intentions of community users' group in order to receive advertising in online social networking communities.
- This study also suggests about the effective method of how the community member's perceptions can be affected by this specific type of community users' group intention. In addition, this study outlines a suggestion about how this particular form of community users' group intention could affect to the perceived advertising value judgements in terms of utilizing such online advertising. This study from Zeng, Huang & Dou (2009) describes possible mechanisms in terms of overall response effect between community members and community advertising. For instance, this study suggests that this mechanisms process can be done by investigating which community members may affect to respond effectively to online community advertising.
- This study from Zeng, Huang & Dou (2009) suggests that as an experiment sample, the total number of samples about 327 popular social networking community users in China has been selected and implemented to use in this particular study by testing the relevant theoretical framework.
- This study from Zeng, Huang & Dou (2009) suggests that an effective use of structural model analysis can be conducted by applying and implementing the idea of maximum likelihood estimation. Zeng, Huang & Dou (2009) addresses that this structural model can also be used to primarily test the main significant hypotheses which are comprehensively provided from this study according to Zeng, Huang & Dou (2009). As indicated in Table 4 below, this particular table from Zeng, Huang & Dou (2009) are indicated as providing an overall summary of total 9 hypothesises used in this study. (i.e. from H1 to H9). Furthermore, it is important note to the reader that the detailed description of these 9 main hypothesises as displayed in Table 4 is directly quoted and obtained from the study of Zeng, Huang & Dou (2009, p. 3-4).

Hypothesis	Description
H1	"H1: Social identity relates positively to group intentions to accept advertising in online social networking communities". (Zeng, Huang & Dou (2009, p. 3)
H2	"H2: Group benefit norms relate positively to group intentions to accept advertising in online social networking communities". (Zeng, Huang & Dou (2009, p. 3)
H3	"H3: Social identity relates positively to perceived ad relevance in online social networking communities". (Zeng, Huang & Dou (2009, p. 3)
H4	"H4: Group benefit norms relate positively to perceived ad relevance in online social networking communities". (Zeng, Huang & Dou (2009, p. 3)
H5	"H5: Group intentions to accept advertising in online social networking communities relates positively to perceived ad relevance in community sites". (Zeng, Huang & Dou (2009, p. 4)
H6	"H6: Group intentions to accept advertising in online social networking communities relates positively to perceived ad value in community sites". (Zeng, Huang & Dou (2009, p. 4)
H7	"H7: Perceived ad relevance relates positively to perceived ad value". (Zeng, Huang & Dou (2009, p. 4)
H8	"H8: Perceived ad relevance relates positively to behavioural intentions to accept advertising in online social networking communities". (Zeng, Huang & Dou (2009, p. 4)
H9	"H9: Perceived ad value relates positively to behavioural intentions to advertising in online social networking communities". (Zeng, Huang & Dou (2009, p. 4)

Table 4: An overall summary of 9 main hypothesises used in this particular study.

(Source directly quoted from: (Zeng, Huang & Dou (2009, p. 3-4)).

According to the previous-related study from Price, Arnould & Tierney (1995), (as cited in Zeng, Huang & Dou (2009)) has identified that the structural model as displayed in Figure 6 below is consisted of mainly six structural model constructs or concepts. This study from Price, Arnould & Tierney (1995), (as cited in Zeng, Huang & Dou (2009)) also suggests that the summated scales are indicated by the form of presenting single indicators. In addition, the previous-related study from Hibbard, Kumar & Stern (2001), (as cited in Zeng, Huang & Dou (2009)) has identified that the process of measurement scale for each structural model construct can be established through the appropriate method of fixing the corresponding error term to one minus its reliability.

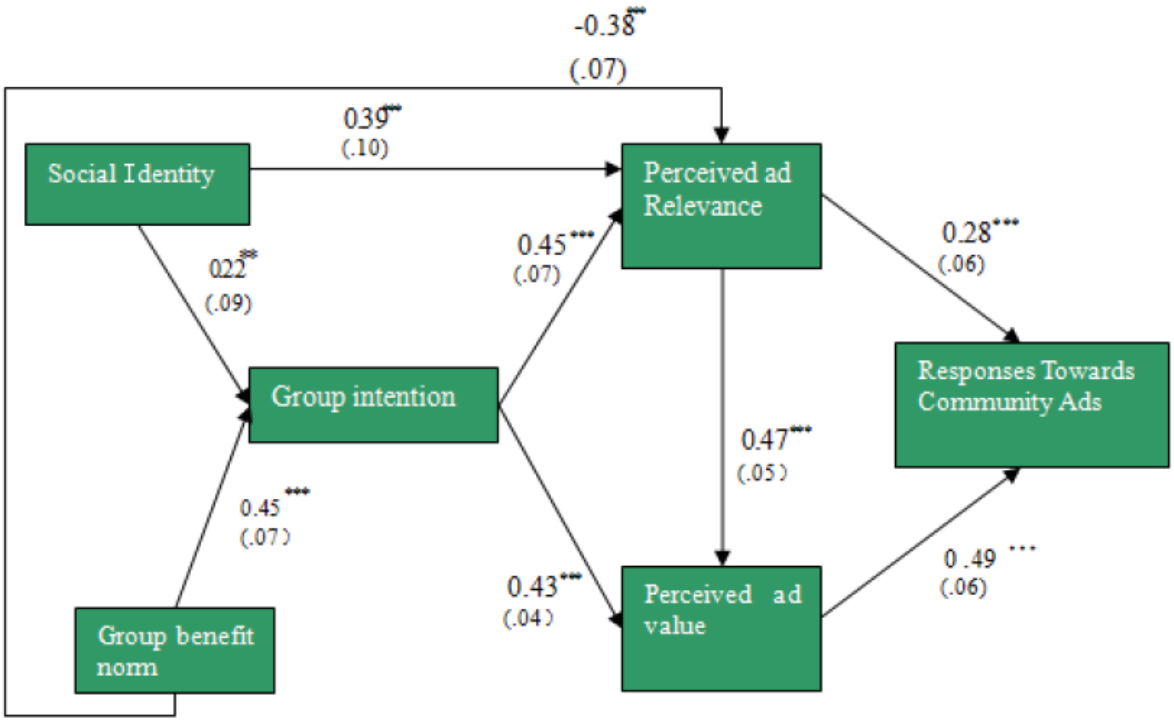


Figure 6: An Overall Illustration of Parameter Estimates for Final Structural Model

(Source from: Zeng, Huang & Dou (2009)).

As illustrated in Figure 6 above, Zeng, Huang & Dou (2009) suggests that Figure 6 essentially represents about all the related-parameter estimates in regards to obtaining final structural model in this study. According to the study from Zeng, Huang & Dou (2009), an interpretation of these is provided as follows:

- Based on the structural model information of Figure 6, this study shows that social identity appeared to have a positive effect in relation to group intention. Therefore, this result (i.e. γ-value: 0.22, p-value < 0.01) from Figure 6 implies that this correlation between social identity and group intention clearly contains a sufficient evidence to support H1 as shown in Table 4.
- In Figure 6, this study shows that group norms appeared to have a positive effect in relation to group intention. Therefore, this result (i.e. γ-value: 0.45, p-value < 0.001) from Figure 6 implies that this correlation between group norms and group intention definitely contains a sufficient evidence to support H2 as shown in Table 4.
- This study shows that social identity clearly appeared to leads to the outcome of positive perceived ad relevance. Therefore, this result (i.e. γ-value: 0.39, p-value < 0.001) from Figure 6 implies that this correlation between social identity and perceived ad relevancy definitely contains a sufficient evidence to support H3 as shown in Table 4.
- As illustrated in Figure 6, this study shows that there is a path starting from group benefit norms to perceived ad relevance. This result (i.e. γ-value: -0.038, p-value <

0.001) from Figure 6 is indicated as negative although this outcome is significant in this study. Hence, it implies that H4 as shown in Table 4 is not received support by the correlation between group norms and perceived ad relevance.

- As illustrated in Figure 6, this study shows that group intention appeared to have a positive effect in relation to perceived ad relevance. Therefore, this result (i.e. β-value: 0.45, p-value < 0.001) from Figure 6 implies that this correlation between group intention and perceived ad relevance clearly contains a sufficient evidence to support H5 as shown in Table 4.
- As illustrated in Figure 6, this study shows that group intention appeared to have a positive effect in relation to perceived ad value and it is significant. Hence, this result (i.e. β-value: 0.40, p-value < 0.001) from Figure 6 implies that this correlation between group intention and perceived ad value definitely contains a sufficient evidence to support H6 as shown in Table 4.
- Figure 6 shows that the direction starting from perceived ad relevance to perceived ad value is clearly indicated as positive result and it is also significant. Therefore, this result (i.e. β-value: 0.47, p-value < 0.001) from Figure 6 implies that this correlation between perceived ad relevance and perceived ad value clearly contains a sufficient evidence to support H7 as shown in Table 4.
- Figure 6 shows that the relationship between perceived ad relevance and responses towards community advertising is observed as both positive and significant outcome. Hence, this result (i.e. β-value: 0.28, p-value < 0.001) from Figure 6 implies that this correlation between perceived ad relevance and responses towards community advertising clearly contains a sufficient evidence to support H8 as shown in Table 4.
- Finally, Figure 6 shows that the relationship between perceived ad value and responses towards community advertising is discovered as both positive and significant result. Hence, this result (i.e. β-value: 0.49, p-value < 0.001) from Figure 6 implies that this relationship between perceived ad value and responses towards community advertising clearly contains a sufficient evidence to support H9 as shown in Table 4.

Based on the above significant results as described (i.e. Figure 6 and Table 4), this study from Zeng, Huang & Dou (2009) therefore, suggests that the following six social factors are observed as key important factors in relation to online social networking communities as below:

(1) Social Identity, (2) Group Intention, (3) Group Norms, (4) Perceived Advertising Relevance, (5) Perceived Advertising Value and (6) Responses towards Community Advertisings.

According to the study from Lee & Lee (2011), this study suggests that the following main social factors are significant in relation to the intention of watching online video advertising as below:

- The previous-related study from Ajzen & Fishbein (1980) and Fishbein & Ajzen (1975), (as cited in Lee & Lee (2011)) has identified that the theory of reasoned action (TRA) is a core theoretical framework used in this particular study that can be useful to the researchers to forecast the number of key important factors in relation to affecting consumer's intention on online video advertising (OVA). (as cited in Lee & Lee (2011)).
- This study from Lee & Lee (2011) suggests that the following three significant elements are directly involved to influence consumer intention for watching or viewing OVAs as below:
  (1) **The attitudes to watching OVAs:** this study indicates that the attitude of the participants for viewing OVAs is more positive and participants had also recognised or perceived the social pressure by watching these ads. Furthermore, this study shows that the higher consumer's intention to watch these particular OVAs is found.
  (2) **Subjective norm to watching OVAs:** this study suggests that subjective norm provide a positive impact on viewer's attitudes to watching or viewing OVAs.
  (3) **Prior frequency of watching OVAs:** this study also suggests that prior frequency of watching OVAs provide a positive impact on the consumer or viewer's intention of seeing OVA.
- This study from Lee & Lee (2011) identifies that the following six significant outcomes were basically found which allows the viewers or consumers to be experienced by viewing or accessing OVAs as below:
  (1) Information, (2) Entertainment, (3) Social Interaction, (4) Passing Time, (5) Escape and (6) Relaxation.

  Therefore, this study from Lee & Lee (2011) suggests that practical insights can be provided effectively to online advertisers in terms of these types of significant outcomes as mentioned above. In addition, this study implies that the aim of having these outcomes as above from the standpoint of advertisers is to provide the consumer's satisfaction depending on their individual requirements in relation to watching or accessing OVAs.

  According to Lee & Lee (2011), this study additionally shows an overall illustration of modified model and hypotheses testing in regard to watching OVAs as displayed in Figure 7 below.

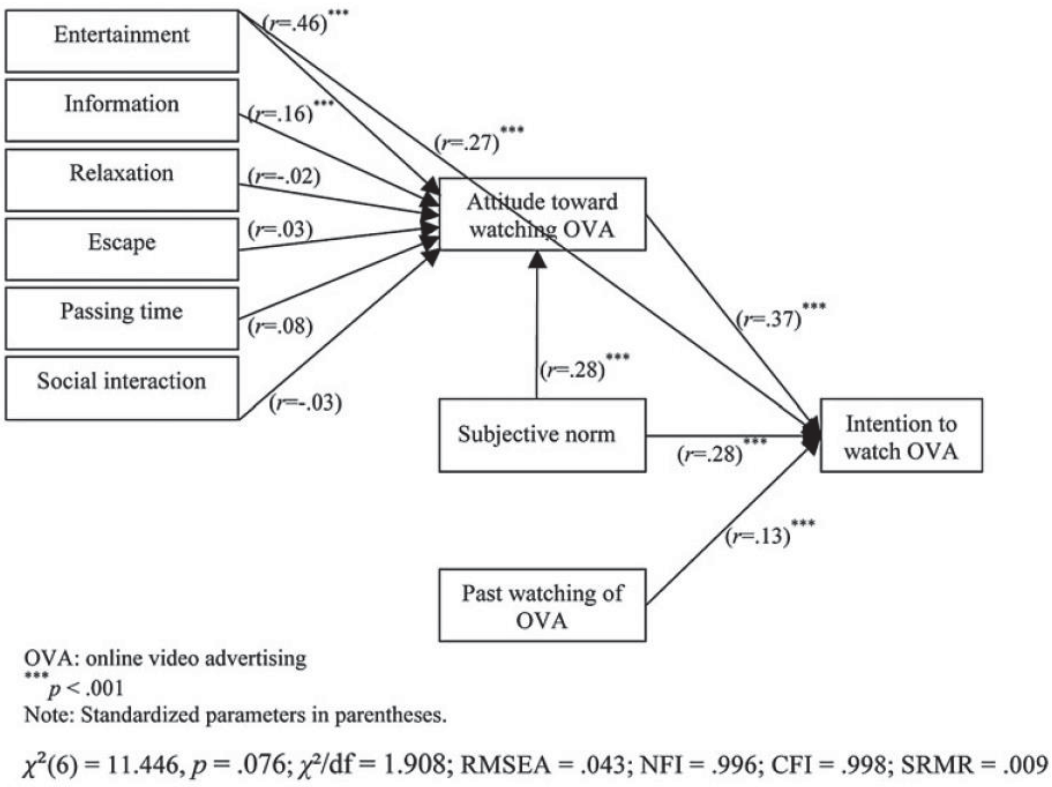


Figure 7: An Overall Structure of Hypotheses Testing and Modified Model in regard to watching OVA

(Source from: Lee & Lee (2011)).

2.1.7 Online Advertising and Positioning

This particular section introduces about the reviews of the-related literatures in regards to online advertising and positioning.

According to the study from Agarwal, Hosanagar & Smith (2011), the following significant features and correlations between online advertising and positioning were highlighted comprehensively as below:

- This study investigates that the significant pattern of direct correlation was analysed and observed between the profitability of online advertising and the effect of ad placement (i.e. ad preferred position) on profits and revenues obtained from the method of using sponsored search in online advertising.
- This study suggests that the primary data has been used for numerous hundred keywords through the activity of an online retailer for the purpose of doing ad campaign.

- This study from Agarwal, Hosanagar & Smith (2011) identifies that the following key factors of advertising placement has been measured in terms of both (a) conversion rate for these particular keywords as above and (b) click-through-rate (CTR). This study also suggests that the main theoretical concept used in this measurement is based on utilizing a hierarchical Bayesian model.
- This study discovers that the conversion rate initially begin to show a pattern of increasing and then decreasing pattern is observed with ad position for the case of using longer ad keywords while CTR basically decreases with ad placement or position at the same time.
- This study from Agarwal, Hosanagar & Smith (2011) indicates that the common sense or conventional knowledge which is widely well-known in both the industry and the public is that ad placement at the topmost position could be persuasive enough to derive into the revenue-maximizing status in terms of sponsored search ads. However, this study suggests that ad placement locating at the top higher position is not necessarily appeared to reflect the revenue-maximizing effect due to a greater rate of decrease in cost with ad placement in the case of sponsored search ads.
- This study also shows that the revenue of advertising has initially increased in accordance with the position of the longer keywords at the beginning.
- This study suggests that for online advertisers, it is better selection for them in the short term to place less weight to obtain a high position in order to maximise transactional profits.
- This study suggests that online advertising in the middle location or position is efficient way to leading to the purchase of the consumer without paying extra cost for the peak position in online advertising.
- Furthermore, this study from Agarwal, Hosanagar & Smith (2011) suggests that the following two types of countermeasures as below are possible in relation to solving potential inefficiencies in the auction system which is prevalently used by a number of most common search engines:

  (1) This study indicates that one possible solution from search engines is an active investment of technologies for the purpose of tracking online consumer's behaviour or action based on the outcome of their previous post-clicks.

  (2) In addition, this study from Agarwal, Hosanagar & Smith (2011) suggests that this particular solution from using search engines is basically designed to charge online advertisers per conversion which is normally known as the concept of pay per auction (PPA). According to InformationWeek (2007), (as cited in Agarwal, Hosanagar & Smith (2011)) this particular information also shows that various search engines have been tested and involved in regard to PPA auctions continuously.

According to the study from Goodrich (2010), the following significant features and correlations between online advertising and positioning were highlighted comprehensively as below:

- This study suggests that the following factors are important components in online advertising environment: (a) Attention, (b) Brand Attitude and (c) Aided Recall.
  In addition, this study investigates about the related-impact of the vertical positioning and type of online advertising based on these three important factors as suggested.
- Furthermore, a previous-related study from Previc (1990), (as cited in Goodrich (2010)) has primarily identified that the result of higher attention for stimuli in the lower visual field (LVF) was obtained compared to the case of upper visual field (UVF).
- The significant outcomes of this particular study from Goodrich (2010) implies that the following three critical implications can be achieved in relation to online advertising:
  (1) This study shows that ad placement at the lower page may significantly be able to affect increases ad attention. Hence, this result from Goodrich (2010) provides a clear evidence to support the previous-related research from Previc (1990) regarding both LVF and UVF research. (as cited in Goodrich (2010)).
  (2) This study from Goodrich (2010) also indicates that the correlation between brand attitude and online advertising attention is inversely related. Therefore, according to Goodrich (2010), this particular result provides a clear evidence to support mere exposure theory.
  (3) This study suggests that the type of online advertising provides a definite influence on the following key components:
  (a) Aided Recall, (b) Brand Attitude and (c) Attention.

2.2 High-Risk Advertising

This particular section introduces about the reviews of the related-literatures in relation to High-Risk Advertising.

According to the study from Austin & Reed (1999), the following significant features and correlations of both targeting children online and key ethics issues from online advertising in terms of high-risk advertising were comprehensively described as below:

- In many recent years, this study indicates that advertising websites on the basis of targeting children has shown a fast improving market. This study also shows that various forms of concerns are also involved for parents in regard to online advertising on these children-based websites. It is a fact that as children are easily exposed to access the various types of contents or materials from advertising websites through link to advertisers, therefore this study from Austin & Reed (1999) indicates that this would clearly lead into the incredible amount of expenditure power from children through the kid-based websites in relation to online advertising. Therefore, this study implies that these issues or concerns have also affected the children to be a critical target group for consumer businesses due to the fact that there are tremendous amount of expenditure power from them.
  According to the previous-related study from Azoulay (1998), (as cited in Austin & Reed (1999) has identified the following significant figures that in 1995, children under the age of 12 appeared to expend US$14 billion, teenagers appeared to spend another US$67 billion and this previous-related from Azoulay (1998) indicates that together they basically appeared to affect the remarkable figure of US$160 billion from their parents' earnings. Hence, this study from Austin & Reed (1999) implies that many reviewers appear to have a question mark over the suitability of targeting children in online advertising in recent years.
- This study from Austin & Reed (1999) proposes that the active movement of utilizing and implementing the various types of legislations and regulatory standards have been widely applied to insure that children's welfare is effectively taken into account in online advertising. This study suggests that the following three significant legislations are core regulatory issues that are appropriate to address in regard to online advertising to children:
  (1) **The Federal Trade Commission (FTC) Act:** this study suggests that FTC Act is primarily designed to prohibit the various types of activities from unjust or deceptive advertising in any forms of information media. This study from Austin & Reed (1999) proposes that the feature of FTC is to specifically inform or remind advertisers that children can be more distorted compared to adults and advertisers should also be careful not to mislead the products displayed particularly to children. According to the previous-related study from Azoulay (1998), (as cited in Austin & Reed (1999), has also identified that the FTC has the ongoing plan to continually monitor a review of the websites.

(2) **The Communications Decency Act (CDA):** This study from Austin & Reed (1999) suggests that CDA is basically know as a website legislation for the protection of children and minor from obscene, indecent and offensive speech clearly transmitted via computer networks or online communications. However, this study indicates that CDA was determined as an unconstitutional in terms of liberty of the press from the official judgement of the United States Supreme Court on June 27, 1997.

(3) **The Child Online Protection Act:** This study from Austin & Reed (1999) also suggests that this proposed legislation (i.e. The Child Online Protection Act) is primarily designed to require from those who are involved in selling or transferring any form of harmful materials to minors over the Internet. Hence, this study suggests that the proper compliance of following this particular legislation from those can be implemented effectively to restrict the access to these harmful elements or materials by minors.

- Furthermore, this study from Austin & Reed (1999) suggests that the following 4 main Internet adverting ethics issues in relation to targeting children online are basically highlighted as shown below:

  (1) Practices of Information Sharing or Information Gathering:

  This study from Austin & Reed (1999) indicates that adults with sufficient maturity can easily be able to disregard or manage the information breaches of their own privacy but children may be simply tempted in this same case by obtaining easy opportunities of free giveaways or a club membership offer etc.

  In addition, this study indicates that some websites often have an advanced technology for recording user's personal data by tracing and accumulating the online travel record of the certain users. (e.g. cookies). Therefore, this study from Austin & Reed (1999) suggests that children may also appear to lose incredibly a large amount of valuable personal data or information like the adults.

  (2) Marketing Practices:

  This study from Austin & Reed (1999) indicates that one-to-one marketing is known as one of the most concerning marketing strategies that allow the companies or organisations to use market in the online business environment.

  This study also indicates that most companies mention they do not use a one-to-one marketing strategy to target children without their parental consent. However, according to Austin & Reed (1999), this study suggests that Nabisco basically proposes a guest book upon request in order to provide a diversity of user's personal information. Moreover, this study indicates that many of the children-targeted websites are distinguished between the selling standpoint of the website and the games that are primarily offered on the basis of entertainment.

  (3) Appropriateness of both terminology and contents which are used on the webpages:

  This study from Austin & Reed (1999) indicates that many online advertising sites are prevalently connected with other websites' link and this study shows that there are in fact children accesses available from using or clicking some of these online advertising links which may lead them into sites that involve inappropriate or harmful content and language to children. Therefore, this study from Austin &

Reed (1999) implies that these two important elements (i.e. contents and language) that are essentially appropriate for children should carefully be taken into account from the perspective of online advertisers at the same time.

(4) Selling products through the use of kid's clubs:
According to Austin & Reed (1999), this study indicates that the use of kid's club is one method that advertisers can be used on their products to create or produce brand royalty. This study from Austin & Reed (1999) suggests that some online advertising websites simply do not contain any educational or entertainment value particularly for children but they appeared to provide advertisements only for the purpose of advertising their products or services. Therefore, this study implies that the utilisation of kid's club is unsuitable methods to obtain information in terms of mainly marketing purpose.

According to the recent empirical study from Auger et al (2015), the following significant features in regard to high risk online advertising were observed between online advertising effect and youth in a Canadian Sports Marketing Campaign as below:

- This study from Auger et al (2015) evaluates that there is highly sufficient potential for affecting unintended and harmful messages in online advertisements which are primarily used to target particularly youth through the random data selection process of 20 secondary school classes based in Montreal, Canada. This study indicates that this unique experiment in regard to identifying the effect between online advertisements and youth group has been conducted by utilising the example of the following Canadian sports marketing campaign called "Light It Up" from one of the leading sports corporations.
- This study from Auger et al (2015) suggests that the following methods are used respectively: (1) Based on the randomly selected sample or trial of overall 20 secondary school classes in Canada, a cluster randomised experiment has been undertaken. (2) This study from Auger et al (2015) indicates that these 20 classes were primarily designed to allocate randomly for viewing either one of the following: (a) an advertisement displayed as "Light It Up" or (b) a neutral comparison advertisement. (3) In addition, this study also indicates that the numbers of these two distinct groups are classified into the following: (a) Exposure Advertisement (i.e. N = 205) and (b) Comparison Advertisement (i.e. N = 192). As a significant implication from this particular experiment, this study from Auger et al (2015) implies that the main measurements were found as a self-imagery impact of having illicit drug messages through the practical experience of watching certain online advertisements or online advertisement slogans.
- As illustrated in Table 5 below, this study from Auger et al (2015) generates the following key significant outcomes in regard to watching a comparison advertisement and an online advertisement displayed as "Light It Up".

Outcome	Exposure advertisement (N = 205), n (%)	Comparison advertisement (N = 192), n (%)	Relative risk (95% confidence interval)	*p* value
Students reported that the				
Slogan refers to illicit drugs	17 (8.3)	3 (1.6)	5.3 (1.8–15.9)	.0045
Advertisement contains images of illicit drug products	47 (22.9)	2 (1.0)	22.0 (6.5–74.9)	.0009
Advertisement is promoting illicit drugs	25 (12.2)	10 (5.2)	2.3 (1.1–5.1)	.038
Any report of illicit drugs	55 (26.8)	14 (7.3)	4.0 (2.4–6.4)	<.0001

Table 5: An overall result of relative risks for describing illicit drug me in the exposure and comparison advertisements (Source from: Auger et al (2015)).

- As illustrated in Table 5, of the students, this study from Auger et al (2015) shows that 22.9% of them responded that the exposure advertisement (i.e. ad displayed as "Light It Up") contains images of illicit drug products compared to only 1.0% student responses for the case of comparison advertisement. Moreover, this study from Auger et al (2015) also indicates that the corresponding values of both relative risk and 95% confidence interval are shown as 22.0 and 6.5-74.9 respectively in this particular case.
- According to Auger et al (2015), of the students, Table 5 also shows that 8.3% of them responded that the exposure advertisement (i.e. ad displayed as "Light It Up") refers to slogan of illicit drugs compared to only 1.6% student responses for the case of comparison advertisement. In addition, this study from Auger et al (2015) also indicates that the corresponding values of both relative risk and 95% confidence interval are shown as 5.3 and 1.8-15.9 respectively in this particular case.
- As illustrated in Table 5, of the students, this study from Auger et al (2015) shows that 12.2% of them responded that the exposure advertisement (i.e. ad displayed as "Light It Up") is promoting illicit drugs compared to only 5.2% student responses for the case of comparison advertisement. Furthermore, this study from Auger et al (2015) also indicates that the corresponding values of both relative risk and 95% confidence interval are shown as 2.3 and 1.1-5.1 respectively in this particular case.
- As illustrated in Table 5, of the students, this study from Auger et al (2015) shows that 26.8% of them responded that the exposure advertisement (i.e. ad displayed as "Light It Up") is any report of illicit drugs compared to only 7.3% student responses for the case of comparison advertisement. In addition, this study from Auger et al

(2015) also indicates that the corresponding values of both relative risk and 95% confidence interval are shown as 4.0 and 2.4-6.4 respectively in this particular case.

- In conclusion, this study from Auger et al (2015) implies that the purpose of this particular advertisement campaign "Light It Up" is to promote the sport effectively from the standpoint of major sports corporation or advertisers. However, the corresponding important result from youth group in this study unexpectedly appeared to believe that this specific advertisement is directly associated with illicit drugs. Moreover, according to Auger et al (2015), this advertisement campaign shows how unwanted behaviours of youth can be affected by experiencing or viewing these types of online advertisements inadvertently, which may eventually lead into the significant high risk outcome to particularly youth or children.
- This study from Auger et al (2015) suggests that marketing to target youth via online advertising environment is necessary to have a strong attention from the perspectives of related-health authorities and researchers.

According to Cai & Zhao (2013), this study indicates that the following significant features and related-effects in terms of online advertising were observed between online advertising and popular children's websites as below:

- This study investigated about advertisements which are primarily published from popular websites for children.
- This study from Cai & Zhao (2013) shows that the experiments were undertaken based on the following data: a total of 117 commercial websites for children, 933 distinctive ads and 813 advertising websites. Furthermore, this study indicates that these primary data components as above were contained in the sample for this study.
- The significant outcomes of this study show that most of the websites for children contained online advertisements. In addition, one of the critical findings in this study from Cai & Zhao (2013) shows that there are significantly large number of Google advertisements published on the websites for children. This study indicates that one in three from the overall advertisements was basically appeared as Google advertisements.
- As a result, this study shows that primary website compliance with the Children Online Privacy Protection Act (COPPA) was appeared in less than half of the websites for children (i.e. 47%) and Cai & Zhao (2013) also shows that approximately a quarter portion of the online advertising websites was appeared and identified as properly complying with COPPA (i.e. 24%) when personal information regarding children was basically accumulated by these websites as above.
- Furthermore, this study implies that the positive efforts and comprehensive involvements of parents, advertisers, educators and policy makers should be undertaken effectively to protect children's online safety.

2.3 Summary

In this chapter, a review of the relevant literature for online advertising and the potential risk it poses to children and other vulnerable groups was undertaken. This chapter demonstrates that the key themes of the following topics such as intrusiveness, obtrusiveness, online advertising and Behavioral Targeting (BT), potential for harm, economic factors of online advertising, online advertising and social factors, lack of control and High-Risk advertising and its potential harmful influence (e.g. such as banking malware, online gambling, sex industry and scams) to the public especially children etc. all support the research questions that were posed at the end of Chapter 1.

CHAPTER THREE

Research Methodology

3.1 Background

In this chapter, the fundamental research methodology or primary technique applied and implemented in this research will be introduced and discussed comprehensively through this main chapter as follows:

First, it is necessary to view and identify the certain number of selected DMCA reported complaints or requests which have been randomly received from the various copyright owners (i.e. rightsholders), governments and representative organisations. According to Google Transparency Report (2015), copyright owners and their representative organisations regularly send requests to Google to delete or remove content from search results that may involve a high possibility of connecting particular link to material or contents that is allegedly violated copyrights. Furthermore, Google Transparency Report (2015) indicated that each request primarily names particular URLs to be deleted and then Google provides a list of the domain parts or segments of the requested URLs for processing content removal procedure under a specifically designated domain.

Therefore, the outcome from above implies the fact that these selected DMCA complaints or requests are primarily submitted from various copyright owners, individual rightsholder and any reporting organisations regularly to Google Transparency Report, in order to protect their intellectual property or full rights of their ownership for the digital contents that they basically own against any type of copyright violation or infringement. Furthermore, the complaints or requests reporting mechanism which is submitted from copyright owners, individual rightsholder and any reporting organisations to Google Transparency Report is allowed to provide the facilities of reporting or informing the contents that is requested by them to be fully removed from any contents, information or services generally provided by Google's products and services.

3.2 Data Collection

The next important procedure in this research methodology is to search for the specified domains (or URLs) which the corresponding copyright owners (i.e. rightsholders) for these domains have officially reported the selected DMCA complaints or requests to Google Transparency Report (2015). This particular procedure through Google Transparency Report (2015) can be available for the user/reader to view or check by accessing or clicking the following key three steps below:

Step One - Link to Google Transparency Report

(URL: https://www.google.com/transparencyreport/?hl=en_US)

Step Two - Link to Requests by copyright owners to remove search results

(URL: https://www.google.com/transparencyreport/removals/copyright/?hl=en)

Step Three - Link to Specified domains

(URL: https://www.google.com/transparencyreport/removals/copyright/domains/?r=all-time)

In the section of specified domains, there are Top Twenty specified domains that have been reported or submitted DMCA reported most complaints by two groups (i.e. copyright owners and reporting organisations). In this particular link (i.e. Specified domains), it also shows that the comprehensive information of these Top Twenty specified domains as displayed in Table 6 below is basically listed in four different tables in terms of the following timeframe respectively: (a) All Available, (b) Past Year, (c) Past Month and (d) Past Week.

In this thesis, the scope of timeframe for these Top Twenty specified domains will be used and implemented as “All Available” category. In addition, Google Transparency Report (2015) also indicates that a distinct percentage symbol (%) in Table 6 below is basically referred to the meaning that “URLs requested to be removed as a percentage of the specified domain’s indexed URLs”.

Specified domains

Past Week | Past Month | Past Year | **All Available**

Specified Domain	Copyright Owners	Reporting organizations	URLs	%
rapidgator.net	4,345	1,052	16,255,433	< 50%
4shared.com	4,094	1,012	13,531,123	< 1%
filestube.com	4,186	1,324	11,532,908	< 5%
dilandau.eu	1,460	141	11,436,314	< 5%
uploaded.net	4,136	856	10,265,761	≥ 50%
zippyshare.com	4,738	762	9,991,922	< 50%
gosong.net	2,423	260	6,592,919	< 5%
torrentz.eu	5,313	1,862	6,129,570	< 1%
torrenthound.com	4,221	1,078	5,855,589	< 50%
rapidlibrary.com	2,926	831	4,535,370	< 5%
filetram.com	3,173	665	4,402,138	< 5%
limetorrents.com	3,474	817	4,279,575	< 50%
mrtzcmp3.net	2,809	244	3,994,536	< 10%
beemp3.com	1,990	176	3,813,283	< 1%
downloads.nl	1,096	153	3,810,870	< 10%
bitsnoop.com	4,289	1,417	3,770,946	< 5%
bittorrent.am	3,182	755	3,677,249	< 50%
seedpeer.me	3,494	696	3,641,701	< 50%
muzofon.com	2,366	213	3,551,623	< 5%
myfreemp3.eu	3,058	311	3,499,941	< 5%

Table 6: Top 20 Specified Domains of DMCA Reported Complaints by Copyright Owners

as of 28/04/2015

(Source from: Google Transparency Report (2015))

The next procedure in this research methodology is that it is possible to obtain and retrieve more detailed domain information and various types of numerical and statistical reports about each of these top twenty specified domains or URLs as illustrated in Table 6 above by accessing and utilising a specific website statistics and website-worth valuation site such as triPHP (URL: http://spy.triphp.com/website-worth) or CuteStat.com (URL: http://www.cutestat.com/).

According to CuteStat.com (2015), the following main features and services regarding these prevalent website statistics and website valuation sites over the Internet (i.e. triPHP and CuteStat.com) are highlighted as below:

- This distinct web service is designed to provide a wide range of effective and useful information for various types of general Internet users including webmasters and owners of the sites. This particular web service is also designed to help particularly these users by retrieving information or data in regard to the following categories or technologies below:
  (1) Information in relation to Search Engine Optimisation (SEO) and Web Server
  (2) Information in relation to Internet Protocol (IP) Address and Domain Name
- This particular web service can be described as an effective tool that is basically designed to offer various web valuation outcome and statistical reports regarding any website that users normally have an interest to seek for. In this particular web service, the following information, services or statistical reports can also be provided to the users as below:
  (1) Providing Web Traffic Reports, (2) Providing Host Information, (3) Providing Social Network Engagement, (4) Providing Website Valuation for any website, (5) Providing information about page speed, (6) Providing reports about search engine, (7) Providing information about online safety, (8) Providing a Domain WHOIS and (9) There are also much more additional services that can be offered by this particular web service for web user through these websites statistics and website valuation sites such as CuteStat.com and triPHP as displayed in Figure 8 and Figure 9 respectively.

Figure 8 below shows a main screenshot about this particular website statistics and valuation service provider (i.e. CuteStat.com), which currently generates and provides various website statistics and website valuation outcome depending on the specific domain or URL.

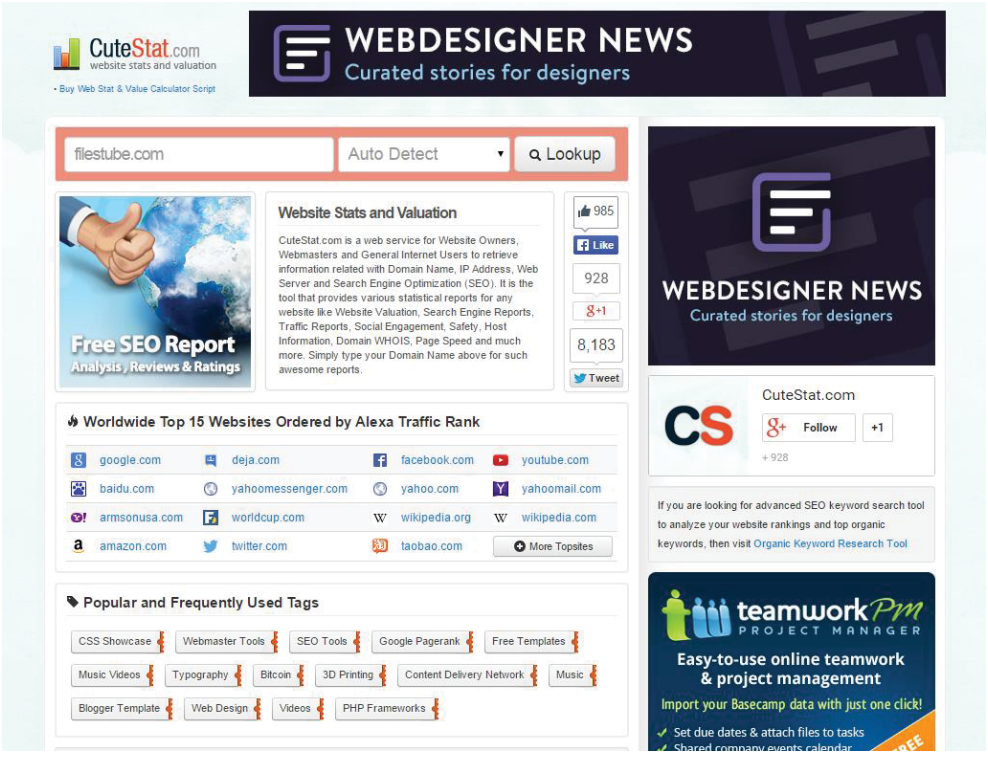


Figure 8: An Example of Website Statistics and Website Valuation from CuteStat.com (Source from: CuteStat.com (2015)).

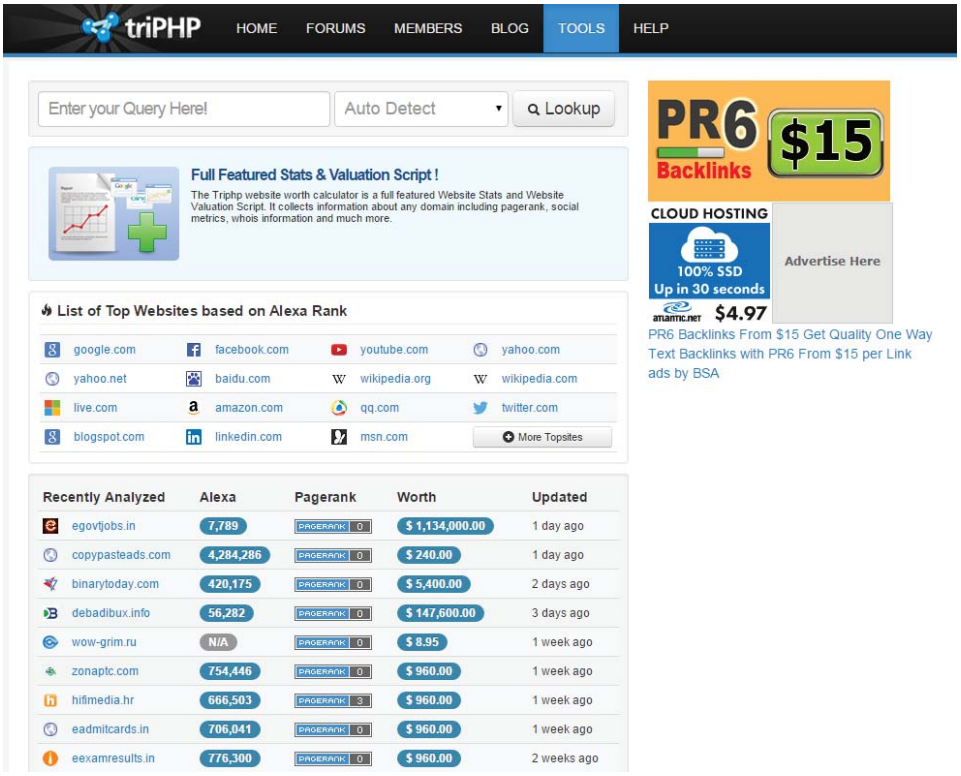


Figure 9: An Example of Website Statistics and Website Valuation from TriPHP

(Source from: Website Valuation – TriPHP (2015)).

In this research methodology, the name of website which is used for retrieving and obtaining detailed information in website statistics and website valuation is based on CuteStat.com (2015). Therefore, the primary raw data is basically obtained or collected from the result of website statistics and website valuation generated from CuteStat.com (2015) and this raw data will also be used and implemented for all the related-analysis and data comparison comprehensively in the entire thesis. The initial date of this data collection process from CuteStat.com was done on 11th of August in 2014. Hence, it is significant to note the fact that there are huge value differences in terms of specified domains ranking and overall numerical and categorical values in website statistics and website valuation due to the different timeframe that this research data observation for retrieving values from CuteStat.com (2015) was initially collected and implemented on 11th of August in 2014.

The next procedure in this research methodology is to obtain the actual Top Twenty specified domain statistics and domain valuation by typing and retrieving the specified domain URLs as previously displayed in Table 6 data (e.g. filestube.com or 4shared.com etc.) basically obtained from Google Transparency Report (2015). For instance, the users may be able to type the name of a specified domain (e.g. filestube.com) from these Top Twenty domains (i.e. Table 6) and then retrieve the corresponding domain statistics and its related-valuation information accordingly by looking up the specified URL as displayed in Figure 10 below.

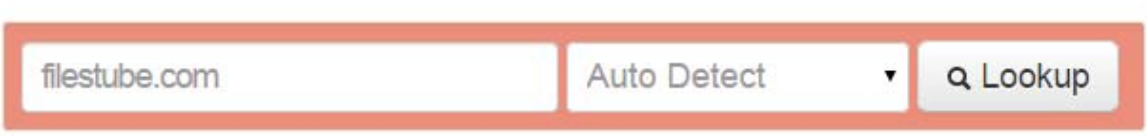


Figure 10: A Screenshot of Search Field with specified domain name (e.g. filestube.com)

(Source from: CuteStat.com (2015))

In Figure 11 below, the following comprehensive website statistical information and website valuation data regarding any specified domain (e.g. filestube.com) from Top Twenty domains can be generated and displayed every time when the user has retrieved the specified URL after clicking the Lookup button as displayed in Figure 10.

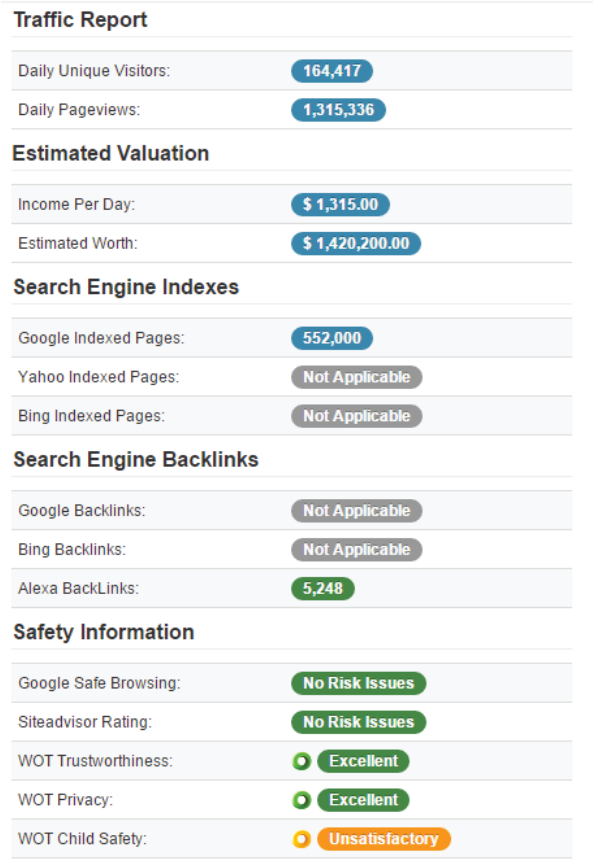


Figure 11: An Example Output of Domain Statistics and Domain Valuation retrieved from filestube.com (i.e. a specified domain) as of 28/04/2015.

(Source from: CuteStat.com (2015))

As displayed in Figure 11 above, the primary categories based on a specified domain (e.g. filestube.com) have been generated and classified into certain titles such as Estimated Valuation, Safety Information, Search Engine Indexes, Traffic Report and Search Engine Backlinks etc. In each main title, there are also a number of other related-variables associated with it. (e.g. income per day, estimated worth, daily pageviews, daily unique visitors, Google

indexed pages, Google backlinks, Alexa backlinks, Bing indexed pages and Yahoo indexed pages etc.).

The next procedure in this research methodology is to produce an Excel table or worksheet based on 20 instances (i.e. Top 20 Rogue Websites in rows) and 55 attributes (i.e. Entire Variables in columns) which have basically obtained or collected from the corresponding search results of website statistics and website valuation generated from CuteStat.com as of 11/08/2014. The following outcomes below from Table 7 to Table 16 respectively show the comprehensive raw data information for this research (i.e. 20 Instances (Top 20 Rogue Websites) and 55 Attributes (Entire Variables used)).

LABELS	INCOME_PER_DAY	ESTIMATED_WORTH	DAILY_UNIQUE_VISITORS	DAILY_PAGEVIEWS	GOOGLE_BACKLINKS
(1) filestube.com	$33,112.00	$35,760,960.00	4,139,050	33,112,400	88
(2) dilandau.eu	$3,741.00	$4,040,280	467,679	3,741,432	55
(3) rapidgator.net	$6,122.45	$8,147,981	2,040,817	27,913,864	Not Applicable
(4) 4shared.com	$16,853.93	$22,429,834	5,617,978	20,280,786	Not Applicable
(5) zippyshare.com	$9,404.39	$12,515,707	3,134,797	22,593,264	Not Applicable
(6) gosong.net	$788.00	$567,360.00	65,675	394,050	Not Applicable
(7) torrentz.eu	$50,486.00	$54,524,880.00	6,310,774	50,486,192	Not Applicable
(8) uploaded.net	$34,365.00	$37,114,200.00	4,295,569	34,364,552	Not Applicable
(9) rapidlibrary.com	$1,163.00	$1,256,040.00	145,385	1,163,080	Not Applicable
(10) torrenthound.com	$2,803.00	$3,027,240.00	350,358	2,802,864	51
(11) limetorrents.com	$497.00	$357,840.00	41,432	248,592	Not Applicable
(12) beemp3.com	$5.00	$1,200.00	762	1,524	Not Applicable
(13) filetram.com	$5,301.00	$5,725,080.00	662,570	5,300,560	14
(14) downloads.nl	$1,180.00	$1,274,400.00	147,461	1,179,688	14
(15) mrtzcmp3.net	$554.00	$398,880.00	46,184	277,104	6
(16) nakido.com	$299.00	$215,280.00	24,886	149,316	Not Applicable
(17) bitsnoop.com	$4,541.00	$4,904,280.00	567,654	4,541,232	64
(18) mp3juices.com	$933.00	$1,007,640.00	116,626	933,008	Not Applicable
(19) bittorrent.am	$533.00	$383,760.00	44,403	266,418	Not Applicable
(20) myfreemp3.eu	$1,044.00	$751,680.00	87,015	522,090	Not Applicable

Table 7: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes)

ALEXA_BACKLINKS	BING_BACKLINKS	GOOGLE_PAGERANK	ALEXA_RANK	DMOZ_LISTING	HOSTED_IP_ADDRESS	HOSTED_COUNTRY
13,655	5	6	247	No	78.140.188.239	Utrecht, The Netherlands
Not Applicable	32	4	2,186	No	176.31.230.116	France, EU
23,329	17	4	490	No	195.211.221.116	Russian Federation
138,292	Not Applicable	6	178	Yes	74.117.178.54	Virgin Islands, British
33,044	Not Applicable	5	362	No	46.105.114.15	France, EU
282	Not Applicable	3	14,651	No	108.162.195.17	United States
3,900	Not Applicable	6	162	No	68.71.55.19	Canada
25,849	Not Applicable	5	238	Yes	8.31.174.249	United States
2,133	Not Applicable	3	7,032	No	213.174.158.79	United States
Not Applicable	36	6	2,918	No	88.80.7.41	Sweden
Not Applicable	28	5	23,224	No	88.80.13.32	Sweden
2,932	Not Applicable	5	631,436	No	46.229.170.195	Netherlands
Not Applicable	25	4	1,543	No	212.124.114.226	United States
Not Applicable	31	4	6,933	No	213.206.91.18	Netherlands
Not Applicable	35	3	20,834	No	176.31.236.165	France, EU
1,358	Not Applicable	3	38,665	No	50.97.219.34	United States
1,805	Not Applicable	5	1,801	No	46.19.137.82	Switzerland
418	Not Applicable	4	8,766	No	85.17.23.6	Netherlands
671	Not Applicable	4	21,670	No	108.162.198.117	United States
326	Not Applicable	3	11,058	No	141.101.117.116	United States

Table 8: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

LOCATION_LATITUDE	LOCATION_LONGITUDE	FACEBOOK_SHARES	FACEBOOK_LIKES	FACEBOOK_COMMENTS	TWITTER_COUNT
52.090800000000	5.122220000000	12506	3105	3236	6
48.856700000000	2.350990000000	1	Not Applicable	Not Applicable	1
60.000000000000	100.000000000000	287	Not Applicable	Not Applicable	3
18.416700363159	-64.616699218750	169220	49896	36557	4282
50.694200000000	3.174560000000	8275	1703	3049	8192
38.949800000000	-77.227800000000	2637	594	726	2902
45.508800000000	-73.587800000000	39106	2670	8018	29410
38.895100000000	-77.036400000000	1715	271453	1808	24792
40.218300000000	-79.487800000000	1342	2364	256	1966
59.333000000000	18.050000000000	1	Not Applicable	Not Applicable	1
59.333000000000	18.050000000000	1	Not Applicable	Not Applicable	1
52.374000000000	4.889690000000	15039	3562	6520	19090
40.478100000000	-80.973100000000	1	Not Applicable	Not Applicable	1
52.350000000000	4.917000000000	1	Not Applicable	Not Applicable	1
48.856700000000	2.350990000000	1	Not Applicable	Not Applicable	1
32.796100000000	-96.802400000000	158	Not Applicable	Not Applicable	181
47.366700000000	8.550000000000	2069	273	631	3700
52.374000000000	4.889690000000	5179	5046	1417	9958
38.949800000000	-77.227800000000	95	12	27	10
34.052200000000	-118.244000000000	3310	4715	1008	6104

Table 9: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

LINKEDIN_SHARES	DELICIOUS_SHARES	GOOGLE_PLUS	DOMAIN_REGISTRAR	WEBSITE_OWNER
240	7062	47087	Instra Corporation Pty, Ltd.	Unknown
Not Applicable	Not Applicable	Not Applicable	EURID	Unknown
Not Applicable	Not Applicable	Not Applicable	INTERNET.BS CORP.	Unknown
1339	5843	57499	GODADDY.COM, LLC.	Alex Lunkov
63	240	182	TUCOWS DOMAINS INC.	Unknown
3	Not Applicable	26	NAME.COM, INC.	Unknown
97	2425	27672	EURID	Unknown
33	50	189	ENOM, INC.	Unknown
14	3725	466	PDR LTD. D/B/A PUBLICDOMAINREGISTRY.COM	Unknown
Not Applicable	Not Applicable	Not Applicable	WEB COMMERCE COMMUNICATIONS LIMITED DBA WEBNIC.CC	Unknown
Not Applicable	Not Applicable	Not Applicable	ENOM, INC.	Unknown
46	4181	674	INTERNET.BS CORP.	Unknown
Not Applicable	Not Applicable	Not Applicable	GODADDY.COM, INC.	Unknown
Not Applicable	Not Applicable	Not Applicable	Stichting Internet Domeinregistratie NL.	Unknown
Not Applicable	Not Applicable	Not Applicable	GANDI SAS	Unknown
Not Applicable	Not Applicable	Not Applicable	ENOM, INC.	Unknown
7	1329	197	EURODNS S.A	Unknown
2	Not Applicable	1822	INTERNET.BS CORP.	Unknown
Not Applicable	315	17	ISOCAM	Unknown
87	40	6435	EURID	Unknown

Table 10: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

OWNERS_EMAIL_ADDRESS	IP_ADDRESS_NUMBER_ONE	COUNTRY_ONE	IP_ADDRESS_NUMBER_TWO	COUNTRY_TWO
Unknown	149.13.65.167	United States	91.121.176.25	France
Unknown	176.31.230.116	France, EU	208.76.59.2	United States
Unknown	195.211.221.116	Russian Federation		
Unknown	74.117.178.54	Virgin Islands, British		
Unknown	46.105.114.15	France, EU	62.129.252.30	Poland
Unknown	173.245.58.129	United States	173.245.59.149	United States
Unknown	68.71.55.17	Canada	31.7.58.170	Switzerland
Unknown	204.13.251.21	United States	208.78.71.21	United States
qualifiedmarketing@googlemail.com	207.226.173.74	United States	88.208.58.214	Netherlands
Unknown	193.104.214.194	Sweden	88.80.30.194	Sweden
Unknown	67.212.93.181	Canada	67.212.93.181	Canada
Unknown	207.226.173.74	United States	88.208.58.214	Netherlands
Unknown	72.233.72.143	United States	174.36.237.98	United States
Unknown	212.204.192.252	Netherlands	212.204.207.192	Netherlands
Unknown	173.246.97.2	United States	217.70.184.40	France
Unknown	50.97.219.34	United States	208.53.183.232	United States
Unknown	80.92.65.2	Luxembourg	80.92.89.242	Luxembourg
Unknown	109.201.142.225	Netherlands	77.247.183.137	Netherlands
lion@freemail.fm	66.148.74.30	United States	66.235.184.104	United States
Unknown	173.245.58.112	United States	173.245.59.112	United States

Table 11: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

IP_ADDRESS_NUMBER_THREE	COUNTRY_THREE	IP_ADDRESS_NUMBER_FOUR	COUNTRY_FOUR	IP_ADDRESS_NUMBER_FIVE	COUNTRY_FIVE
75.126.229.92	United States				
208.76.56.76	United States	208.76.60.2	United States	208.76.58.2	United States
95.211.105.225	Netherlands	62.129.252.41	Poland		
85.195.102.27	Germany	31.7.58.114	Switzerland	94.242.253.62	Luxembourg
204.13.250.21	United States	208.78.70.21	United States		
62.250.7.3	Netherlands				
217.70.182.20	France				
80.92.95.42	Luxembourg	192.174.68.100	Austria		
108.61.12.163	United States				
85.17.19.180	Netherlands	91.121.10.227	France		

Table 12: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

IP_ADDRESS_NUMBER_SIX	COUNTRY_SIX	GOOGLE_SAFE_BROWSING	SITEADVISOR_RATING	WOT_TRUSTWORTHINESS	WOT_PRIVACY
		No Risk Issues	Minor Risk Issues	Excellent	Excellent
208.76.61.2	United States	No Risk Issues	Not Applicable	Very Poor	Very Poor
		No Risk Issues	No Risk Issues	Very Poor	Very Poor
		No Risk Issues	No Risk Issues	Excellent	Excellent
		No Risk Issues	Not Applicable	Excellent	Excellent
		No Risk Issues	Not Applicable	Excellent	Excellent
		No Risk Issues	Minor Risk Issues	Excellent	Excellent
		No Risk Issues	Not Applicable	Excellent	Excellent
		No Risk Issues	Minor Risk Issues	Good	Good
		No Risk Issues	No Risk Issues	Very Poor	Very Poor
		No Risk Issues	No Risk Issues	Very Poor	Very Poor
		No Risk Issues	Not Applicable	Excellent	Excellent
		No Risk Issues	No Risk Issues	Very Poor	Very Poor
		No Risk Issues	No Risk Issues	Very Poor	Very Poor
		No Risk Issues	No Risk Issues	Very Poor	Very Poor
		No Risk Issues	No Risk Issues	Very Poor	Very Poor
		No Risk Issues	No Risk Issues	Excellent	Excellent
		No Risk Issues	Minor Risk Issues	Good	Good
		No Risk Issues	Minor Risk Issues	Excellent	Excellent
		No Risk Issues	Not Applicable	Good	Good

Table 13: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

WOT_CHILD_SAFETY	GOOGLE_INDEXED_PAGES	YAHOO_INDEXED_PAGES	BING_INDEXED_PAGES	JAVASCRIPT_KILOBYTES	HTML_KILOBYTES
Unsatisfactory	78,900,000	42	2,790,000	783.1	76
Very Poor	8,980,000	467,000	42		
Very Poor	Not Applicable	51,400	5,260		
Excellent	Not Applicable	51,400	5,260	1007.1	92.6
Excellent	Not Applicable	Not Applicable	7	811.9	34.2
Excellent	Not Applicable	Not Applicable	Not Applicable	239.9	44.5
Excellent	Not Applicable	Not Applicable	Not Applicable	93.5	2.6
Excellent	Not Applicable	Not Applicable	Not Applicable	515.1	11.7
Poor	407,000,000	Not Applicable	Not Applicable	965.7	202.5
Very Poor	5,260,000	1,450,000	42		
Very Poor	2,770,000	141,000	38		
Excellent	Not Applicable	Not Applicable	Not Applicable	172.4	28.6
Very Poor	49,600,000	2,680,000	264,000		
Very Poor	13,900,000	548,000	42		
Very Poor	15,400,000	16,500	7		
Very Poor	6,410,000	553,000	40		
Excellent	63,100,000	Not Applicable	Not Applicable	119.6	5.5
Good	Not Applicable	Not Applicable	Not Applicable	158.1	6.5
Poor	16,400,000	Not Applicable	Not Applicable	19	118.9
Good	Not Applicable	Not Applicable	Not Applicable	1710	162.4

Table 14: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

IMAGES_KILOBYTES	CSS_KILOBYTES	KILOBYTES_TEXT	OTHER_KILOBYTES
38.9	7		0.1
240.1	149.1		72.8
20.4	96.8		
33.6			0.6
2.2	15.1		
144.8	39.2	88.8	0.1
134.6	26.1		
232.9	39.7		0.3
4.8	10.2		
122	6.8		0.2
202.3	4.4		0.4
42.6	83		208.3

Table 15: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

TOTAL_LINKS_OF_INTERNAL_LINKS	NO_FOLLOW_LINKS_OF_INTERNAL_LINKS	TOTAL_LINKS_OF_EXTERNAL_LINKS	NO_FOLLOW_LINKS_OF_EXTERNAL_LINKS
40	8	4	3
35	1	9	8
23	0	0	0
170	0	0	0
8	0	3	0
8	0	0	0
0	0	0	0
6	0	6	4
20	3	1	0
6	0	1	0
291	0	18	7
34	1	3	0

Table 16: The Comprehensive Raw Data retrieved and collected from CuteStat.com as of 11/08/2014 (20 Instances, 55 Attributes) continued.

Once the fundamental process of the entire raw data collection (i.e. Instances (Rows): 20, Attributes (Columns): 55) is fully done in an Excel-based worksheet retrieved and obtained from the corresponding search results of CuteStat.com (2015) as of 11/08/2014 as shown

above from Table 7 to Table 16 respectively, this basic raw data can be used to implement for the main analysis in this research, such as **Cluster Analysis**. The next section in this chapter (i.e. 3.3 (Cluster Analysis)) will be discussed in more detail regarding the key features and definitions of this particular analysis, basic methods of how these analysis techniques can be used and implemented and the description of the reason behind why this particular analysis technique is used for this research by utilising a simple k-means clustering analysis. (i.e. Cluster Analysis is basically used in this thesis).

3.3 Cluster Analysis

In this section of the thesis, the research methodology of Cluster Analysis will be discussed based on the raw dataset which has been collected on 11/08/2014. One of the main objectives in this research is to develop a robust approach to grouping together which presents the highest revenue risk to rightsholders. Cluster Analysis can be used to group together websites based on a wide range of attributes, such as income earned per day, estimated worth, daily unique visitors, etc. The attributes of high earning and low earning websites could also give some insight into policy options which might be effective in reducing earnings by pirate or rogue websites. For instance, are all low value websites based in a country with effective Internet controls? In this case, one of the data mining techniques such as a decision tree analysis could help the rightsholders to interpret these attributes. K-means is the most commonly used clustering algorithm in the field which it why it was selected.

Hypothesis:

As previously mentioned in the beginning of this thesis, the main hypothesis in this thesis is that there are basically two distinct groups of rogue websites which are most complained about (i.e. high value rogue websites and low value rogue websites). If hypothesis is true, then the appropriate law enforcement and rightsholders should only be targeting high value websites. Because these represent the greatest risk. In addition, the null hypothesis is that the most complained about websites are all high value rogue websites.

3.3.1 Theoretical Framework of Cluster Analysis

In this particular section of the thesis – 3.3.1, the comprehensive overview about the theoretical frameworks and definitions of clusters and clustering based on various related-studies and publications will first be described. The brief introduction of a clustering algorithm used in this thesis will also be discussed (i.e. a simple K-means algorithm). Furthermore, the description of Cluster Analysis in a general context will be discussed. The following particular section (i.e. section 3.3.3) will also show the fundamental key procedures of Cluster Analysis based on the raw data as shown previously from Table 7 to Table 16 by using the practical implementation of the following application: Waikato Environment for Knowledge Analysis (WEKA) 3.6. In the main contents of this practical demonstration, cluster analysis based on a simple K-means algorithm will be used and implemented in WEKA 3.6 Explorer data mining environment.

According to Witten, Frank & Hall (2011), the following fundamental features are significantly indicated about clustering in general:

- The utilisation of clustering method can be used and applied when no class is existed only for the purpose of using prediction. However, this study from Witten, Frank & Hall (2011) indicated that the instances can be classified into distinct or natural groups in general.
- This particular study suggested that these clusters would seem to apply or probably generate a significant reflection on some mechanism which is currently valid to operate in the domain where the instances are basically induced. This study from Witten, Frank & Hall (2011) indicated that there is also a mechanism which is primarily designed to influence some instances to bring a higher similarity between these particular instances except for their response to the rest of the instances.

Furthermore, Witten, Frank & Hall (2011) also indicated that one of the main clustering methods is an iterative distance-based clustering technique such as a simple K-means clustering algorithm. The main features of this particular clustering technique or method are highlighted comprehensively by Witten, Frank & Hall (2011) as follows:

- This study indicated that the conventional clustering method is widely well known as K-means clustering. The first significant process is to determine regarding how many clusters are basically needed to be sought beforehand: This process showed that this refers to the value of the parameter which is indicated as K. After the value of K is determined or assigned, the next process is that the random selection of K points was implemented on the basis of centres for cluster. This study from Witten, Frank & Hall (2011) also indicated that all the instances can be determined to the position where they are most closely located around the cluster centre in accordance with the primary theory of Euclidean distance metric.
- This study from Witten, Frank & Hall (2011) indicated that the cluster's centroid from the corresponding instances in regard to each clusters is computed. In addition, this

study indicated that the values of these centroids from the instances are considered as a new mean or centroid value for their specific clusters.

- Witten, Frank & Hall (2011) highlighted that the entire process in regard to K-means clustering algorithm is iterated along with new centre values for their specific clusters. Witten, Frank & Hall (2011) also indicated that repetition basically proceeds until the equivalent or equal points are allocated to each cluster in a series of rounds. This particular study implied that the centres of the cluster have well-maintained and this distinct iteration process from K-means clustering technique will continue to remain continually.
- This study from Witten, Frank & Hall (2011) suggested that utilisation of using a clustering technique such as K-means clustering analysis can be comprehensible and effective method. This study also suggested that the value of the overall squared distance between the cluster's each distinct point and its centres can be minimised effectively by selecting the relevant cluster centre in regard to becoming the centroid. Once the repetition process has well-maintained or stabilised based on the foundation of K-means clustering, Witten, Frank & Hall (2011) indicated that each distinct point can be allocated to the closest location of cluster centre. Therefore, this study implied that the overall squared distance from all points to cluster centres can be achieved to obtain the following significant effect: the minimisation of the overall squared distance between them.

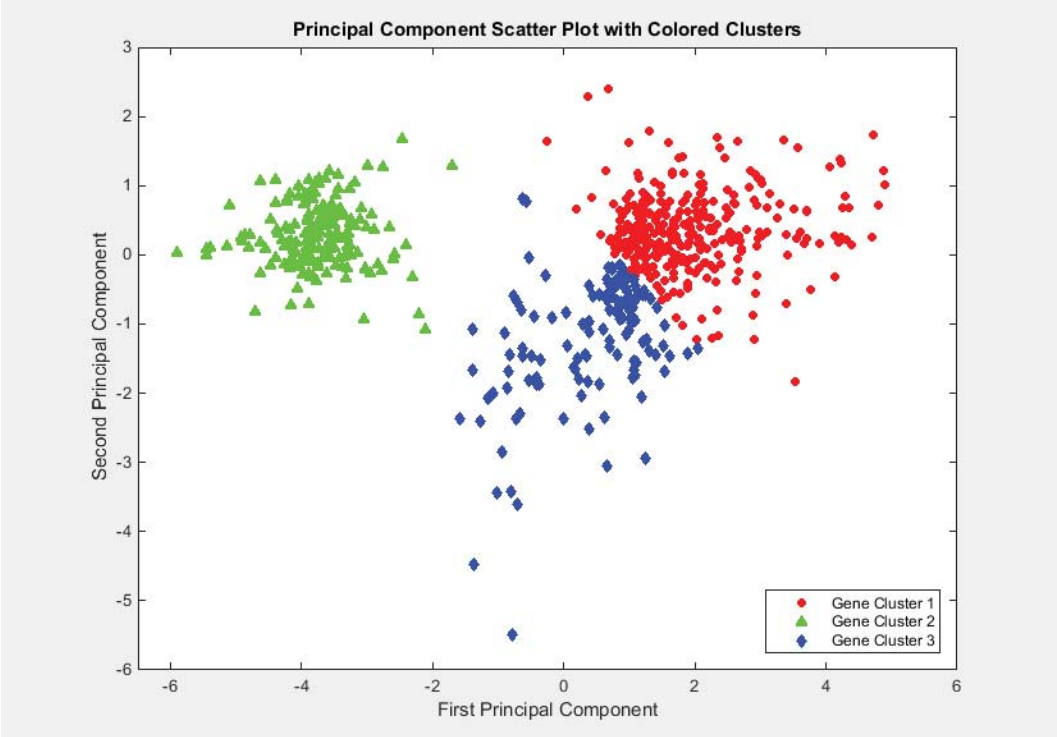


Figure 12: An Example of General Clustering

(Source from: MathWorks (2015) – au.mathworks.com)

According to Suh (2012), the following theoretical properties of clusters and clustering are also indicated in general as below:

- In this study, Suh (2012) indicated that the meaning of clustering can be defined as the method of categorising extensive group of data components into smaller groups which primarily interchange the equivalent or identical characteristics.
- This particular study from Suh (2012) indicated that the various utilisation of clustering has been broadly used and demonstrated as an effective cluster analysis method in the case of handling with enormous amounts of data, due to the fact that this method is very useful and efficient in terms of both (1) data simplification and (2) data classification. Hence, this study from Suh (2012) implied that the basic interaction between cluster analysis and clustering is appeared to be compatible.
- This study suggested that various types of clustering algorithms have been utilised and evolved for the purpose of mainly enhancing the effectiveness of both data classification and data simplification. In addition, Suh (2012) indicated that clustering analysis is a prevalent method based on several purposes that is primarily designed to perform the analysis of numerous amounts of datasets in many different areas or fields.
- This study from Suh (2012) indicated that clustering can be considered as a fundamental element of classification from originally unclassified information or data on the basis of common characteristics. Moreover, this particular study suggested that it is significant to comprehend the diverse aspects of the clusters that can be very helpful to comprehend the essential algorithms (e.g. a simple K-means clustering algorithm) used in cluster analysis in detail.
- In this study, Suh (2012) implied that there are largely two significant points in regard to clustering analysis as follows:
  (1) One of the significant points is known as similarity. It can be reflected in the measured distance between the two specific points.
  (2) The second significant point is known as classification. It presents the aim of clustering.
  Hence, this study from Suh (2012) on the basis of these two significant points above indicated that clustering is a procedure to examine or inspect a group of data in accordance with any similar properties and also to establish classification between groups.
- Finally, this study showed that the primary aim of clustering is to mainly examine groups from data that enables to satisfy one of these two requirements:
  (1) Similarity is essentially applied to the members in specific group.
  (2) The classification of groups should be clearly distinguished between groups.

Furthermore, a related-information from Everitt (1980), (as cited in Suh (2012)) highlighted that the following theoretical descriptions are involved in relation to a cluster as below:

Everitt (1980) specified that a cluster is a group of similar objects or a group of not similar objects from different clusters. In addition, this particular study from Everitt (1980) indicated that a cluster can be defined as a combination from specific points which implies that extent between any two particular points within the cluster is less than the extent between any internal point within cluster and any external point outside cluster. This study also indicated that a cluster is fundamentally equivalent to the connected area or segment from full dimensional space comprising of highly compact points. In addition, this study from Everitt (1980) suggested that a cluster is a group of adjacent components associated with statistical population.

3.3.2 A Simple K-means Clustering Algorithm

This particular section introduces and describes about the theoretical framework of a simple K-means clustering algorithm which is primarily used as a key cluster analysis technique for research in this thesis.

In this research, one of the most commonly used data mining techniques like Cluster Analysis can be used to group together websites based on a wide range of attributes or variables (e.g. income earned per day, estimated worth, daily unique visitors and daily page-views etc.) in order to mainly identify groups of data that have the two certain patterns (i.e. similarity-within criterion and separation-between criterion etc.) as indicated by Suh (2012). This research will be focused on the implementation of a commonly used clustering algorithm (i.e. a simple K-means clustering is used) based on the raw data collected in regards to Top Twenty rogue websites retrieved from Google Transparency Report (2015). According to Ahlemeyer-Stubbe & Coleman (2014), this study indicated that the foundation of K-means clustering algorithm is primarily designed to create or construct clusters based on the clusters of K seed initially. This study from Ahlemeyer-Stubbe & Coleman (2014) also specified that the basic distance measurement of this clustering method (i.e. cluster analysis) is on the basis of the Euclidean distance which can be used to find the differences between mean values from each variable for the purpose of input. In addition, Kantarzic (2011) indicated that K-means clustering algorithm is a method of the most commonly used clustering analysis associated with the following two features: (1) The implementation of clustering is simple and easy and (2) It is relatively small clustering in terms of both time and space complexity.

Furthermore, the related-study from Johnson & Wichern (1998), (as cited in Olson & Shi (2007)) has identified that the following six main consecutive steps are highlighted about the key procedures of a simple K-means clustering algorithm as below:

Step One: Selecting the preferable number of a given cluster k (i.e. iteration can be implemented from two to the maximum number of preferable clusters).

Step Two: Selecting k which refers to observed values initially as given seed values (i.e. this may be randomly in this step but if these given values as seed were basically as far apart as possible, the clustering algorithm would be able to perform better).

Step Three: The computation of average cluster values can be implemented on each distinct variable (In the case of initial repetition or iteration process, this will clearly be indicated as initial observed seed values).

Step Four: Each of the other training observed values (i.e. observations) can be allocated to the closest cluster, as computed by squared distance (i.e. In this particular step, the metric of using squared distance is conventional method. However, other types of relevant metrics can possibly be used in regard to this step).

Step Five: The recalculation process of cluster averages is implemented on the basis of the assignments or arrangements from Step Four as above.

Step Six: Iteration process is needed between Step Four and Step Five above until the identical set of arrangements or assignments are attained.

In addition, this recent study from Ahlemeyer-Stubbe & Coleman (2014) highlighted that K-means clustering can be described as a very rapid technique or method that is primarily designed to handle or cope with enormous numbers of both cases and input variables. Ahlemeyer-Stubbe & Coleman (2014) also suggested that this K-means clustering technique is faster than hierarchical clustering method. Because this study from Ahlemeyer-Stubbe & Coleman (2014) indicated that the number of core comparisons required to determine is smaller than other clustering methods in regard to identifying which groups to join or collaborate together.

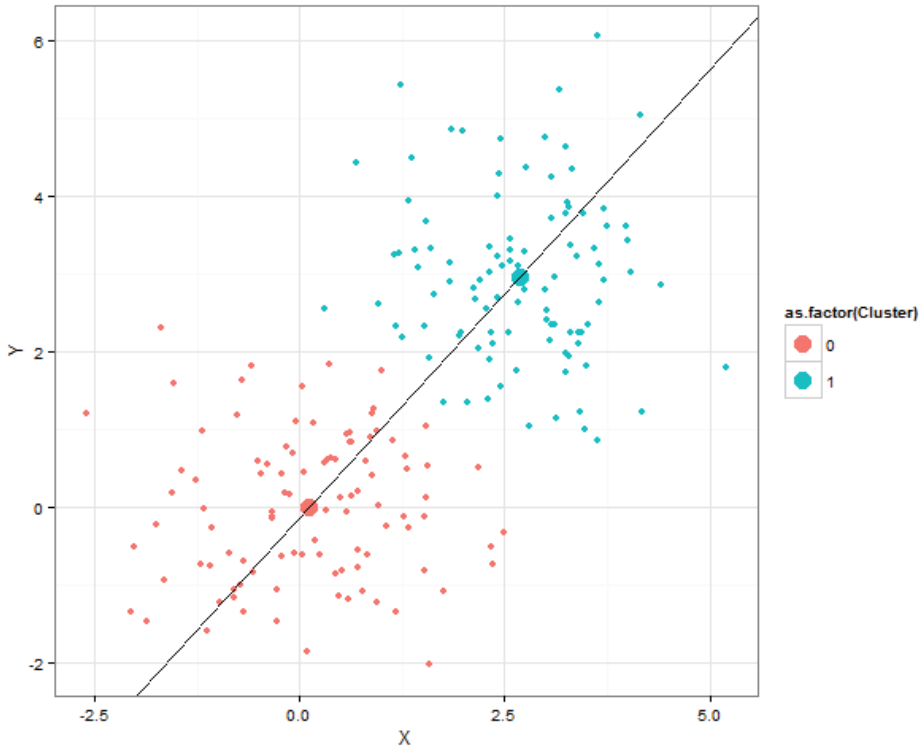


Figure 13: An Example of a Simple K-means clustering (Source from: StackExchange (2015) – stats.stackexchange.com)

3.3.3 Project Implementation by using K-means Clustering Analysis

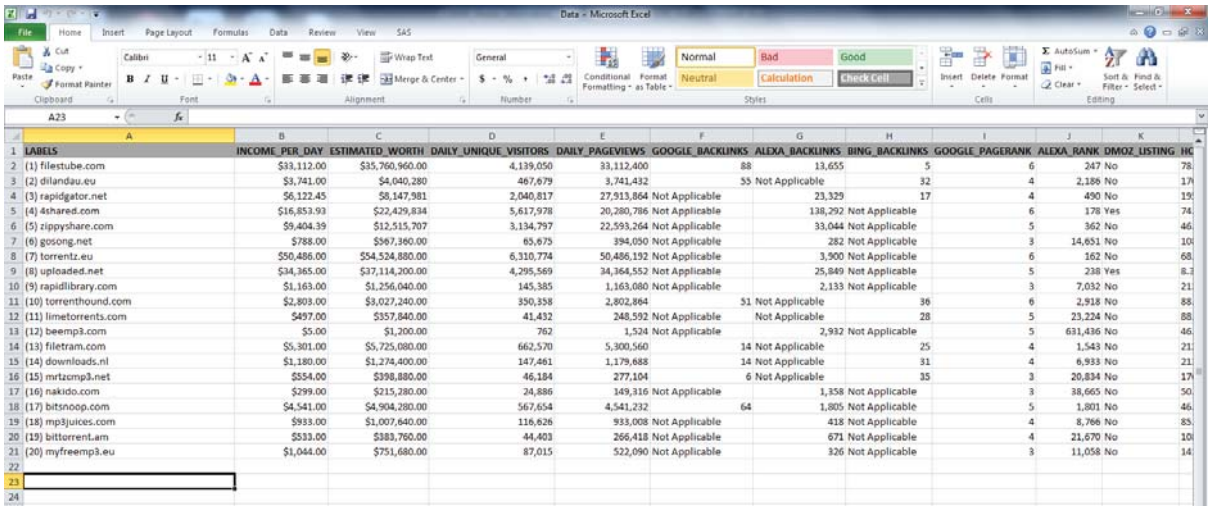
This particular section introduces and describes about the project implementation by using K-means clustering analysis for research used in this thesis as below:

The next procedure will demonstrate regarding how the K-means Cluster Analysis can be used and implemented from the raw data of this research as of 11/08/2014 (i.e. from Table 7 to Table 16) which has been initially retrieved from the corresponding website valuation results of CuteStat.com (2015) based on the Top 20 rogue website details of Google Transparency Report (2015). In this particular procedure, the implementation of WEKA 3.6 is used for this raw data as of 11/08/2014 (i.e. from Table 7 to Table 16).

According to the Department of Computer Science of The University of Waikato (2008), it indicates that the following significant features regarding an ARFF (Attribute-Relation File Format (ARFF) file are highlighted for utilisation with the WEKA 3.6 machine learning application as below:

- This study indicated that an ARFF is a file that stands for American Standard Code for Information Interchange (ASCII) text file that is primarily designed to interpret or explain a list of basic instances which have the features of sharing based on a set of certain attributes.
- This study also showed that the foundation of this particular ARFF files was originally created and evolved by the development of Machine Learning Project at the Department of Computer Science of The University of Waikato (2008) particular for effective data mining utilisation and implementation by using WEKA 3.6 machine learning software.
- This information highlighted that each unique instance can be basically expressed or indicated in terms of a single line form, indicating the end of the instance in association with carriage returns.
- This study showed that in the case of the instance data, **missing values** are primarily indicated or represented by applying form of **a single question mark**.
- This study also highlighted that it is case-sensitive in the case of both nominal values and string values and if there are any that include space, and then it should be quoted.
- Furthermore, this study emphasised that dates should be indicated in the data segment using the string-based description defined in the declaration of attribute.

Therefore, to import and run the basic raw data (i.e. from Table 7 to Table 16) successfully in WEKA 3.6 environment, first it is one of the core requirements that all the not-applicable (NA) or missing values contained in any Excel cell are initially necessary to be fully replaced by inserting a single question mark in an Microsoft Excel worksheet format. In the basic raw data as of 11/08/2014 (i.e. Data.xlsx – as seen previously from Table 7 to Table 16) for this research as displayed in Figure 14 below, all the unknown values or inputs in any Excel cell are also necessary to be replaced by entering a single question mark instead.



LABELS	INCOME_PER_DAY	ESTIMATED_WORTH	DAILY_UNIQUE_VISITORS	DAILY_PAGEVIEWS	GOOGLE_BACKLINKS	ALEXA_BACKLINKS	BING_BACKLINKS	GOOGLE_PAGERANK	ALEXA_RANK	DMOZ_LISTING
(1) filestube.com	$33,112.00	$35,760,960.00	4,139,050	33,112,400	88	13,655	5	6	247	No
(2) dilandau.eu	$3,741.00	$4,040,280	467,679	3,741,432	55	Not Applicable	32	4	2,186	No
(3) rapidgator.net	$6,122.45	$8,147,981	2,040,817	27,913,864	Not Applicable	23,329	17	4	490	No
(4) 4shared.com	$16,853.93	$22,429,834	5,617,978	20,280,786	Not Applicable	138,292	Not Applicable	6	178	Yes
(5) zippyshare.com	$9,404.39	$12,515,707	3,134,797	22,593,264	Not Applicable	33,044	Not Applicable	5	362	No
(6) gosong.net	$788.00	$967,360.00	65,675	394,050	Not Applicable	282	Not Applicable	3	14,651	No
(7) torrentz.eu	$50,486.00	$54,524,880.00	6,310,774	50,486,192	Not Applicable	3,900	Not Applicable	6	162	No
(8) uploaded.net	$34,365.00	$37,114,200.00	4,295,569	34,364,552	Not Applicable	25,849	Not Applicable	5	238	Yes
(9) rapidlibrary.com	$1,163.00	$1,256,040.00	145,385	1,163,080	Not Applicable	2,133	Not Applicable	3	7,032	No
(10) torrenthound.com	$2,803.00	$3,027,240.00	350,358	2,802,864	51	Not Applicable	36	6	2,918	No
(11) limetorrents.com	$497.00	$357,840.00	41,432	248,592	Not Applicable	Not Applicable	28	5	23,224	No
(12) beemp3.com	$5.00	$1,200.00	762	1,524	Not Applicable	2,932	Not Applicable	5	631,436	No
(13) filetram.com	$5,301.00	$5,725,080.00	662,570	5,300,560	14	Not Applicable	25	4	1,543	No
(14) downloads.nl	$1,180.00	$1,274,400.00	147,461	1,179,688	14	Not Applicable	31	4	6,933	No
(15) mrtzmp3.net	$554.00	$598,880.00	46,184	277,104	6	Not Applicable	35	3	20,834	No
(16) nakido.com	$299.00	$215,280.00	24,886	149,316	Not Applicable	1,358	Not Applicable	3	38,665	No
(17) bitsnoop.com	$4,541.00	$4,904,280.00	567,654	4,541,232	64	1,805	Not Applicable	5	1,801	No
(18) mp3juices.com	$933.00	$1,007,640.00	116,626	933,008	Not Applicable	418	Not Applicable	4	8,766	No
(19) bittorrent.am	$533.00	$383,760.00	44,403	266,418	Not Applicable	671	Not Applicable	4	21,670	No
(20) myfreemp3.eu	$1,044.00	$751,680.00	87,015	522,090	Not Applicable	326	Not Applicable	3	11,058	No

Figure 14: A Screenshot of the basic raw data file in a Microsoft Excel worksheet

(i.e. Data.xlsx)

As previously mentioned, all the not-applicable (NA) or missing values as well as any unknown input values (i.e. unknown) can be replaced all by entering a single question mark and choosing the option "replace all" as follows:

Step One: First choose the Find and Replace window by entering the following buttons (CTRL + F) in the raw data file (i.e. Data.xlsx) as shown in Figure 15 below.

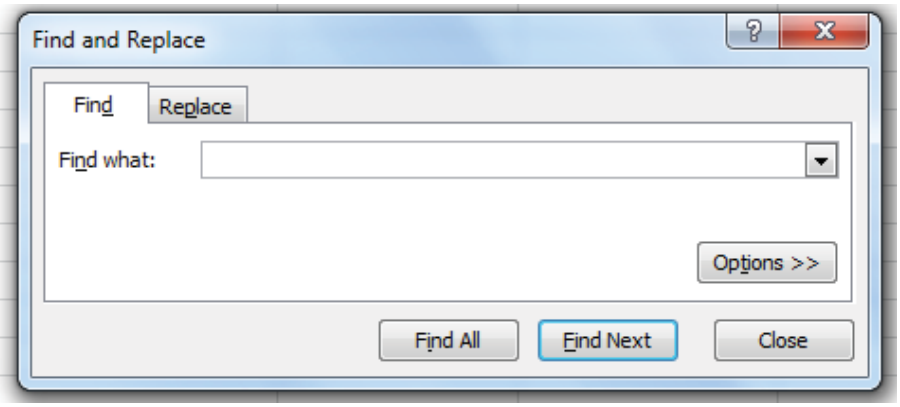


Figure 15: A Screenshot of Find and Replace window in Data.xlsx (Raw data file)

Step Two: As illustrated in Figure 16 below, Select the Replace menu in Find and Replace window and then type in "Not Applicable" in the text field – *Find what* and also type in a single question mark as "?" in the text field – *Replace with*. Once all the input details such as "Not Applicable" and a single question mark "?" are entered, the next step is to click the button "Replace All".

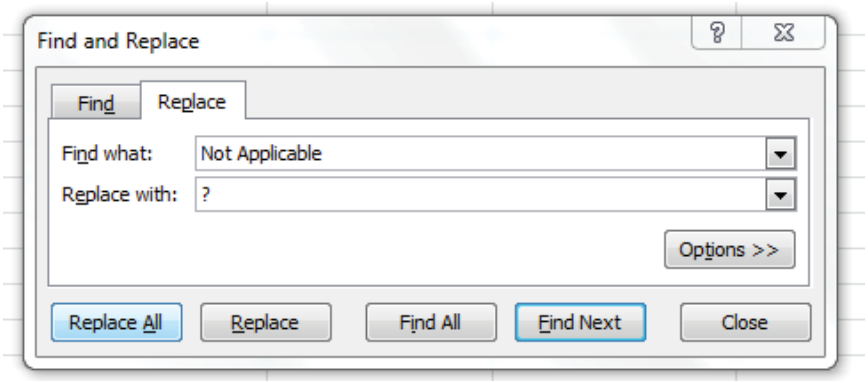


Figure 16: A Screenshot of Replace menu containing the two required input text fields

(i.e. *Find what* (Not Applicable) and *Replace with* (a single question mark)).

Step Three: As illustrated in Figure 17 below, Select the Replace menu in Find and Replace window and then type in "Unknown" in the text field – *Find what* and also type in a single question mark as "?" in the text field – *Replace with*. Once all the input details such as "Unknown" and a single question mark "?" are entered, the next step is to click the button "Replace All".

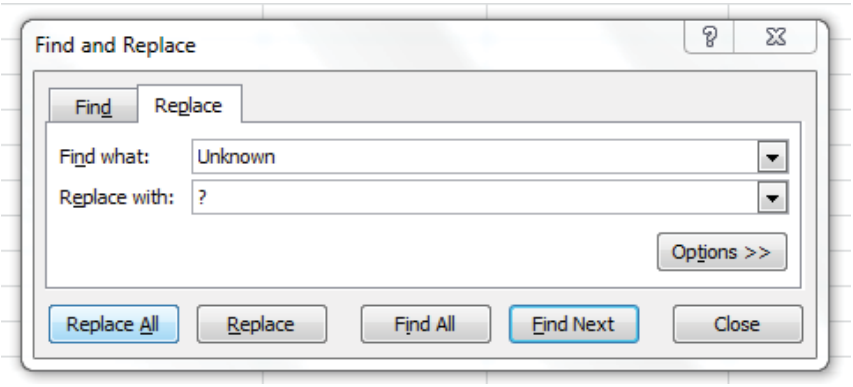


Figure 17: A Screenshot of Replace menu containing the two required input text fields

(i.e. *Find what* (Unknown) and *Replace with* (a single question mark)).

Step Four: Within a Microsoft Excel worksheet file (Data.xlsx), the next step is to choose the following options in order to save the basic raw data file as a Comma Separated Values (CSV) file (i.e. Data.csv). Because, a CSV is only a valid type of imported file format available in WEKA 3.6 Explorer Environment. Therefore, it is important to save the raw data file as "Data.csv" from the existing file name "Data.xlsx" as an Excel spreadsheet format in an Microsoft Excel worksheet environment. As illustrated in Figure 18 below, it is also needed to save the file as the following type – CSV (Comma Delimited).

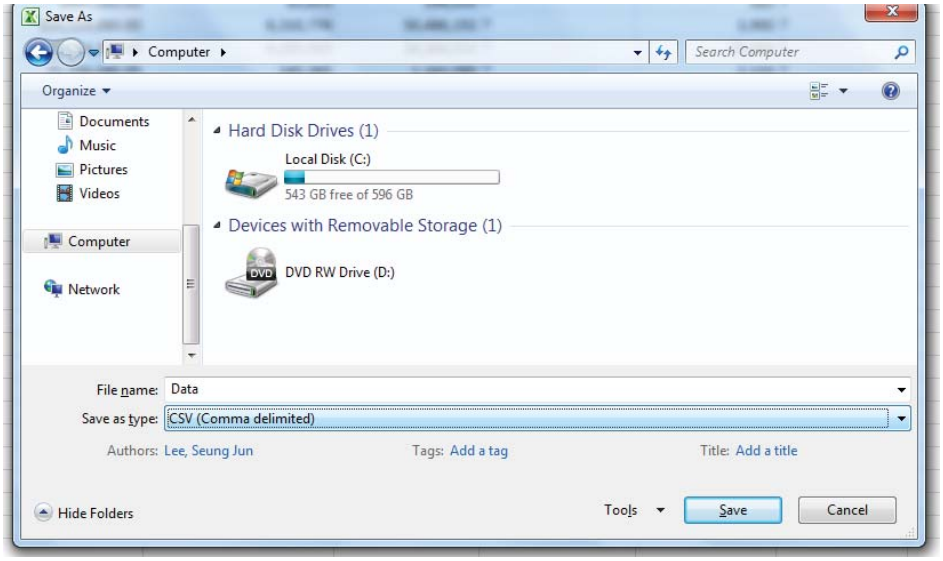


Figure 18: A Screenshot of saving the basic raw data file as CSV (Comma Delimited) type in an Microsoft Excel worksheet environment (i.e. Data.csv)

Step Five: The next procedure is to import this CSV file from WEKA 3.6 Explorer Environment in order to use and implement the Cluster Analysis (i.e. a simple K-means clustering algorithm). To do this, the following additional steps are basically undertaken respectively:

(1) Click the Windows menu at the bottom of the main screen

(2) Choose the All Programs

(3) Choose the program folder called – Weka 3.6.12

(4) Choose the executable program – Weka 3.6 and then the following Graphical User Interface (GUI) Chooser of WEKA 3.6 data mining program will be displayed as shown in Figure 19 below:

Figure 19: A Screenshot of WEKA 3.6 Data Mining Program GUI Chooser

(Source from: Department of Computer Science of The University of Waikato (2014)).

(5) In WEKA 3.6 GUI Chooser, the next procedure is to select the option called "Explorer" in Applications menu.

(6) In WEKA 3.6 Explorer, select the button called "Open file" in the "Preprocess" menu. Then browse the raw CSV data file which has been saved as "Data.csv" from a Microsoft Excel worksheet as previously described. Figure 20 below displays an illustration of opening the main data file "Data.csv" in WEKA 3.6 Explorer environment.

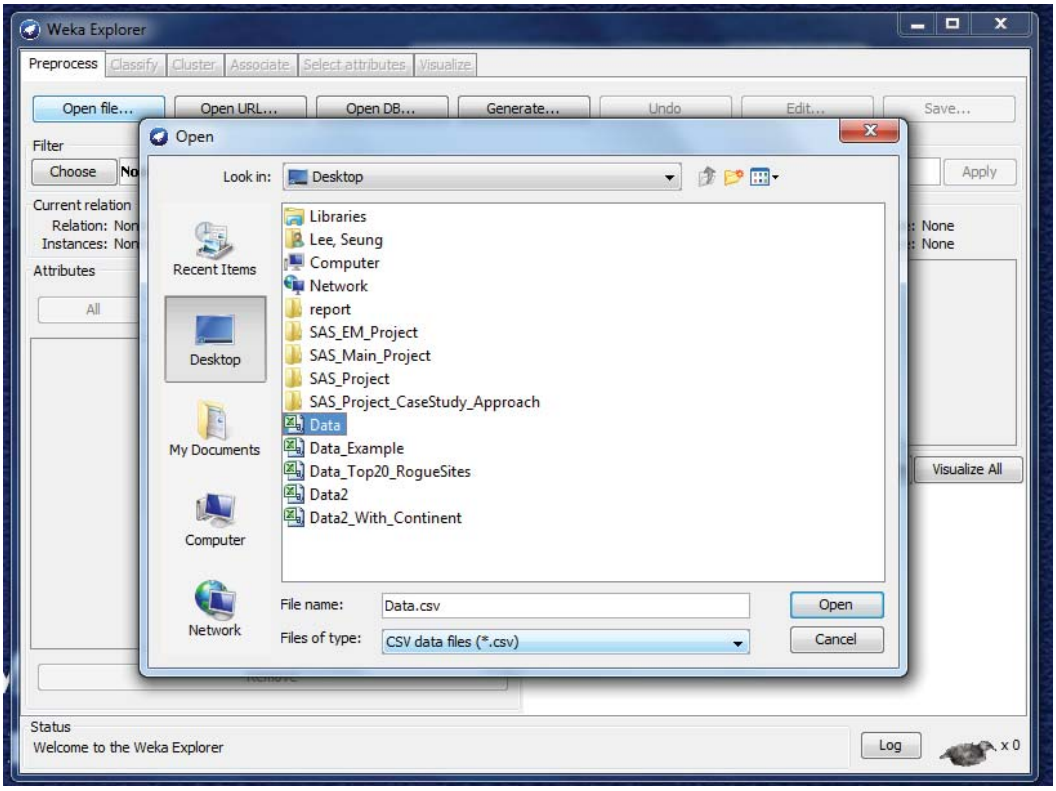


Figure 20: A Screenshot of opening the main data file "Data.csv" in WEKA 3.6 Explorer.

(7) The following GUI as shown in Figure 21 below will be displayed with 20 instances (i.e. rows) and 55 attributes (i.e. columns).

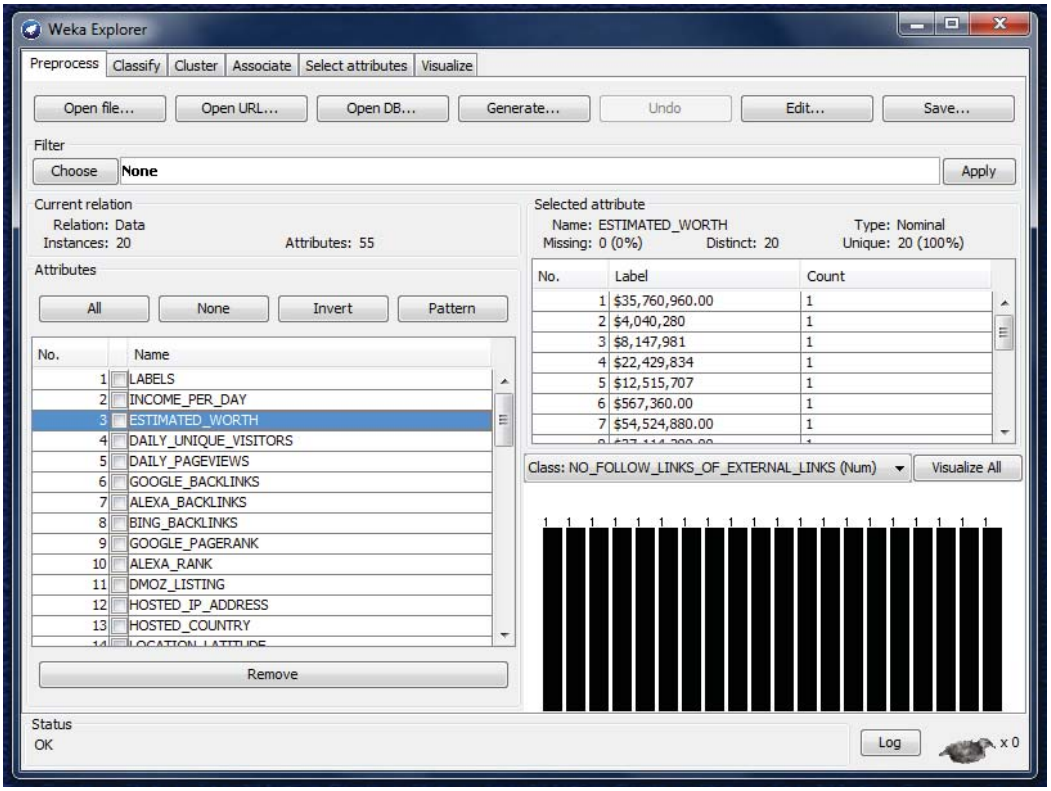


Figure 21: A Screenshot of WEKA 3.6 Explorer after opening the main data file (Data.csv).

(8) The next procedure is to set up the feature of implementing the *ReplaceMissingValues*. According to Witten, Frank & Hall (2011), this particular study indicated that *ReplaceMissingValues* (i.e. unsupervised attribute filter) is primarily designed to substitute for all missing values for both numeric and nominal attributes based on the mode for nominal attributes and the mean of numeric attributes.

Witten, Frank & Hall (2011) also emphasised in *ReplaceMissingValues* that in the case of a class is set, missing values in regard to that particular attribute are not basically substituted for attributes by default. However, this can be altered or varied.

This particular function can be used by implementing the following few steps respectively: (a) Click the button called choose in the Filter menu, (b) Open the folder called *filters* under the upper folder called *weka*, (c) Open the sub-folder called *unsupervised* under the upper folder – *filters*, (d) Open the sub-folder called *attribute* under the upper folder – *unsupervised* and (e) Select the required function (i.e. *ReplaceMissingValues*) within the folder – *attribute*. Figure 23 as displayed below basically shows an entire illustration of how these steps have been achieved in order to implement one of the filtering algorithms such as *ReplaceMissingValues* etc. Figure 22 as illustrated below also shows that there is now an Unsupervised Attribute Filters function available (i.e. ReplaceMissingValues) after implementing and applying this particular function successfully. To apply this filtering algorithm (i.e. ReplaceMissingValues) properly on the current data file (Data.csv), it is also important to ensure that the button called "Apply" should be selected in Figure 22.

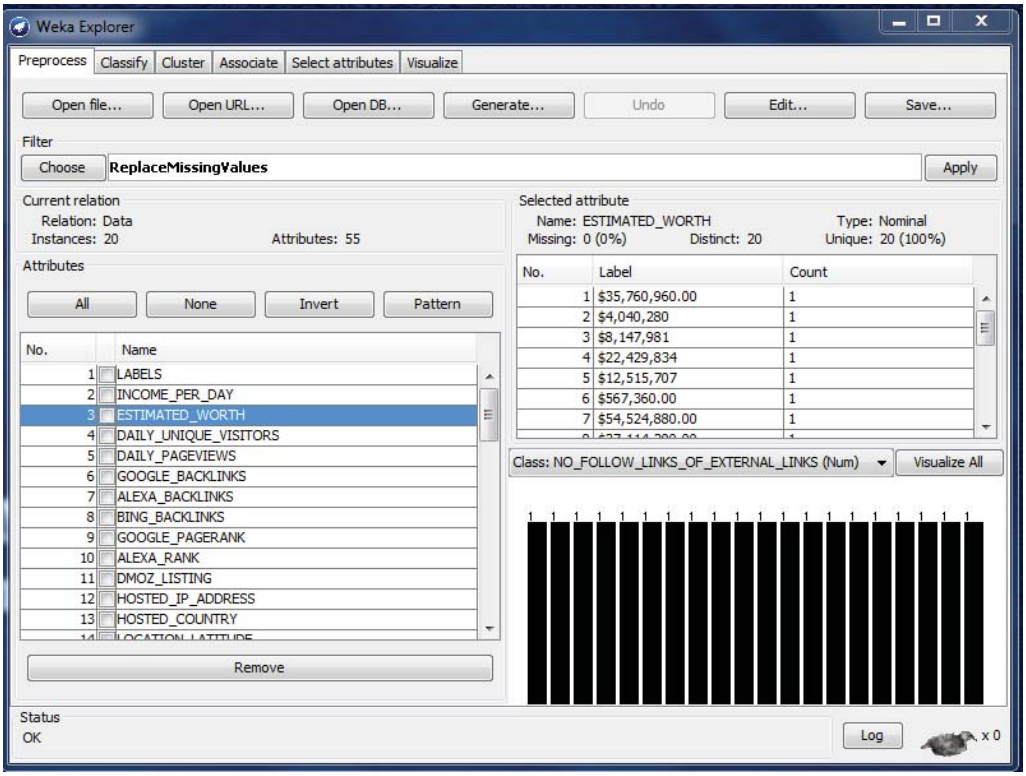


Figure 22: A Screenshot of WEKA 3.6 filtering algorithm (i.e. *ReplaceMissingValues*) after implementing and applying this particular function successfully.

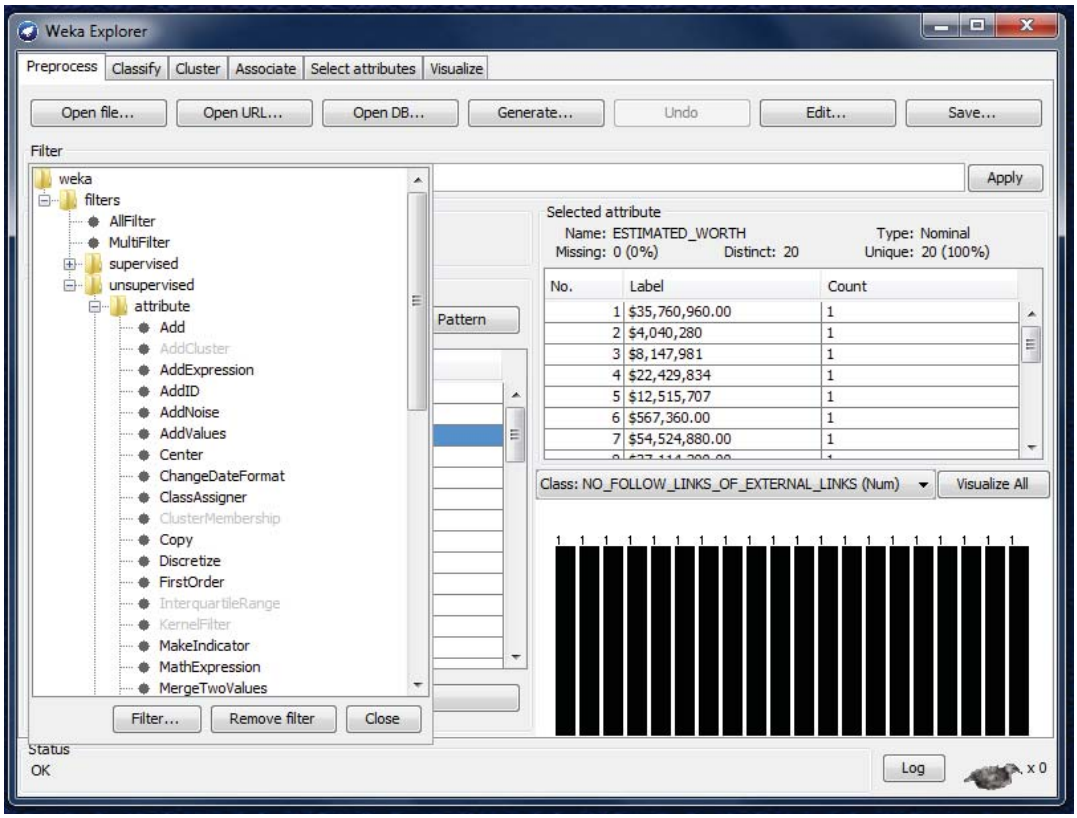


Figure 23: A Screenshot of WEKA 3.6 filtering algorithm in WEKA Explorer

(i.e. Unsupervised Attribute Filters – *ReplaceMissingValues*).

(9) The next procedure is to implement a Cluster Analysis (i.e. a simple K-means clustering method in WEKA 3.6 Explorer) by choosing the following steps as shown below:

(a) In WEKA 3.6 Explorer, the first step is to select one of the main menus called "Cluster" and then click the button called "Choose" in the Clusterer section.

(b) The next step is to select a clusterer type called "SimpleKMeans". The function of this particular clusterer is to cluster data and implement cluster analysis by using a simple K-means clustering algorithm.

According to the description of WEKA 3.6 Explorer (2014) regarding this clustering algorithm, this application showed that simple K-means clustering can be used and classified into the following two distinct types: (1) the Euclidean distance by default and (2) the Manhattan distance. The description of WEKA 3.6 Explorer (2014) also indicated that if the Manhattan distance is selected and used, then the computation of centroids is basically implemented in terms of the component-wise median rather than mean. Figure 24 below shows a basic Cluster UI and it can be seen in Figure 24 that a simple K-means clustering algorithm for this research is selected in WEKA 3.6 Explorer (2014) environment as below.

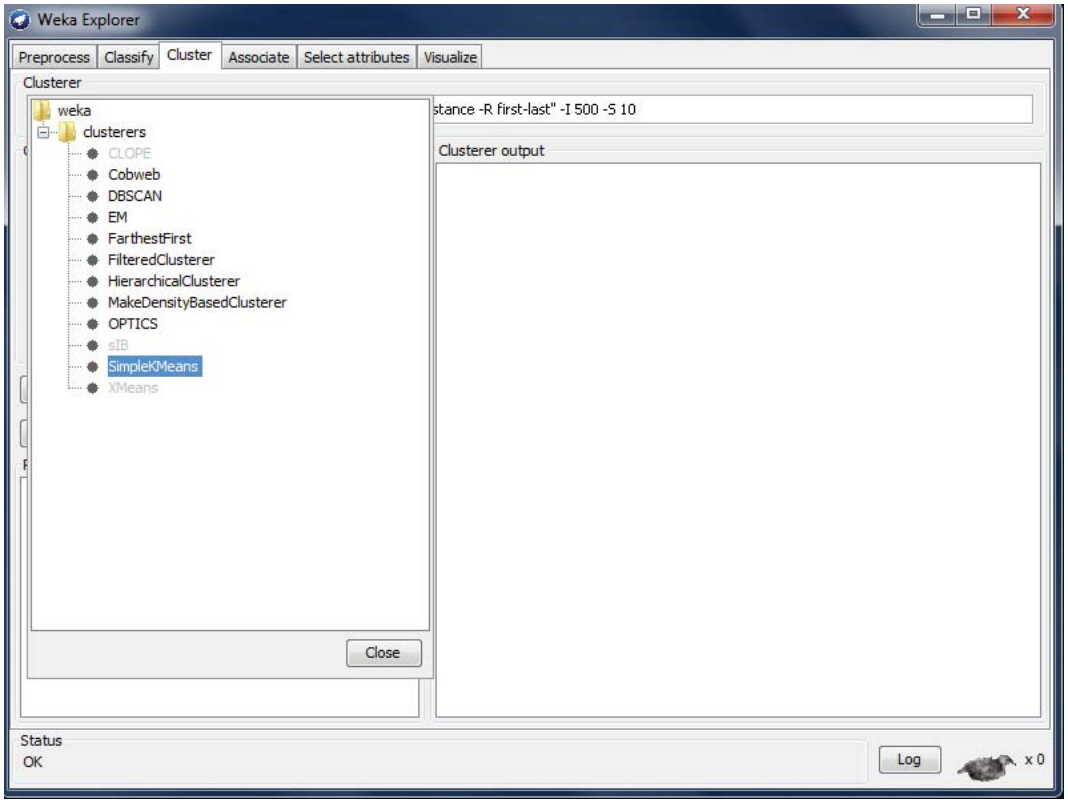


Figure 24: A Screenshot of selecting a simple K-means clustering algorithm in WEKA 3.6 Explorer.

(c) The next step is to select an executable button called "Start" using the simple K-means algorithm based on the primary data file (i.e. Data.csv). Figure 25 as shown below displays an illustration of the clustering output after implementing the simple K-means clustering algorithm successfully.

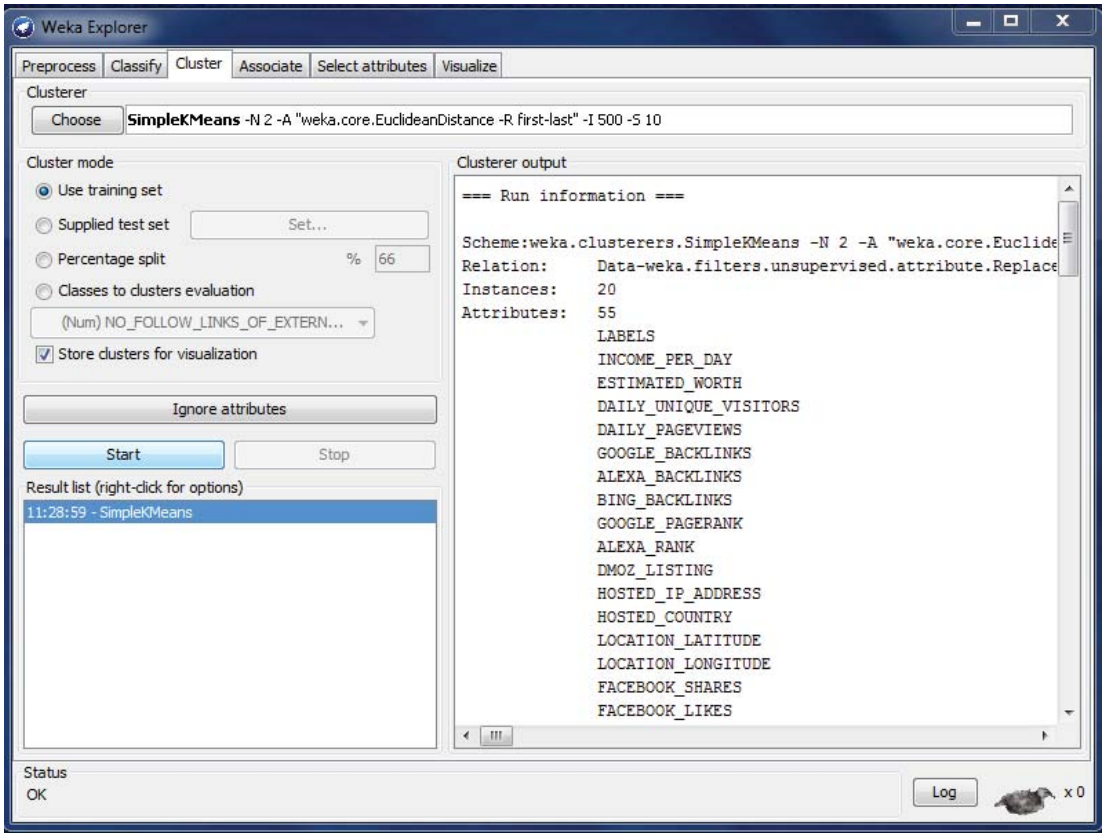


Figure 25: A Screenshot of the clustering output after the implementation of a simple K-means clustering algorithm is successfully processed.

In addition, the comprehensive interpretation/description of experiments and results regarding this particular clustering output based on a simple K-means clustering algorithm will be introduced and discussed more details in the next chapter of the thesis (i.e. Chapter 4 – Experiments and Results).

3.4 Descriptive Statistics

In this particular section (i.e. 3.4 Descriptive Statistics), the main theoretical framework regarding the key properties of variation and shape such as skewness and kurtosis will be introduced and discussed in terms of normally distributed data perspective.

In addition, this particular section also demonstrates that the key significant statistical results such as interval variable summary statistics (i.e. mean, median, maximum, minimum and standard deviation etc.), finding variation and shape from normally distributed data (i.e. skewness and kurtosis) in this thesis will be produced and implemented based on the output which has been primarily generated from the implementation of Statistical Analysis System (SAS) Enterprise Miner 13.1. data mining application.

Detailed interpretation in regard to the significant results of five-number summary will be introduced and discussed comprehensively in the following section in Chapter Four: 4.1 – Descriptive Statistics. Furthermore, detailed interpretation in regard to the significant results of both Skewness and Kurtosis will also be introduced and discussed in the following section in Chapter Four: 4.1 – Descriptive Statistics.

Finding Variation and Shape from Normally Distributed Data

According to Fernandez (2010), this study highlighted that the following main properties of these two components (i.e. skewness and kurtosis) from normally distributed data in relation to this research are comprehensively indicated as below:

- ***Skewness*:** This study from Fernandez (2010) defined that the skewness is a measured value to quantify the direction and degree of asymmetry of the frequency distribution. Fernandez (2010) indicated that if a frequency distribution is displayed or shown evenly to the left and right direction from the centre point, the frequency distribution of a continuous variable primarily appears to be symmetric distribution. This study also showed that data based on positively skewed frequency distribution appears to have relevant values that contain the basic feature of clustering together below the value of mean. (i.e. it is skewed to the right direction). However, this positively skewed frequency distribution appeared to maintain a long tail above the value of mean. On the other hand, Fernandez (2010) showed in this study that data based on negatively skewed frequency distribution appears to have relevant values that contain the basic feature of clustering together above the value of mean. (i.e. it is skewed to the left direction). However, Fernandez (2010) indicated that this negatively skewed frequency distribution appeared to maintain a long tail below the value of mean. In addition, this study indicated that skewness estimate in the case of a normal distribution is equivalent to zero. This study from Fernandez (2010) also highlighted

that the data appears to be skewed to the left direction in the case of estimate from a negative skewness. Therefore, in this study, this estimate based on a negative skewness implies that the left tail of normal distribution is heavier compared to the right tail of this normal distribution. Moreover, Fernandez (2010) also highlighted in this study that the data appears to be skewed to the right direction in the case of estimate from a positive skewness. Therefore, in this study, this estimate based on a positive skewness implies that the right tail of normal distribution is heavier compared to the left tail of this particular normal distribution.

- ***Kurtosis*:** This study from Fernandez (2010) defined that the kurtosis is a measure that is primarily designed to quantify about data's pointed degree (i.e. whether data is flat or peaked) of the normal distribution based on the central tendency values. According to Fernandez (2010), this study indicated that the datasets associated with large kurtosis of distribution appears to clearly show a unique peak around the mean and also appears to maintain the distinct feature of heavy tails as well as decreased rather rapidly. On the other hand, Fernandez (2010) indicated in this study that the datasets associated with low kurtosis of distribution appears to clearly show a unique flat top shape near the mean instead of a sharp peak shape. This study from Fernandez (2010) highlighted that kurtosis can be consisted of both positive and negative kurtosis and also indicated that distributions associated with the prevalent features of positive kurtosis appear to clearly maintain or have the typical characteristics of heavy tails. Furthermore, this study from Fernandez (2010) specified that the basic presence of outliers is important factor to affect the outcome of both kurtosis and skewness estimates. Because, this study showed that these two particular estimates are appeared to be very sensitive depending on the existence of outliers in distributions as indicated by Fernandez (2010). Hence, this study from Fernandez (2010) implied that these estimates above based on kurtosis and skewness can be directly affected by the presence of outstanding observed patterns or shapes such as outliers, which is typically appeared in the tail segment of the distribution.

The next procedure is to produce the descriptive statistics output which is generated from the implementation of SAS Enterprise Miner 13.1 data mining software. To generate the descriptive statistics, the following key steps are needed to undertake as below: which is based on the basic raw data file as of 11/08/2014 (Data.xlsx) in the form of a Microsoft Excel worksheet as previously displayed in Figure 14.

Step One: First the following main welcome screen can be displayed by executing this application - SAS Enterprise Miner Workstation 13.1 as seen in Figure 26 below.

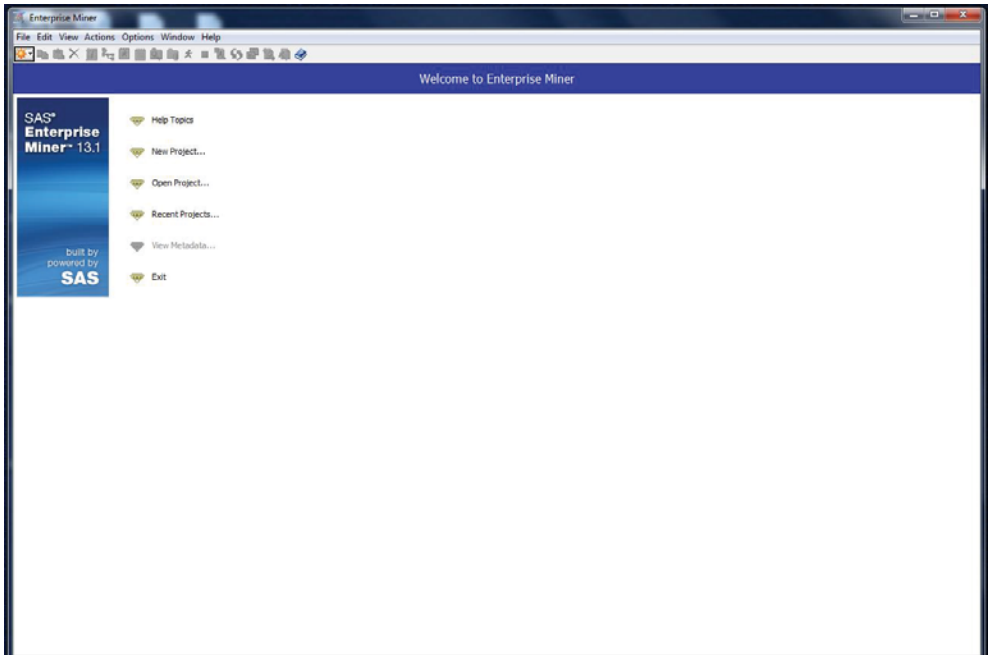


Figure 26: A Screenshot of SAS Enterprise Miner 13.1 Main Welcome Screen.

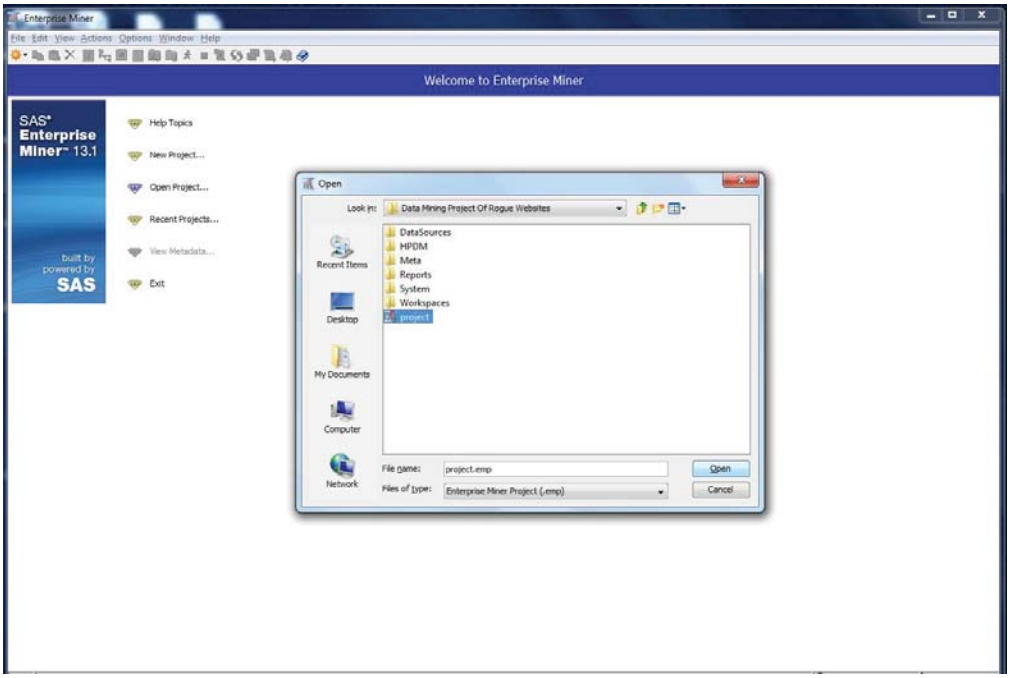


Figure 27: A Screenshot of Opening the Project in SAS Enterprise Miner 13.1.

Step Two: As displayed in Figure 27 above, the screen of opening the project can be displayed in SAS Enterprise Miner 13.1 project environment. The next procedure is to select a project called "project" within the following project folder – *Data Mining Project Of Rogue Websites*.

Step Three: To view the project diagrams, the next procedure is to select and click the project diagram called "Data Mining Project of Rogue Websites" in the Project Panel (i.e. left upper section of the main project screen) of SAS Enterprise Miner 13.1 project environment. Figure 28 as displayed below shows a screenshot of this particular step.

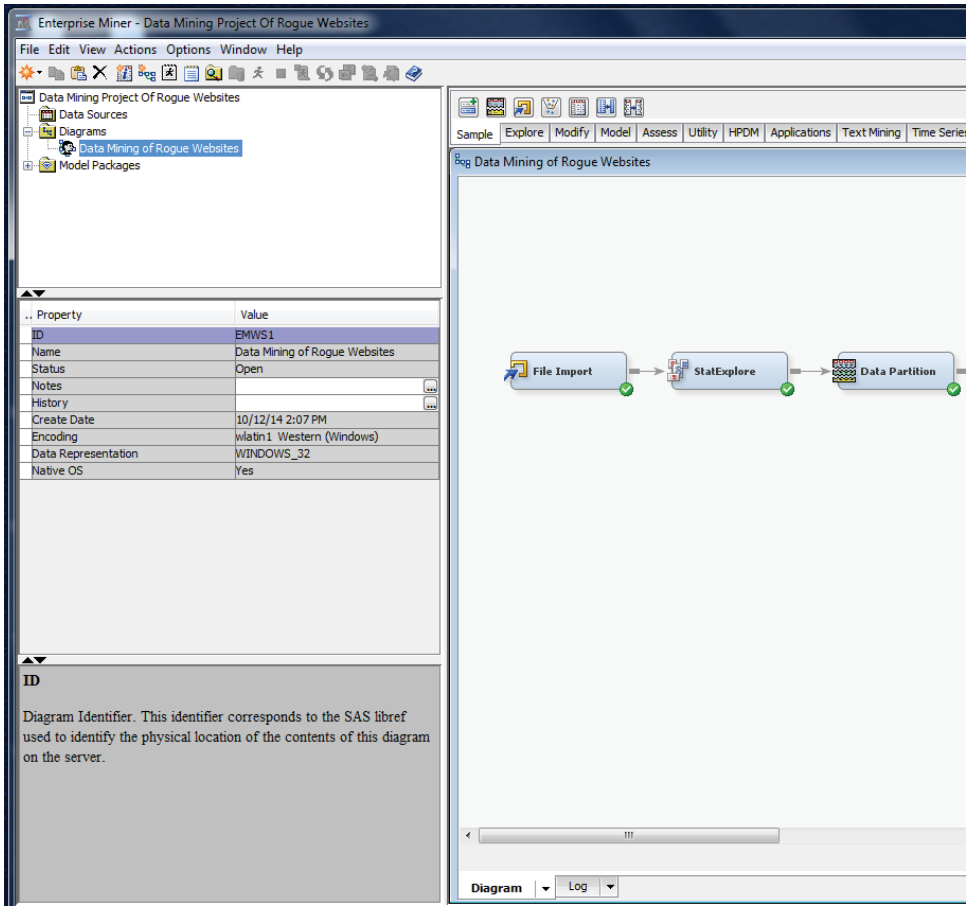


Figure 28: A Screenshot of selecting and opening the main project diagram in the Project Panel of SAS Enterprise Miner 13.1 environment.

Step Four: in the following SAS Enterprise Miner project created – "Data Mining of Rogue Websites" as illustrated in Figure 29 below, the corresponding descriptive statistics output as displayed in Table 17 from the following section in Chapter 4 (i.e. Experiments and Results - 4.1 Descriptive Statistics) will be generated by selecting the further option called "Results" from *StatExplore Node* in SAS Enterprise Miner 13.1 project diagram workspace as displayed in Figure 29 below.

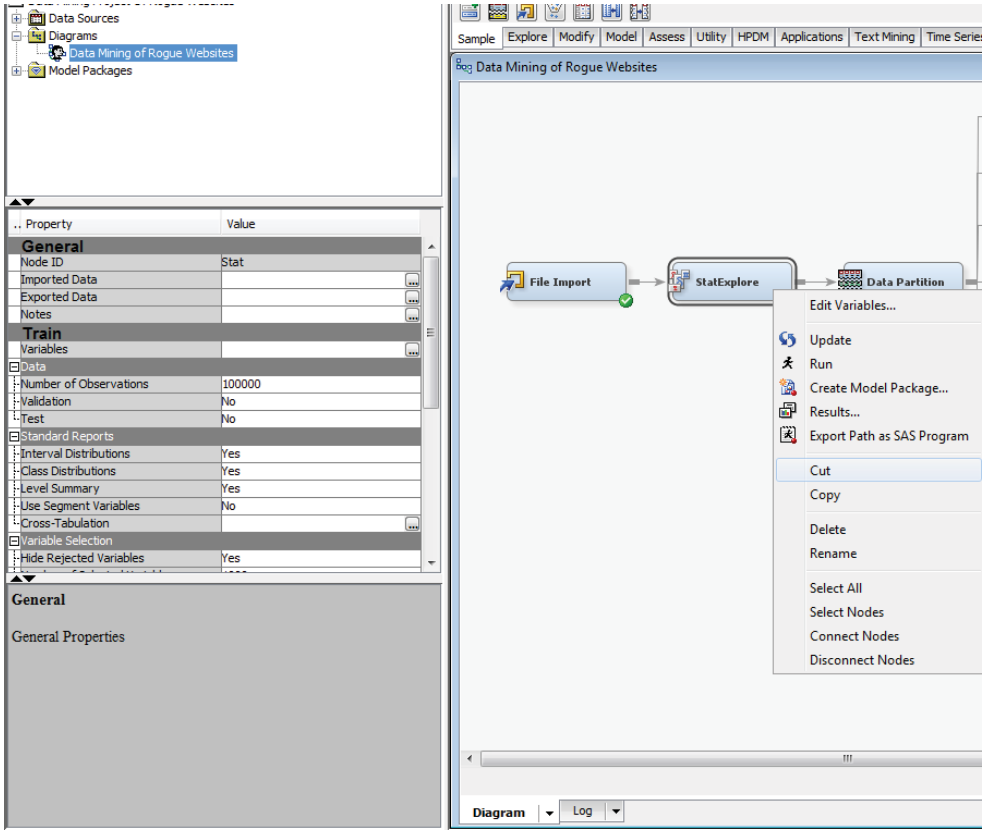


Figure 29: A Screenshot of selecting and displaying the option called "Results" from *StatExplore Node* in SAS Enterprise Miner 13.1 project diagram workspace.

Regarding the final descriptive statistics output generated from SAS Enterprise Miner 13.1 project diagram workspace as displayed in Figure 29 above, more detailed information about this corresponding final result from descriptive statistics output will be introduced and discussed comprehensively in the following section in Chapter 4 (i.e. Experiments and Results - 4.1 Descriptive Statistics).

3.5 Summary

In this chapter, it demonstrates that the basic data collection used for the thesis research (i.e. 20 instances and 55 attributes) was fully undertaken and produced in the form of a Microsoft Excel worksheet. This chapter also shows that this basic data collection was initially obtained and implemented on 11th of August 2014 from the corresponding search results from the CuteStat.com (i.e. a website providing website valuation data and website statistical information) regarding the Top Twenty rogue websites in terms of their website valuation information and related-statistical data. Furthermore, the main details of these Top Twenty rogue websites in all time have been obtained from the following link of specified domains (https://www.google.com/transparencyreport/removals/copyright/domains/?r=all-time) within Google Transparency Report (2015) https://www.google.com/transparencyreport/?hl=en_US.

Once the fundamental raw data collection process for Top Twenty rogue websites was fully done and processed in a Microsoft Excel worksheet format, this chapter also demonstrates that the following main data mining method or technique has been comprehensively used, analysed and implemented from the basic data source file (i.e. Data.csv) regarding top twenty rogue websites in this research by using the practical implementation of Waikato Environment for Knowledge Analysis (WEKA) 3.6 data mining software: Cluster Analysis based on using a simple K-means clustering algorithm.

In addition, this chapter demonstrates that the key significant statistical results such as interval variable summary statistics (i.e. mean, median, maximum, minimum and standard deviation etc.), finding variation and shape from normally distributed data (i.e. skewness and kurtosis) in this thesis can be produced and implemented effectively based on the output which has been primarily generated from the implementation of Statistical Analysis System (SAS) Enterprise Miner 13.1 data mining application.

CHAPTER FOUR

Experiments and Results

4.1 Descriptive Statistics

In this particular section of the thesis, the research results based on the data collection of Top 20 rogue websites (i.e. 20 instances and 55 attributes) as previously discussed in Chapter 3.2 (i.e. Data Collection) which have been initially retrieved and obtained from Google Transparency Report (2015) will be discussed and interpreted in this particular section comprehensively in terms of descriptive statistics through the implementation of Statistical Analysis System (SAS) Enterprise Miner 13.1 in this research.

In the contents of descriptive statistics, the main significant results generated from the perspective of interquartile range (IQR) will be introduced and interpreted including the further significant findings of variation and shape such as skewness and kurtosis from normally distributed data that we have obtained through the implementation of Statistical Analysis System (SAS) Enterprise Miner 13.1 in this research.

In addition, the final results of five-number summary statistics will be introduced and interpreted based on the output which has been mainly generated from the implementation of Statistical Analysis System (SAS) Enterprise Miner 13.1. data mining application.

Furthermore, this particular section (i.e. 4.2 Cluster Analysis) will be introduced and interpreted comprehensively about the final results in this thesis in terms of cluster analysis within this entire chapter 4 (i.e. Experiments and Results). Detailed interpretation of this particular section (i.e. 4.2 Cluster Analysis) in this thesis will also be comprehensively discussed in terms of the following contents respectively below:

(1) Income Per Day and Estimated Valuation

(2) Daily Unique Visitors and Daily Page-views

(3) Search Engine Backlinks (e.g. Google Backlinks, Alexa Backlinks and Bing Backlinks)

(4) Website Ranks and Scores

(5) Location Latitude and Location Longitude

(6) Social Network Engagement (e.g. Facebook Shares, Facebook Likes and Twitter Count)

(7) Search Engine Indexes

(8) Page Resources Breakdown

(9) Homepage Links Analysis

(10) Online Safety Information

The next procedure is to interpret the descriptive statistics output which is primarily generated from the implementation of SAS Enterprise Miner 13.1 data mining software. In the following SAS Enterprise Miner project created – "Data Mining of Rogue Websites" as illustrated in Figure 30 below, the corresponding descriptive statistics output as displayed in Table 17 in the next page can be generated by selecting the further option called "Results" from *StatExplore Node* in SAS Enterprise Miner 13.1 project diagram workspace.

As displayed in Figure 30 below, this particular screenshot shows a corresponding screen of selecting and displaying the option called "Results" from *StatExplore Node* in SAS Enterprise Miner 13.1 project diagram workspace.

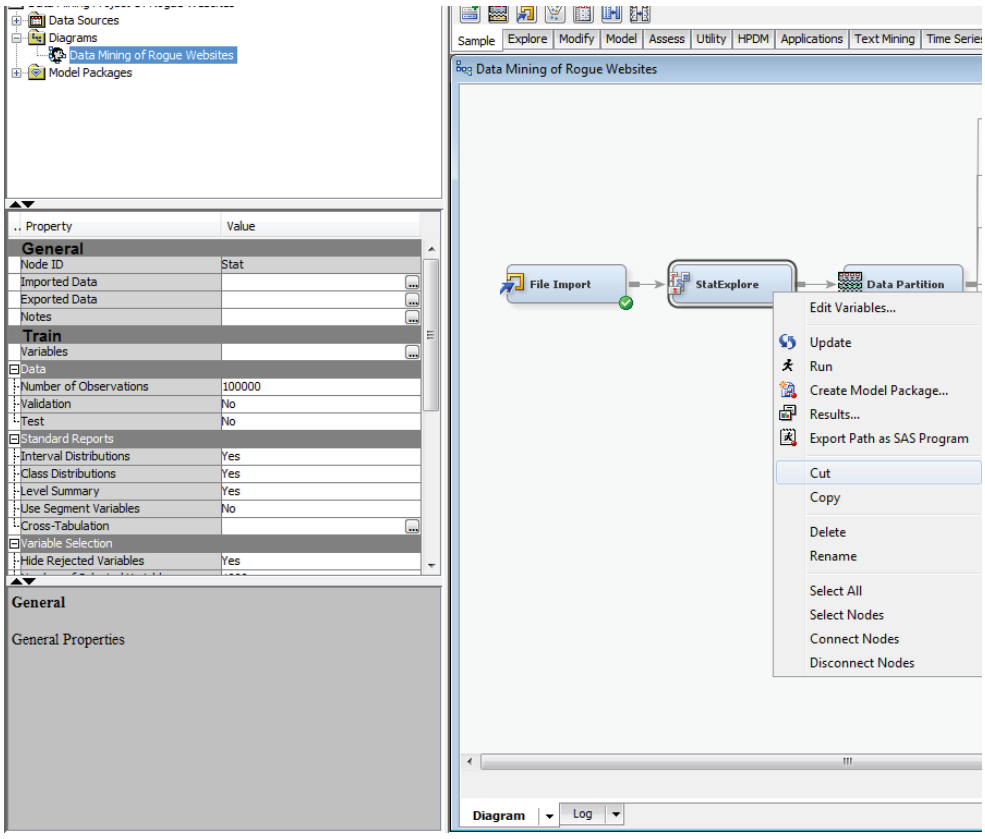


Figure 30: A Screenshot of selecting and displaying the option called "Results" from *StatExplore Node* in SAS Enterprise Miner 13.1 project diagram workspace.

Interval Variable Summary Statistics
(maximum 500 observations printed)

Data Role=TRAIN

Variable	Role	Mean	Standard Deviation	Non Missing	Missing	Minimum	Median	Maximum	Skewness	Kurtosis
ALEXA_RANK	INPUT	39719.7	139666.3	20	0	162	2918	631436	4.431197	19.73941
CSS_KILOBYTES	INPUT	43.4	46.91639	11	9	4.4	26.1	149.1	1.389653	1.25286
HTML_KILOBYTES	INPUT	65.5	66.44819	12	8	2.6	34.2	202.5	1.018212	0.018039
IMAGES_KILOBYTES	INPUT	101.6	89.34823	12	8	2.2	42.6	240.1	0.439815	-1.43545
JAVASCRIPT_KILOBYTES	INPUT	549.6167	515.1714	12	8	19	239.9	1710	1.032249	0.673957
LOCATION_LATITUDE	INPUT	45.5947	10.35968	20	0	18.4167	47.3667	60	-0.80462	0.982214
LOCATION_LONGITUDE	INPUT	-28.6429	55.30603	20	0	-118.244	2.35099	100	0.289215	-0.32828
NO_FOLLOW_LINKS_OF_EXTERNAL_LINK	INPUT	1.833333	2.979729	12	8	0	0	8	1.380889	0.510016
NO_FOLLOW_LINKS_OF_INTERNAL_LINK	INPUT	1.083333	2.35327	12	8	0	0	8	2.743687	7.852152
TOTAL_LINKS_OF_EXTERNAL_LINKS	INPUT	3.75	5.29365	12	8	0	1	18	2.056451	4.580299
TOTAL_LINKS_OF_INTERNAL_LINKS	INPUT	53.41667	87.62052	12	8	0	20	291	2.328881	5.056323
TWITTER_COUNT	INPUT	5530.1	8826.557	20	0	1	181	29410	1.838495	2.489306
DAILY_PAGEVIEWS	TARGET	10513601	15202649	20	0	1524	1179688	50486192	1.427481	1.015724
DAILY_UNIQUE_VISITORS	TARGET	1415354	2078411	20	0	762	147461	6310774	1.3909	0.544024
ESTIMATED_WORTH	TARGET	9720226	15481140	20	0	1200	1274400	54524880	1.922771	2.940569
FACEBOOK_SHARES	TARGET	13047.2	37881.41	20	0	1	1342	169220	4.08799	17.35708
GOOGLE_PAGERANK	TARGET	4.4	1.095445	20	0	3	4	6	0.149476	-1.22033
INCOME_PER_DAY	TARGET	8686.289	14149.52	20	0	5	1180	50486	2.054107	3.474016

Table 17: A comprehensive information of Interval Variable Summary Statistics (i.e. descriptive statistics output) generated from the implementation of SAS Enterprise Miner 13.1 project (Data Mining of Rogue Websites).

The descriptive statistics output as shown from Table 17 above implies that almost all the variables have a positive or right-skewed distribution of the data except for one variable (i.e. Skewness of this variable (LOCATION_LATITUDE): -0.80462) which has a negative or left-skewed distribution of the data. This outcome also means that the mean of this particular variable (i.e. LOCATION_LATITUDE) is less than the value of its median. On the other hand, all the remaining variables have the mean which is greater than their median values. According to Berenson, Levine & Krehbiel (2009), the following main properties of the distribution of data have also highlighted in terms of skewness as below:

- If the data is involved in either a left-skewed or negative distribution, this study from Berenson, Levine & Krehbiel (2009) indicated that the greater part of the values is appeared to locate in the upper segment of the distribution. This study also showed that both left-skewed distribution and a long tail to the left direction can be primarily affected by the presence of enormously small values observed. Berenson, Levine & Krehbiel (2009) also indicated in this study that the presence of these enormously small values is to basically influence the mean value descending; therefore the value of mean becomes less than the value of the median due to this effect of the mean descending or downward.

- If the data is involved in a symmetrical distribution, this study from Berenson, Levine & Krehbiel (2009) indicated that this particular outcome can be primarily observed or appeared as a mirror or reflection image between each half of the corresponding curve and the other half of the corresponding curve. This study from Berenson, Levine & Krehbiel (2009) also highlighted that low value and high value for the scale balance and the value of mean are appeared to be equivalent to the median value.

- If the data is involved in either a right-skewed or positive distribution, this study from Berenson, Levine & Krehbiel (2009) indicated that the greater part of the values is appeared to locate in the lower segment of the distribution. This study also showed that both right-skewed distribution and a long tail to the right direction can be primarily affected by the presence of enormously large values observed. Berenson, Levine & Krehbiel (2009) also indicated in this study that the presence of these enormously large values is to basically influence the mean value ascending; therefore the value of mean becomes greater than the value of the median due to this effect of the mean ascending or upward.

Therefore, the descriptive statistics outcome from Table 17 above also implies that these two particular variables such as ALEXA_RANK and FACEBOOK_SHARES have the following large values of skewness respectively: 4.431197 and 4.08799 so that these two large significant skewness values can be clearly observed or considered as a main key factor to influence to the mean which is greater than the median on the basis of related-theory above from Berenson, Levine & Krehbiel (2009).

Furthermore, the outcomes from Table 17 above show that almost all the variables have a positive (i.e. right-skewed) distribution of the data in terms of the skewness value despite there are also eight variables in which each of them have 8 or 9 missing values (or not-applicable values) obtained from the initial data collection process through CuteStat.com (2015) as of 11/08/2014. It can also be observed through the descriptive output from Table 17 that these 8 or 9 missing values from each of these eight rogue website were mostly appeared due to the not-existing numerical (or not-applicable) values of the allocation of languages and other components allocated for their particular websites (e.g. CSS, HTML, JavaScript and Images allocated) at the initial data collection process through CuteStat.com (2015) as of 11/08/2014. All the external and internal links within these 8 rogue websites (e.g. NO_FOLLOW_LINKS_OF_EXTERNAL_LINKS, TOTAL_LINKS_OF_INTERNAL_LINKS and TOTAL_LINKS_OF_EXTERNAL_LINKS etc.) have the same reason of having a significant number of missing values. Because the values of these variables also have non-existing or missing values at the initial data collection process through CuteStat.com (2015) as of 11/08/2014.

In addition, from the standpoint of kurtosis, the descriptive statistics output from Table 17 implies that these particular variables such as ALEXA_RANK and FACEBOOK_SHARES with extremely large kurtosis values (i.e. 19.73941 and 17.35708 respectively) can be observed to have or maintain a unique peak around the mean as well as maintaining thick or heavy tails at the same time, in relation with information from Fernandez (2010). On the other hand, the same descriptive statistics output from Table 17 also implies that these five particular variables such as HTML_KILOBYTES (0.018039),

JAVASCRIPT_KILOBYTES (0.673957), LOCATION_LONGITUDE (-0.32828), NO_FOLLOW_LINKS_OF_EXTERNAL_LINK (0.510016)

and DAILY_UNIQUE_VISITORS (0.544024) respectively with low kurtosis values can be observed to maintain or have a unique flat top shape around the mean instead of a sharp peak shape around the mean (e.g. ALEXA_RANK and FACEBOOK_SHARES), in relation with information from Fernandez (2010). In addition, the distributions of these particular variables with positive kurtosis values, except for this variable with negative kurtosis value (LOCATION_LONGITUDE: -0.32828) can be considered or observed to have heavy tails mostly.

The Five-Number Summary Statistics

According to the information from Fernandez (2010), it indicates that the following main features of the five-number summary statistics are described in a statistical context as below:

- This study from Fernandez (2010) indicated that there are five-number summary of a continuous variable as follows: the maximum value, the third quartile, the median, the first quartile and the minimum value.
- This study from Fernandez (2010) also indicated that the second quartile or the median refers to the middle value of the arranged data.

- In this study, Fernandez (2010) indicated that the first quartile represents the 25th percentile of the arranged data and the third quartile represents the 75th percentile of the arranged data. Furthermore, this study showed that half of the data is primarily included in the range between the first quartile and the third quartile.
- This study from Fernandez (2010) indicated that the interquartile range (IQR) is equivalent to the difference between the third quartile (i.e. the 75th percentile of the arranged data) and the first quartile (i.e. the 25th percentile of the arranged data).
- Therefore, this study from Fernandez (2010) implied that there are largely three main features that can be provided from these five-number summary as follows:
  (1) This study indicated that the full range of statistical variation is available from the minimum value to the maximum value.
  (2) The common range of statistical variation is available from the first quartile to the third quartile.
  (3) This study indicated that a typical value can also be provided from the five-number summary. (i.e. the median)

Because the existing values of the maximum, the minimum and the median have already been generated and summarised as an output shown in Table 17 previously from the implementation of SAS Enterprise Miner 13.1 data mining software, Figure 31 below shows that the values of finding out the first quartile, the third quartile and the IQR can be calculated by using the following method called “Insert Function” under the menu called “Formulas” in Microsoft Excel worksheet environment:

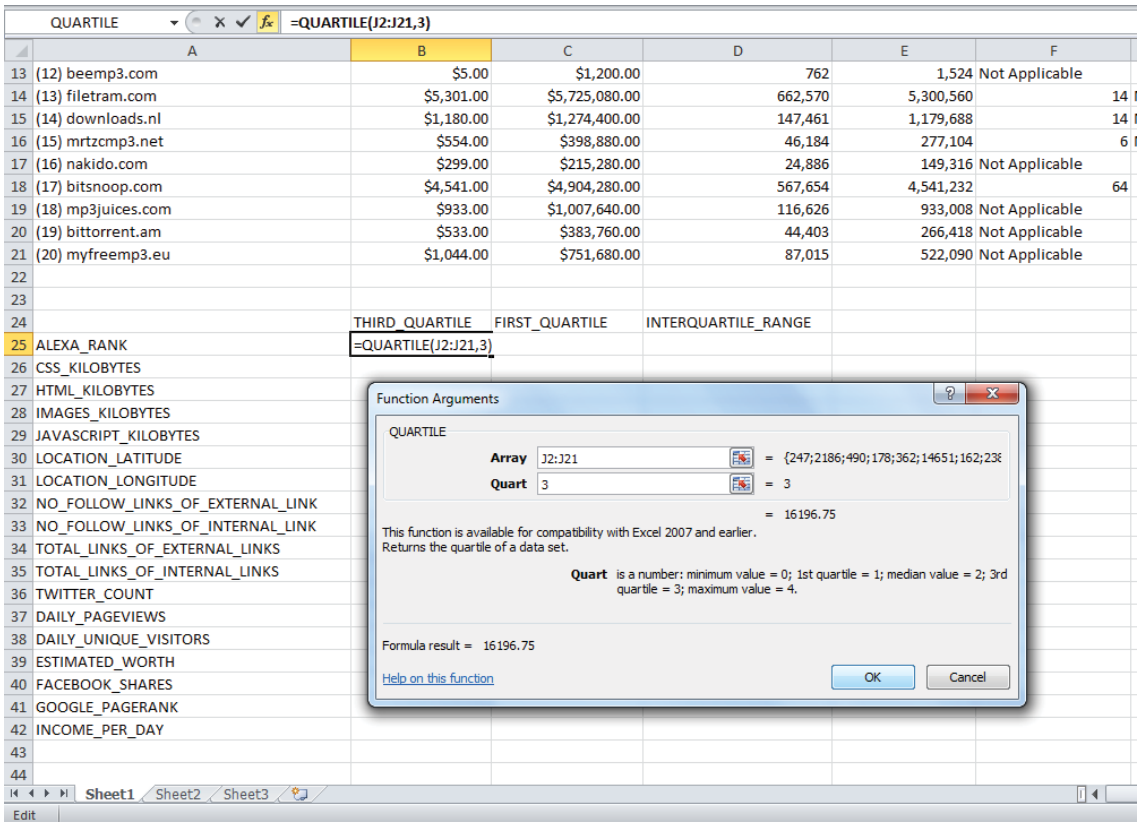


Figure 31: A Screenshot of inserting the function arguments for calculating the third quartile value of an existing variable (i.e. ALEXA_RANK).

Figure 31 above shows how to calculate the quartile values (e.g. the third quartile and the first quartile) by searching for a function in Microsoft Excel environment. The first step is to select the menu called "Formulas" and then choose the further option called "Insert Function". In the text field called "Search for a function", the user needs to enter the name of function called "QUARTILE" and then the above screenshot in Figure 31 can be displayed. In the Array section, the user then needs to highlight the corresponding columns (i.e. ALEXA_RANK: from cell J2 to cell J21) and in the Quart section, the input value of 3 can be entered in this section. To obtain the first quartile value, the user also needs to highlight the same range of Array (i.e. from cell J2 to cell J21 in the case of variable(ALEXA_RANK)) but the input value of entering 1 in Quart section should be required. In Figure 31, the output of calculating the third quartile for this particular variable (ALEXA_RANK) shows the value of 16196.75 after function arguments is fully done. In addition, according to Fernandez (2010) information, this study confirms that the IQR is equivalent to the difference between the third quartile and the first quartile (i.e. IQR = the third quartile - the first quartile). Therefore, the value of IQR can be calculated by doing the subtraction between the third quartile value and the first quartile value as displayed in Table 18.

THIRD_QUARTILE	FIRST_QUARTILE	INTERQUARTILE_RANGE
16196.75	458	15738.75

Table 18: The calculated results of the third quartile, the first quartile and the IQR for an existing variable (i.e. ALEXA_RANK).

Hence, all the remaining variables can also be calculated in a similar method as explained above in order to obtain the values of the third quartile, the first quartile and the IQR.

Table 19 as shown below displays the comprehensive five-number summary statistics including the existing results (i.e. maximum, median and minimum) which have been generated from the implementation of corresponding SAS Enterprise Miner 13.1 project (Data Mining of Rogue Websites) as illustrated previously in Table 17. Furthermore, the output from Table 19 below also shows an overall numerical data of the interquartile range (IQR) which represents the difference between the third quartile (i.e. the 75th percentile) and the first quartile (i.e. 25th percentile) on the basis of the related-theoretical information from Fernandez (2010) as previously indicated.

	MAXIMUM	THIRD_QUARTILE	MEDIAN	FIRST_QUARTILE	MINIMUM	INTERQUARTILE_RANGE (IQR)
ALEXA_RANK	631436	16196.75	2918	458	162	15738.75
CSS_KILOBYTES	149.1	61.35	26.1	8.6	4.4	52.75
HTML_KILOBYTES	202.5	99.175	34.2	10.4	2.6	88.775
IMAGES_KILOBYTES	240.1	159.175	42.6	30.3	2.2	128.875
JAVASCRIPT_KILOBYTES	1710	850.35	239.9	148.475	19	701.875
LOCATION_LATITUDE	60	52.356	47.3667	38.9498	18.4167	13.4062
LOCATION_LONGITUDE	100	4.968305	2.35099	-77.2278	-118.244	82.196105
NO_FOLLOW_LINKS_OF_EXTERNAL_LINK	8	3.25	0	0	0	3.25
NO_FOLLOW_LINKS_OF_INTERNAL_LINK	8	1	0	0	0	1
TOTAL_LINKS_OF_EXTERNAL_LINKS	18	4.5	1	0	0	4.5
TOTAL_LINKS_OF_INTERNAL_LINKS	291	36.25	20	7.5	0	28.75
TWITTER_COUNT	29410	6626	181	1	1	6625
DAILY_PAGEVIEWS	50486192	20858905.5	1179688	364813.5	1524	20494092
DAILY_UNIQUE_VISITORS	6310774	2314312	147461	60802.25	762	2253509.75
ESTIMATED_WORTH	54524880	9239912.5	1274400	525240	1200	8714672.5
FACEBOOK_SHARES	169220	5953	1342	1	1	5952
GOOGLE_PAGERANK	6	5	4	3.75	3	1.25
INCOME_PER_DAY	50486	6942.936	1180	729.5	5	6213.436

Table 19: Comprehensive Five-Number Summary Statistics and Interquartile Range (IQR)

In Table 19 as shown above, one of the most significant observations is that there is a large difference between the total number of links for both external and internal links and the total number of no-follow links for both external and internal links in the five-number summary statistics. As displayed in Table 19, we found the following significant findings based the result of five-number summary statistics: Total number of links from internal links show 291 links as maximum and the total number of links from external links indicate 18 links as maximum which are relatively greater than the numbers of no-follow links from both internal and external links in comparison.

4.2 Cluster Analysis

In this section of the thesis, the final statistical output or results through the implementation of a simple K-means clustering algorithm generated in WEKA 3.6 data mining environment as previously discussed in chapter 3.3 (i.e. Research Methodology section of Cluster Analysis) will be interpreted and discussed comprehensively in this chapter.

kMeans

Number of iterations: 4
Within cluster sum of squared errors: 317.43697031599424
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (20)	Cluster Number 0 (11)	Cluster Number 1 (9)
INCOME_PER_DAY	$33,112.00	$3,741.00	$33,112.00
ESTIMATED_WORTH	$35,760,960.00	$4,040,280	$35,760,960.00
DAILY_UNIQUE_VISITORS	4,139,050	467,679	4,139,050
DAILY_PAGEVIEWS	33,112,400	3,741,432	33,112,400
GOOGLE_BACKLINKS	41.7143	35.4805	49.3333
ALEXA_BACKLINKS	13,655	13,655	13,655
BING_BACKLINKS	26.125	28.0455	23.7778
GOOGLE_PAGERANK	4.4	3.9091	5
ALEXA_RANK	247	2,186	247
DMOZ_LISTING	No	No	No
HOSTED_IP_ADDRESS	78.140.188.239	176.31.230.116	78.140.188.239
HOSTED_COUNTRY	United States	United States	United States
LOCATION_LATITUDE	45.5947	48.0589	42.5829
LOCATION_LONGITUDE	-28.6429	-20.4453	-38.6622
FACEBOOK_SHARES	13047.2	934.7273	27851.3333
FACEBOOK_LIKES	28782.75	22035.1818	37029.7778
FACEBOOK_COMMENTS	5271.0833	4077.2424	6730.2222
TWITTER_COUNT	5530.1	1656.1818	10264.8889
LINKEDIN_SHARES	175.5455	137.0331	222.6162
DELICIOUS_SHARES	2521	2404.9091	2662.8889
GOOGLE_PLUS	11855.5	9415.1818	14838.1111
DOMAIN_REGISTRAR	EURID	EURID	Instra Corporation Pty, Ltd.
WEBSITE_OWNER	Alex Lunkov	Alex Lunkov	Alex Lunkov
OWNERS_EMAIL_ADDRESS			

Table 20: Clustering Output of a Simple K-means clustering algorithm for dataset (Data.csv).

IP_ADDRESS_NUMBER_ONE	207.226.173.74	176.31.230.116	149.13.65.167
COUNTRY_ONE	United States	United States	United States
IP_ADDRESS_NUMBER_TWO	88.208.58.214	88.208.58.214	88.208.58.214
COUNTRY_TWO	United States	United States	United States
IP_ADDRESS_NUMBER_THREE	75.126.229.92	75.126.229.92	75.126.229.92
COUNTRY_THREE	United States	United States	United States
IP_ADDRESS_NUMBER_FOUR	208.76.60.2	208.76.60.2	208.76.60.2
COUNTRY_FOUR	United States	United States	United States
IP_ADDRESS_NUMBER_FIVE	208.76.58.2	208.76.58.2	208.76.58.2
COUNTRY_FIVE	United States	United States	United States
IP_ADDRESS_NUMBER_SIX	208.76.61.2	208.76.61.2	208.76.61.2
COUNTRY_SIX	United States	United States	United States
GOOGLE_SAFE_BROWSING	No Risk Issues	No Risk Issues	No Risk Issues
SITEADVISOR_RATING	No Risk Issues	No Risk Issues	No Risk Issues
WOT_TRUSTWORTHINESS	Excellent	Very Poor	Excellent
WOT_PRIVACY	Excellent	Very Poor	Excellent
WOT_CHILD_SAFETY	Very Poor	Very Poor	Excellent
GOOGLE_INDEXED_PAGES	78,900,000	78,900,000	78,900,000
YAHOO_INDEXED_PAGES	51,400	51,400	51,400
BING_INDEXED_PAGES	42	42	42
JAVASCRIPT_KILOBYTES	549.6167	657.3394	417.9556
HTML_KILOBYTES	65.5	81.4	46.0667
IMAGES_KILOBYTES	101.6	101.0909	102.2222
CSS_KILOBYTES	43.4	42.1	44.9889
KILOBYTES_TEXT	88.8	88.8	88.8
OTHER_KILOBYTES	35.35	47.8773	20.0389
TOTAL_LINKS_OF_INTERNAL_LINKS	53.4167	42.4848	66.7778
NO_FOLLOW_LINKS_OF_INTERNAL_LINKS	1.0833	0.8788	1.3333
TOTAL_LINKS_OF_EXTERNAL_LINKS	3.75	3.0909	4.5556
NO_FOLLOW_LINKS_OF_EXTERNAL_LINKS	1.8333	1.3333	2.4444

Time taken to build model (full training data) : 0.18 seconds

Model and evaluation on training set

Clustered Instances

0 11 (55%)
1 9 (45%)

Table 21: Clustering Output of a Simple K-means clustering algorithm for dataset (Data.csv).

The next procedure is to create and add a specific pre-defined criteria or standard as shown in Table 22 below to provide the final results of the percentage for Cluster Number 1 over Cluster Number 0 based on specific criteria as follows:

Symbol	Meaning
< or >	<50% or >50%
<< or >>	<100% or >100%
<<< or >>>	<1000% or >1000%

Table 22: A table information of pre-defined criteria

kMeans

Number of iterations: 4
Within cluster sum of squared errors: 317.43697031599424
Missing values globally replaced with mean/mode

Cluster centroids:

Attribute	Full Data (20)	Cluster Number 0 (11)	Cluster Number 1 (9)	Percentage for Cluster #1 / Cluster #0	Criteria
INCOME_PER_DAY	$33,112.00	$3,741.00	$33,112.00	885%	>>
ESTIMATED_WORTH	$35,760,960.00	$4,040,280	$35,760,960.00	885%	>>
DAILY_UNIQUE_VISITORS	4,139,050	467,679	4,139,050	885%	>>
DAILY_PAGEVIEWS	33,112,400	3,741,432	33,112,400	885%	>>
GOOGLE_BACKLINKS	41.7143	35.4805	49.3333	139%	>
ALEXA_BACKLINKS	13,655	13,655	13,655	100%	=
BING_BACKLINKS	26.125	28.0455	23.7778	85%	<
GOOGLE_PAGERANK	4.4	3.9091	5	128%	>
ALEXA_RANK	247	2,186	247	11%	<<
DMOZ_LISTING	No	No	No		
HOSTED_IP_ADDRESS	78.140.188.239	176.31.230.116	78.140.188.239		
HOSTED_COUNTRY	United States	United States	United States		
LOCATION_LATITUDE	45.5947	48.0589	42.5829	89%	<
LOCATION_LONGITUDE	-28.6429	-20.4453	-38.6622	189%	>>
FACEBOOK_SHARES	13047.2	934.7273	27851.3333	2980%	>>>
FACEBOOK_LIKES	28782.75	22035.1818	37029.7778	168%	>>
FACEBOOK_COMMENTS	5271.0833	4077.2424	6730.2222	165%	>>
TWITTER_COUNT	5530.1	1656.1818	10264.8889	620%	>>
LINKEDIN_SHARES	175.5455	137.0331	222.6162	162%	>>
DELICIOUS_SHARES	2521	2404.9091	2662.8889	111%	>

GOOGLE_PLUS	11855.5	9415.1818	14838.1111	158%	>>
DOMAIN_REGISTRAR	EURID	EURID	Instra Corporation Pty, Ltd.		
WEBSITE_OWNER	Alex Lunkov	Alex Lunkov	Alex Lunkov		
OWNERS_EMAIL_ADDRESS					
IP_ADDRESS_NUMBER_ONE	207.226.173.74	176.31.230.116	149.13.65.167		
COUNTRY_ONE	United States	United States	United States		
IP_ADDRESS_NUMBER_TWO	88.208.58.214	88.208.58.214	88.208.58.214		
COUNTRY_TWO	United States	United States	United States		
IP_ADDRESS_NUMBER_THREE	75.126.229.92	75.126.229.92	75.126.229.92		
COUNTRY_THREE	United States	United States	United States		
IP_ADDRESS_NUMBER_FOUR	208.76.60.2	208.76.60.2	208.76.60.2		
COUNTRY_FOUR	United States	United States	United States		
IP_ADDRESS_NUMBER_FIVE	208.76.58.2	208.76.58.2	208.76.58.2		
COUNTRY_FIVE	United States	United States	United States		
IP_ADDRESS_NUMBER_SIX	208.76.61.2	208.76.61.2	208.76.61.2		
COUNTRY_SIX	United States	United States	United States		
GOOGLE_SAFE_BROWSING	No Risk Issues	No Risk Issues	No Risk Issues		
SITEADVISOR_RATING	No Risk Issues	No Risk Issues	No Risk Issues		
WOT_TRUSTWORTHINESS	Excellent	Very Poor	Excellent		
WOT_PRIVACY	Excellent	Very Poor	Excellent		
WOT_CHILD_SAFETY	Very Poor	Very Poor	Excellent		
GOOGLE_INDEXED_PAGES	78,900,000	78,900,000	78,900,000	100%	=
YAHOO_INDEXED_PAGES	51,400	51,400	51,400	100%	=
BING_INDEXED_PAGES	42	42	42	100%	=
JAVASCRIPT_KILOBYTES	549.6167	657.3394	417.9556	64%	<
HTML_KILOBYTES	65.5	81.4	46.0667	57%	<
IMAGES_KILOBYTES	101.6	101.0909	102.2222	101%	>
CSS_KILOBYTES	43.4	42.1	44.9889	107%	>
KILOBYTES_TEXT	88.8	88.8	88.8	100%	=

OTHER_KILOBYTES	35.35	47.8773	20.0389	42%	<<
TOTAL_LINKS_OF_INTERNAL_LINKS	53.4167	42.4848	66.7778	157%	>>
NO_FOLLOW_LINKS_OF_INTERNAL_LINKS	1.0833	0.8788	1.3333	152%	>>
TOTAL_LINKS_OF_EXTERNAL_LINKS	3.75	3.0909	4.5556	147%	>
NO_FOLLOW_LINKS_OF_EXTERNAL_LINKS	1.8333	1.3333	2.4444	183%	>>

Time taken to build model (full training data) : 0.18 seconds

Model and evaluation on training set

Clustered Instances

0 11 (55%)
1 9 (45%)

Table 23: Final Clustering Output of a Simple K-means clustering algorithm for dataset (Data.csv) including the results of percentage for Cluster Number 1 over Cluster Number 0 based on the pre-defined criteria.

In summary, the k-means clustering algorithm identified two very distinct clusters, one associated with high income (HIC), cluster 1, and low income (LIC), cluster 0. There were 11 sites in the LIC and 9 sites in the HIC. As discussed below, there were quite significant differences in the centroid attribute values for the LIC and the HIC. An interpretation of these is provided below.

4.2.1 Income Per Day and Estimated Worth

(1) In Table 23 above, there are firstly following two main attributes in terms of estimated valuation: INCOME_PER_DAY and ESTIMATED_WORTH.

They both indicate the percentage for the HIC sites over the LIC sites as 885% (i.e. >>) which implies that the overall values of both income per day and estimated worth of the HIC sites is 885% greater than these two same attributes of the LIC sites. As illustrated in Table 23, these two attributes are also belonged to the category of >150% range based on results criteria, which is much greater than 150%. This supports my hypothesis that there are two distinct groups of the most-complained about sites: one group which generates a lot of revenue, and one which does not. Figure 32 below displays comprehensive bar-graph information of estimated valuation in terms of the HIC sites and the LIC sites. As illustrated in Figure 32, it shows that the HIC sites have much greater values of the estimated worth compared to the LIC sites.

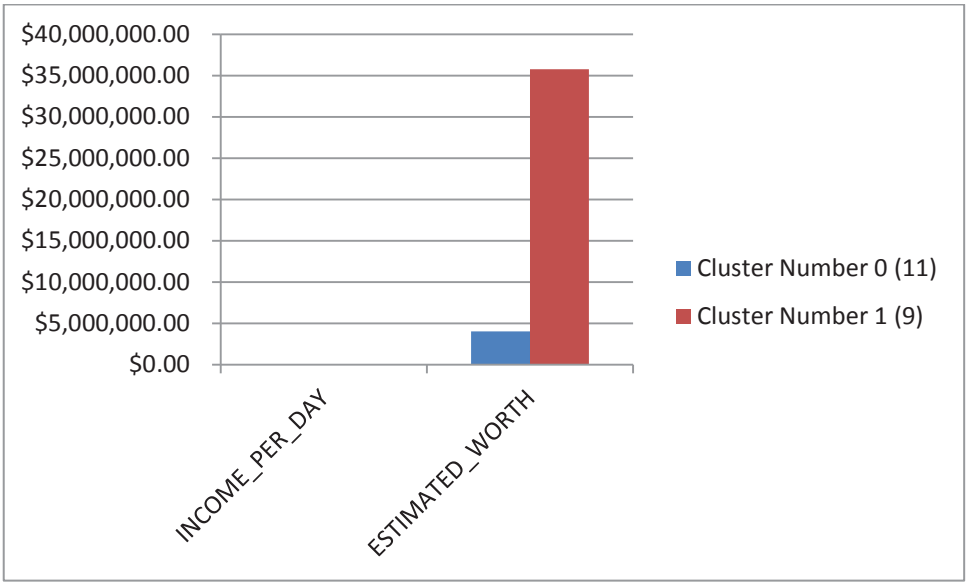


Figure 32: A bar-graph information of income per day and estimated valuation in terms of the HIC sites and the LIC sites.

4.2.2 Daily Unique Visitors and Daily Page-views

(2) In Table 23 above, there are following two attributes in terms of traffic report: DAILY_UNIQUE_VISITORS and DAILY_PAGEVIEWS.

They both show the percentage for Cluster Number 1 over Cluster Number 0 as 885% (i.e. >>) which basically implies that the overall values of both daily unique visitors and daily page-views of the HIC sites is 885% much greater than these two same attributes of the LIC sites. As displayed in Table 23, these two attributes are also belonged to the category of >150% range based on results criteria, which is much greater than 150%. Clearly, HIC sites are much more visited than LIC sites. Figure 33 below displays comprehensive bar-graph information of traffic report (i.e. daily unique visitors and daily page-views) in terms of the HIC sites and the LIC sites. As illustrated in Figure 33, it shows that the HIC sites have much greater values of both daily unique visitors and daily page-views compared to the LIC sites.

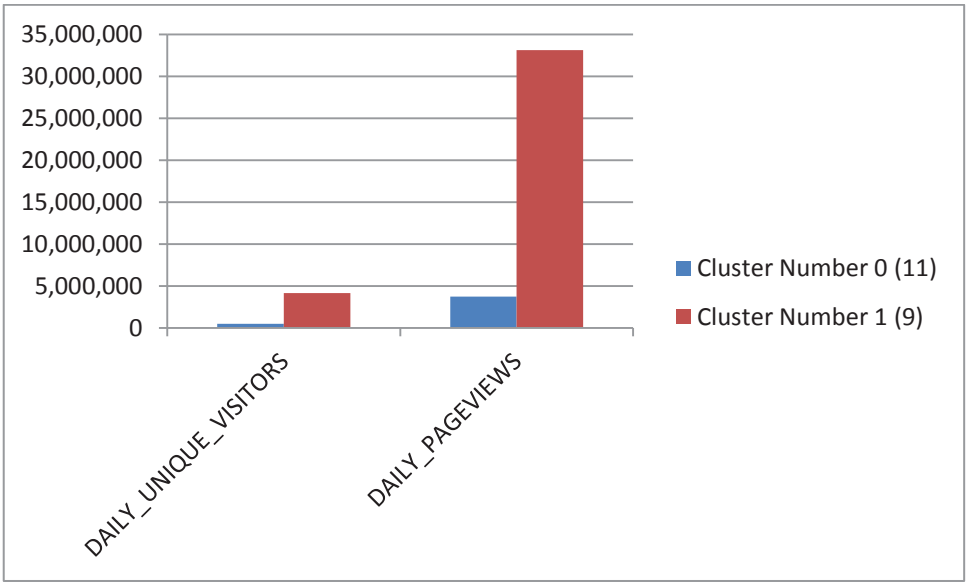


Figure 33: A bar-graph information of daily unique visitors and daily page-views in terms of the HIC sites and the LIC sites.

4.2.3 Search Engine Backlinks

(3) In Table 23 above, there are following three main attributes in terms of search engine backlinks: GOOGLE_BACKLINKS, ALEXA_BACKLINKS and BING_BACKLINKS.

The results criteria output from Table 23 shows that the percentages for Cluster Number 1 over Cluster Number 0 of these three attributes are indicated as 139% (i.e. >), 100% (i.e. =) and 85% (i.e. <) respectively. Therefore, this output implies that the overall numbers of

Google Backlinks regarding the HIC sites is 139% greater than the overall numbers of Google Backlinks for the LIC sites. On the other hands, it also implies that the overall number of Bing Backlinks regarding the HIC sites is 85% less than the overall number of Bing Backlinks for the LIC sites. Furthermore, the overall number of Alexa Backlinks regarding the HIC sites is exactly equal to the overall number of Alexa Backlinks for the LIC sites. Figure 34 below displays comprehensive bar-graph information of Search Engine Backlinks (i.e. Google Backlinks, Alexa Backlinks and Bing Backlinks) in terms of the HIC sites and the LIC sites. As illustrated in Figure 34, it shows that the HIC sites have a greater value of Google Backlinks (i.e. 139%), an equal value of Alexa Backlinks (i.e. 100%) and a less value of Bing Backlinks (i.e. 85%) respectively compared to these three same attributes of the LIC sites, according to our final clustering output as previously shown in Table 23.

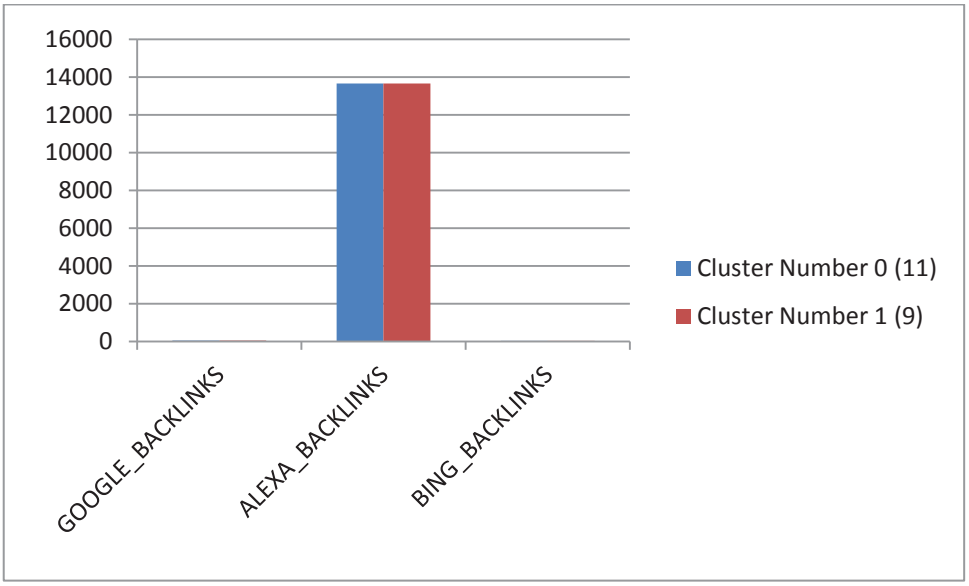


Figure 34: A bar-graph information of Google Backlinks, Alexa Backlinks and Bing Backlinks in terms of the HIC sites and the LIC sites.

4.2.4 Website Ranks and Scores

(4) In Table 23 above, there are following two attributes in terms of website ranks and scores: GOOGLE_PAGERANK and ALEXA_RANK.

The results criteria output from Table 23 indicates that the percentages for Cluster Number 1 over Cluster Number 0 of these two particular attributes are shown as 128% (i.e. >) and 11% (i.e. <<) respectively. Hence, this output implies that the overall value of Google Pagerank regarding the HIC sites is 128% greater than the overall value of Google Pagerank for the LIC sites. In addition, this same output from Table 23 above also implies that the overall value of Alexa Rank regarding the HIC sites is 11% much less than the overall value of Alexa Rank for the LIC sites. Figure 35 below displays bar-graph information of website ranks and scores (i.e. Google Page-rank and Alexa Rank) in terms of the HIC sites and the LIC sites. As illustrated in Figure 35, it shows that the HIC sites have much less values of

Alexa Rank compared to the LIC sites. Furthermore, the HIC sites have slightly greater value of Google Page-rank (i.e. 5) compared to Google Page-rank value of the LIC sites (i.e. 3.9091) according to our final clustering output as previously shown in Table 23.

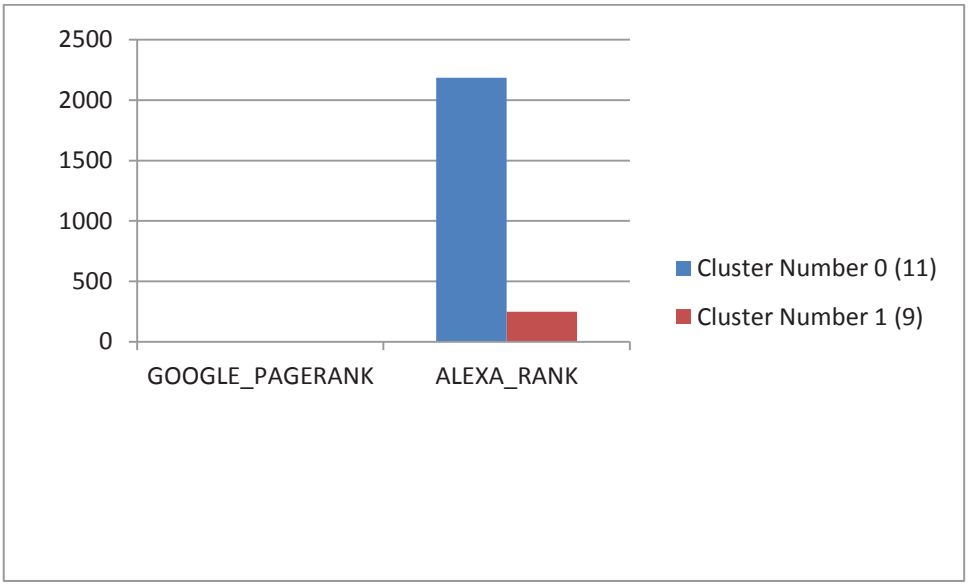


Figure 35: A bar-graph information of Google Page-rank and Alexa Rank in terms of the HIC sites and the LIC sites.

4.2.5 Location Latitude and Location Longitude

(5) In Table 23 above, there are following two attributes in terms of web server information: LOCATION_LATITUDE and LOCATION_LONGITUDE.

The results criteria output from Table 23 above shows that the percentages for Cluster Number 1 over Cluster Number 0 of these two particular attributes are indicated as 89% (i.e. <) and 189% (i.e. >>) respectively. Therefore, this output implies that the overall value of Location Latitude regarding the HIC sites is 89% less than the overall value of Location Latitude for the LIC sites. Furthermore, this output also implies that the overall value of Location Longitude regarding the HIC sites is 189% much greater than the overall value of Location Longitude for the LIC sites. Figure 36 below displays comprehensive bar-graph information of web server information (i.e. Location Latitude and Location Longitude) in terms of the HIC sites and the LIC sites. As illustrated in Figure 36, it shows that the HIC sites have less value of Location Latitude (i.e. 42.5829) compared to the LIC sites (i.e. 48.0589). On the other hands, Figure 36 also indicates that the HIC sites have much greater value of Location Longitude (i.e. -38.6622) compared to the LIC sites (i.e. -20.4453).

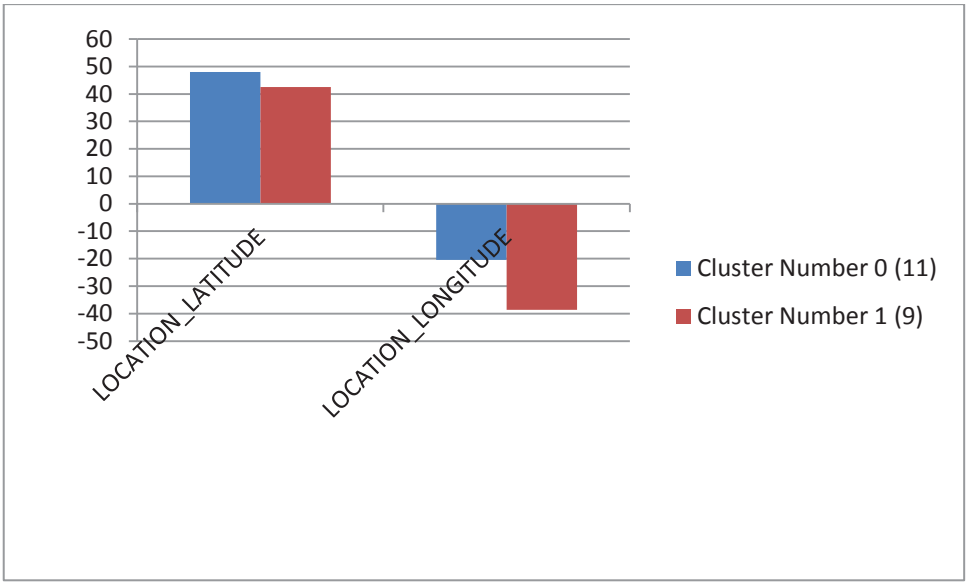


Figure 36: A bar-graph information of Location Latitude and Location Longitude in terms of the HIC sites and the LIC sites.

4.2.6 Social Network Engagement

(6) In Table 23 above, there are following seven main attributes in terms of social network engagement: FACEBOOK_SHARES, FACEBOOK_LIKES, FACEBOOK_COMMENTS, TWITTER_COUNT, LINKEDIN_SHARES, DELICIOUS_SHARES and GOOGLE_PLUS.

The results criteria output from Table 23 above shows that the percentages for Cluster Number 1 over Cluster Number 0 of these seven particular attributes are indicated as 2980% (i.e. >>>), 168% (i.e. >>), 165% (i.e. >>), 620% (i.e. >>), 162% (i.e. >>), 111% (i.e. >) and 158% (i.e. >>) respectively. Hence, this output implies that the overall number of Facebook Shares regarding the HIC sites is 2980% extremely much greater than the overall number of Facebook Shares for the LIC sites. This output also implies that the overall number of both Facebook Likes and Facebook Comments regarding the HIC sites are 168% and 165% much greater than the overall number of Facebook Likes and Facebook Comments for the LIC sites respectively. Furthermore, this criteria output also implies that the overall number of Twitter Count for the HIC sites is 620% much greater than the LIC sites. On the other hands, this criteria output shows that the overall number of both LinkedIn Shares and Google Plus for the HIC sites are 162% and 158% much greater than the overall number of both LinkedIn Shares and Google Plus for the LIC sites respectively compared to the criteria output for the case of Delicious Shares. (i.e. which basically shows that the overall number of Delicious Shares for the HIC sites is 111% greater than the overall number of Delicious Shares for the LIC sites). Figure 37 below displays comprehensive bar-graph information of various social network engagement or social media shares in terms of the HIC sites and the LIC sites. As illustrated in Figure 37, it shows that the HIC sites clearly have much greater values of Facebook Shares, Facebook Likes, Twitter Count and Google Plus compared to these four

same social media attributes of the LIC sites. Furthermore, Figure 37 also shows that the HIC sites have greater values of Facebook Comments, LinkedIn Shares and Delicious Shares compared to the LIC sites.

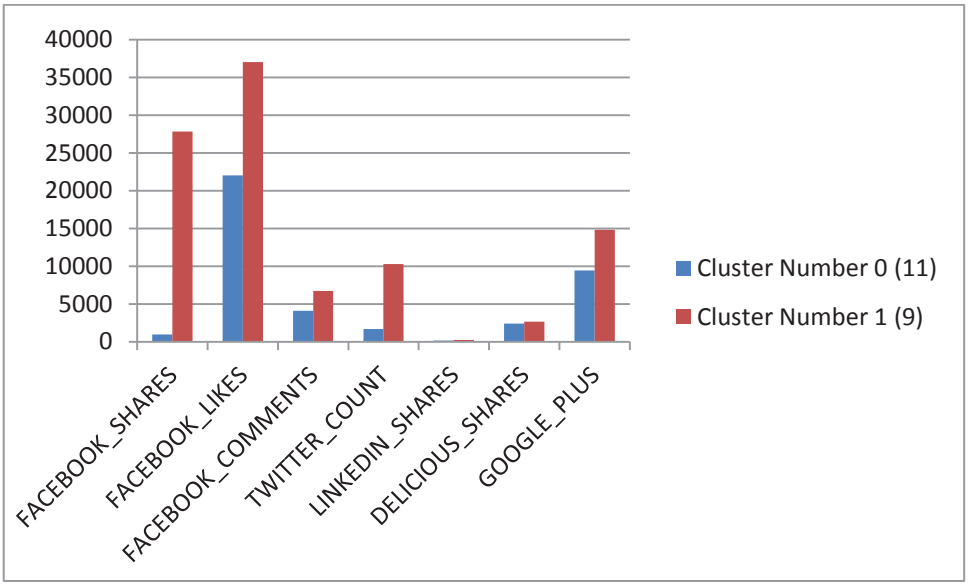


Figure 37: A comprehensive bar-graph information of social network engagement in terms of HIC sites and LIC sites.

4.2.7 Search Engine Indexes

(7) In Table 23 above, there are following three main attributes in terms of search engine indexes: GOOGLE_INDEXED_PAGES, YAHOO_INDEXED_PAGES and

BING_INDEXED_PAGES.

The results criteria output from Table 23 as illustrated previously indicates that the percentages for Cluster Number 1 over Cluster Number 0 of these three attributes are shown as equal to 100% (i.e. =) respectively. Hence, this output based on criteria in Table 23 implies that the overall number of search engine indexes including Google Indexed Pages, Yahoo Indexed Pages and Bing Indexed Pages for the HIC sites is equal to the overall number of Google Indexed Pages, Yahoo Indexed Pages and Bing Indexed Pages for the LIC sites.

4.2.8 Page Resources Breakdown

(8) In Table 23, there are following six main attributes in terms of page resources breakdown in Kilobytes (i.e. KB): JAVASCRIPT_KILOBYTES, HTML_KILOBYTES, IMAGES_KILOBYTES, CSS_KILOBYTES, KILOBYTES_TEXT and OTHER_KILOBYTES.

The results criteria output from Table 23 shows that the percentages for Cluster Number 1 over Cluster Number 0 of these six particular attributes are indicated as 64% (i.e. <), 57% (i.e. <), 101% (i.e. >), 107% (i.e. >), 100% (i.e. =) and 42% (i.e. <<) respectively. Therefore, this output implies that the overall page breakdown resources used in both JavaScript and HTML for the HIC sites are 64% and 57% less than the overall page breakdown resources used in JavaScript and HTML for the LIC sites respectively. This output based on criteria also implies that the overall page breakdown resources used in both Images and Cascading Style Sheet (CSS) for the HIC sites are 101% and 107% greater than the overall page breakdown resources used in Images and CSS for the LIC sites respectively. Furthermore, this output from Table 23 implies that the overall page breakdown resources used in Text (in Kilobytes) for the HIC sites is equal to the overall page breakdown resources used in Text for the LIC sites. Finally, this output also implies that the overall page breakdown resources used in other types of resources (in Kilobytes) for the HIC sites is 42% much less than the overall page breakdown resources used in other types of resources (in Kilobytes) for the LIC sites. Figure 38 below displays comprehensive bar-graph information of various page resources breakdown in Kilobytes in terms of the HIC sites and the LIC sites. As illustrated in Figure 38, it confirms that the HIC sites clearly have significantly less value of both JavaScript and HTML compared to the LIC sites. Furthermore, Figure 38 also shows that the HIC sites have much less value of other types of page resources in Kilobytes (i.e. 42% much less) compared to the LIC sites.

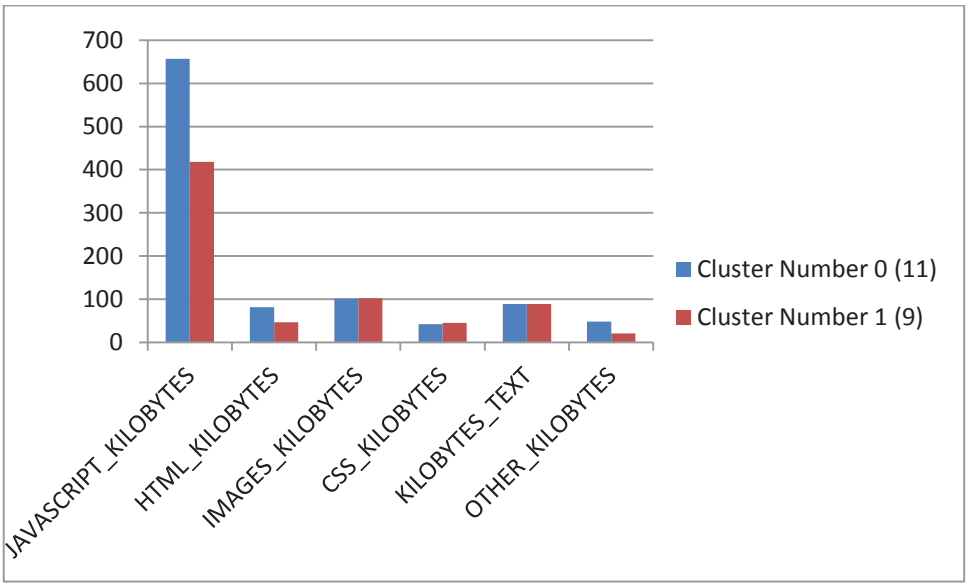


Figure 38: A comprehensive bar-graph information of various page resources breakdown in Kilobytes in terms of the HIC sites and the LIC sites.

4.2.9 Homepage Links Analysis

(9) In Table 23, there are following four attributes in terms of homepage links analysis:
TOTAL_LINKS_OF_INTERNAL_LINKS,
NO_FOLLOW_LINKS_OF_INTERNAL_LINKS,
TOTAL_LINKS_OF_EXTERNAL_LINKS and
NO_FOLLOW_LINKS_OF_EXTERNAL_LINKS.

The results criteria output from Table 23 shows that the percentages for Cluster Number 1 over Cluster Number 0 of these four particular attributes are indicated as 157% (i.e. >>), 152% (i.e. >>), 147% (i.e. >) and 183% (i.e. >>) respectively. Hence, this output implies that the total number of links from their internal links for the HIC sites is 157% much greater than the total number of links from their internal links for the LIC sites. This output also implies that the overall number of no-follow links from their internal links for the HIC sites is 152% much greater than the overall number of no-follow links from their internal links for the LIC sites. Furthermore, the total number of links from their external links for the HIC sites is 147% greater than the total number of links from their external links for the LIC sites. Finally, this output from Table 23 also implies that the total number of no-follow links from their external links for the HIC sites is 183% much greater than the total number of no-follow links from their external links for the LIC sites. Figure 39 below displays comprehensive bar-graph information of homepage links analysis (e.g. total number of links from the internal links, total number of no-follow links from the external links etc.) in terms of the HIC sites and the LIC sites.

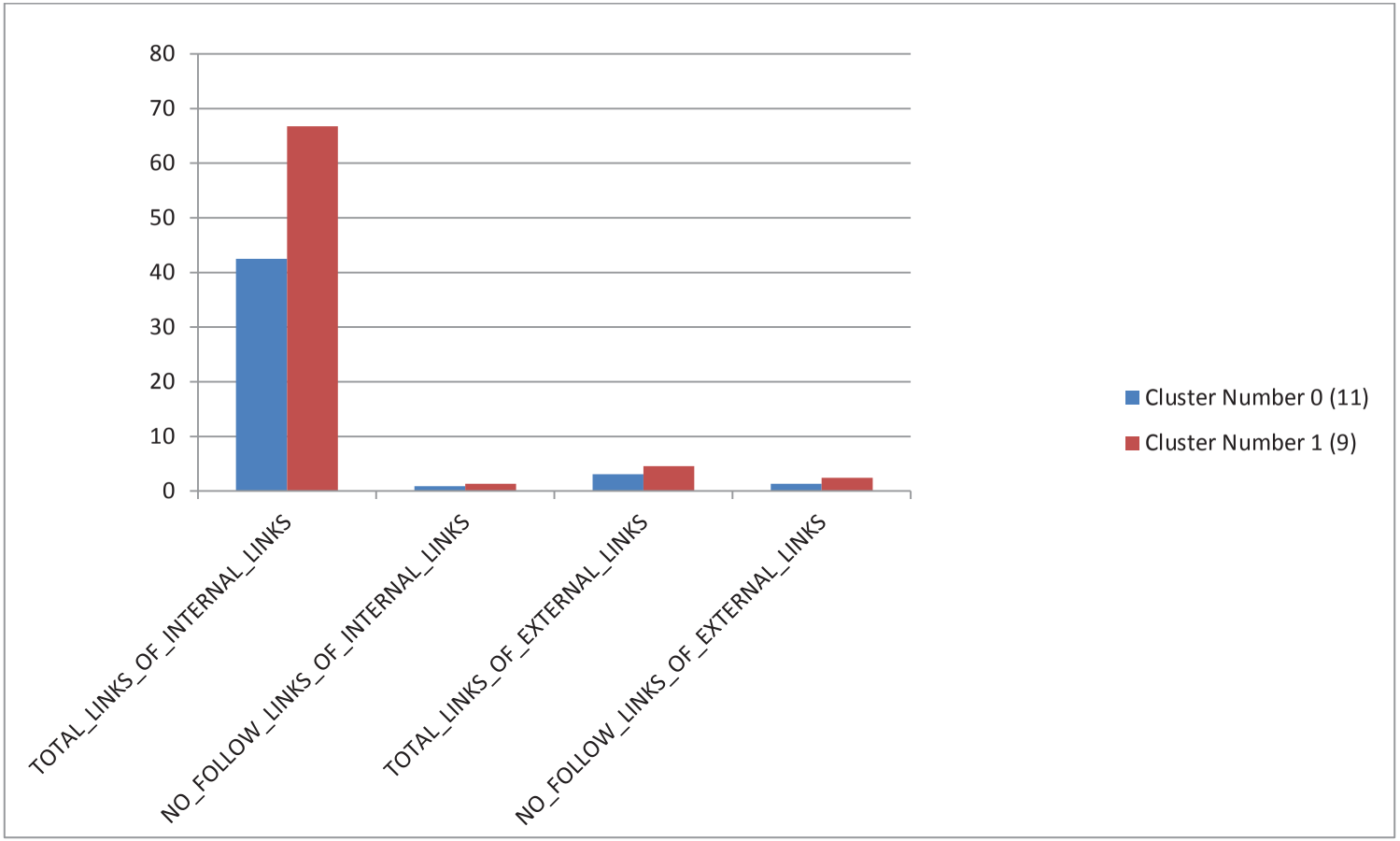


Figure 39: A comprehensive bar-graph information of homepage links analysis in terms of the HIC sites and the LIC sites.

As illustrated in Figure 39 above, it clearly shows that the HIC sites have much greater number of the following 4 links respectively compared to the LIC sites:

(a) TOTAL_LINKS_OF_INTERNAL_LINKS (i.e. the HIC sites have 157% much greater than the LIC sites).
(b) NO_FOLLOW_LINKS_OF_INTERNAL_LINKS (i.e. the HIC sites have 152% much greater than the LIC sites).
(c) TOTAL_LINKS_OF_EXTERNAL_LINKS (i.e. the HIC sites have 147% greater than the LIC sites).
(d) NO_FOLLOW_LINKS_OF_EXTERNAL_LINKS (i.e. the HIC sites have 183% much greater than the LIC sites).

4.2.10 Online Safety Information

(10) In Table 23, there are also following five attributes in terms of online safety information:

- WOT_TRUSTWORTHINESS (WEB OF TRUST TRUSTWORTHINESS)
- WOT_PRIVACY (WEB OF TRUST PRIVACY)
- WOT_CHILD_SAFETY (WEB OF TRUST CHILD SAFETY)
- GOOGLE_SAFE_BROWSING
- SITEADVISOR_RATING

The most significant result from Table 23 clearly shows that the safety ratings of these three particular attributes (i.e. WOT_TRUSTWORTHINESS, WOT_PRIVACY and WOT_CHILD_SAFETY) for the HIC sites are all indicated as “Excellent”. On the other hand, the safety ratings of these three same attributes for the LIC sites are all indicated as “Very Poor”.

4.3 Summary

In this chapter of the thesis, we found the following key findings:

- There are two groups of rogue websites, the low income and high income. In the entire process of this experiment, the k-means clustering algorithm has been used and implemented in order to identify two very distinct clusters: one associated with high income (HIC), cluster 1, and low income (LIC), cluster 0. We also found that there were 9 sites in the LIC and 11 sites in the HIC.
- According to our k-means clustering output, it demonstrates that the LIC sites are visited much less than the HIC sites and the HIC sites also have much more daily page-views on each sites compared to the daily page-views of the LIC sites.
- This chapter also demonstrates that the HIC sites have much greater number of both internal and external links (i.e. including the normal links and no-follow links) compared to the LIC sites.
- The HIC sites have much less data size of page resources used in JavaScript and HTML compared to the LIC sites. However, the HIC sites have greater data size of page resources used in CSS compared to the LIC sites. Furthermore, the HIC sites

have much less data size of page resources based on other types of page resources in Kilobytes compared to the LIC sites.

- The HIC sites have much greater values of social network engagement, especially Facebook Shares, Facebook Likes, Twitter Count, LinkedIn Shares and Google Plus etc. compared to these five same social media shares of the LIC sites. In addition, the HIC sites also have greater number of social network engagement such as Facebook Comments and Delicious Shares etc. compared to these two particular social media shares of the LIC sites.
- The HIC sites have less value of Location Latitude (i.e. 42.5829) compared to the LIC sites (i.e. 48.0589). On the other hands, this chapter demonstrates that the HIC sites have much greater value of Location Longitude (i.e. -38.6622) compared to the LIC sites (i.e. -20.4453).
- The HIC sites have an "Excellent" safety rating in terms of the following components:
  - Web of Trust (WOT) Trustworthiness
  - WOT Privacy
  - WOT Child Safety

  On the other hands, the LIC sites have a "Very Poor" safety rating in terms of these three particular components as shown above.
- The total number of Google Backlinks for the HIC sites has 139% greater than the total number of Google Backlinks for the LIC sites. On the other hands, the total number of Bing Backlinks for the HIC sites has 85% less than the total number of Bing Backlinks for the LIC sites. Furthermore, the overall number of Alexa Backlinks regarding the HIC sites is exactly equal to the overall number of Alexa Backlinks for the LIC sites.
- The overall value of Google Pagerank for the HIC sites have 128% greater than the overall value of Google Pagerank for the LIC sites. On the other hands, the overall value of Alexa Rank for the HIC sites have 11% much less than the overall value of Alexa Rank for the LIC sites.

CHAPTER FIVE

5.1 Conclusion and Discussion

In conclusion, we found the following significant findings in this particular chapter of the thesis as below:

- In this research, there are basically two groups of rogue websites, the HIC sites and the LIC sites.
- Low income rogue websites are visited much less than high income rogue websites in terms of daily unique visitors (i.e. 885% much greater for the HIC sites over the LIC sites) according to our dataset. It shows that the high income rogue websites also have much more daily page-views (i.e. 885% much greater for the HIC sites over the LIC sites).
- The HIC sites have much greater number of both internal and external links basically containing the normal links and no-follow links compared to the LIC sites. For instance, the total number of links from their internal links for the HIC sites is 157% much greater than the total number of links from their internal links for the LIC sites. Furthermore, the total number of both links and no-follow links from their external links for the HIC sites are shown as 147% and 183% respectively much greater than the total number of both links and no-follow links from their external links for the LIC sites.
- The HIC sites have much less data size of page breakdown resources used in JavaScript and HTML (i.e. 64% and 57% respectively) compared to the data size of page breakdown resources used in JavaScript and HTML from the LIC sites. The HIC sites also have 107% greater data size of page breakdown resources used in CSS compared to the LIC sites. Furthermore, HIC sites have 42% much less data size of other types of page resources in Kilobytes compared to the LIC sites.
- The HIC sites have much greater resulting values in terms of social network engagement (or social media shares) such as Facebook Shares, Facebook Likes, Facebook Comments, Twitter Count, LinkedIn Shares, Delicious Shares and Google Plus etc. compared to the resulting values of the LIC sites. For instance, the HIC sites have 2980% extremely much greater Facebook Shares than the LIC sites and the HIC sites also have 620% much greater Twitter Count than the LIC sites. In addition, it shows that the HIC sites have the following results: 168% much greater Facebook Likes, 165% much greater Facebook Comments, 162% much greater LinkedIn Shares, 111% greater Delicious Shares and 158% much greater Google Plus respectively compared to these five particular social media shares of the LIC sites.

- The HIC sites have a safety rating or standard as "Excellent" in terms of the following key components: (a) WOT Trustworthiness, (b) WOT Privacy and (c) WOT Child Safety. On the other hands, the LIC sites have a safety rating or standard as "Very Poor" in terms of these particular three components. Hence, it implies that the HIC sites generally have a robust security protection system to ensure that WOT trustworthiness, WOT privacy and WOT child safety of the HIC sites are all well-maintained and securely protected. However, it indicates that the LIC sites are observed to be vulnerable to the potential exposures or threats to maintain the protection from these particular safety rating components (i.e. WOT Trustworthiness, WOT Privacy and WOT Child Safety).
- The HIC sites have 89% less value of Location Latitude (i.e. 42.5829) compared to Location Latitude of the LIC sites (i.e. 48.0589). On the other hands, the corresponding result from the k-means clustering output (i.e. Table 23) demonstrates that the HIC sites have 189% much greater value of Location Longitude (i.e. -38.6622) compared to Location Longitude of the LIC sites (i.e. -20.4453).
- The total number of Google Backlinks for the HIC sites (i.e. 49.3333) has 139% greater than the total number of Google Backlinks for the LIC sites (i.e. 35.4805). On the other hands, the total number of Bing Backlinks for the HIC sites (i.e. 23.7778) has 85% less than the total number of Bing Backlinks for the LIC sites (i.e. 28.0455). In addition, the overall number of Alexa Backlinks regarding the HIC sites is exactly equal to the overall number of Alexa Backlinks for the LIC sites.
- The overall values of Google Pagerank for the HIC sites (i.e. 5) have 128% greater than the overall values of Google Pagerank for the LIC sites (i.e. 3.9091). On the other hands, the overall values of Alexa Rank for the HIC sites (i.e. 247) have 11% much less than the overall values of Alexa Rank for the LIC sites (i.e. 2186).
- According to the descriptive statistics output from Table 17, it implies that these two particular variables such as ALEXA_RANK and FACEBOOK_SHARES have the following large values of skewness respectively: 4.431197 and 4.08799. Hence, it confirms that these two large significant skewness values can be clearly observed or considered as a main key factor to influence to the mean which is greater than the median.
- According to the descriptive statistics output from Table 17, it also implies that these particular variables such as ALEXA_RANK and FACEBOOK_SHARES with extremely large kurtosis values (i.e. 19.73941 and 17.35708 respectively) can be observed to have a distinct peak near the mean as well as having heavy tails at the same time, in relation with information from Fernandez (2010). On the other hands, the same descriptive statistics output from Table 17 also implies that these five particular variables such as HTML_KILOBYTES (0.018039), JAVASCRIPT_KILOBYTES (0.673957), LOCATION_LONGITUDE (-0.32828), NO_FOLLOW_LINKS_OF_EXTERNAL_LINK (0.510016) and DAILY_UNIQUE_VISITORS (0.544024) respectively with low kurtosis values can be observed to have a flat top near the mean rather than a sharp peak near the mean (e.g. ALEXA_RANK and FACEBOOK_SHARES), in relation with information from Fernandez (2010). In addition, it shows that the distributions of these particular

variables with positive kurtosis values, except for this variable with negative kurtosis value (LOCATION_LONGITUDE: -0.32828) can be considered or observed to have heavy tails mostly.

The purpose of analysing the data in this was to answer three research questions in this thesis. It was found that, as predicted, there were two natural clusters of the most complained about sites (high income and low income). This means that rightsholders should focus their efforts and resources on only high income sites, and ignore the others.

It was also found that the main significant factors or key critical variables for separating high-income vs low-income rogue websites included daily page-views, number of internal and external links, social media shares (i.e. social network engagement) and element of the page structure, including HTML page and JavaScript sizes. Further research should investigate why these factors were important in driving website revenue higher. For example, why is high revenue associated with smaller HTML pages and less JavaScript? Is it because the pages are simply faster to load? A similar pattern is observed with the number of links. These results could form a study looking into what attributes make e-commerce successful more broadly.

It is important to note that this was a preliminary study only looking at the Top 20 rogue websites basically suggested by Google Transparency Report (2015). Whilst these account for the majority of complaints, a different picture may emerge if we analysed more sites, and/or selected them based on different sets of criteria, such the time period, geographic location, content category (software versus movies, for example), and so on.

5.2 Summary

This chapter demonstrates that the following significant factors or key critical variables are important in order to determine between high-income vs low income rogue websites in general:

- Low-income rogue websites are visited much less than high-income rogue websites in terms of daily unique visitors. Furthermore, the high-income rogue websites have much greater number of daily page-views compared to the low-income rogue websites. (i.e. 885% much greater for the HIC sites over the LIC sites in terms of both daily unique visitors and daily page-views).
- The high-income rogue websites have much greater number of both internal and external links (i.e. including the normal links and no-follow links from both internal and external links) compared to the internal and external links of the low-income rogue websites.
- The high-income rogue sites have much less data size of page breakdown resources used in JavaScript and HTML compared to the low-income sites ((i.e. 64% and 57% respectively). The high-income rogue sites have 107% greater data size of page resources used in CSS compared to the low-income rogue sites.

Furthermore, the high-income rogue websites have 42% much less data size of other types of page resources in Kilobytes compared to the low-income rogue websites.

- The high-income rogue websites have much greater resulting values in terms of various social network engagement (or social media shares) such as Facebook Shares (i.e. 2980% extremely much greater), Facebook Likes (i.e. 168% much greater), Facebook Comments (i.e. 165% much greater), Twitter Count (i.e. 620% much greater), LinkedIn Shares (i.e. 162% much greater), Delicious Shares (i.e. 111% greater) and Google Plus (i.e. 158% much greater) etc. compared to these particular same social media shares of the low-income rogue websites.
- The high-income rogue websites have 89% less value of Location Latitude (i.e. 42.5829) compared to Location Latitude of the low-income rogue websites (i.e. 48.0589). On the other hands, demonstrates that the high-income rogue websites have 189% much greater value of Location Longitude (i.e. -38.6622) compared to Location Longitude of the low-income rogue websites (i.e. -20.4453).
- The high-income rogue websites have indicated an “Excellent” safety rating or standard in terms of the following: (a) WOT Trustworthiness, (b) WOT Privacy and (c) WOT Child Safety. On the other hands, the low-income rogue websites have indicated a “Very Poor” safety rating or standard in terms of these three components as above.

CHAPTER SIX

Future Work

Once all our implementations for the data mining analytics methods or techniques (i.e. cluster analysis etc.) used in WEKA 3.6 Explorer environment have been set as previously mentioned and described in Chapter 3 (i.e. Research Methodology) and Chapter 4 (i.e. Experiments and Results), the appropriate steps and use of the most common and innovative data mining application such as "WEKA 3.6" platform can be undertaken and executed further to improve the sustainable, well-integrated and much reliable cyber security environment effectively. These fast-growing days with technology and industry via the Internet and ubiquitous world, we all know that the current reality of rogue websites (or pirate websites) operation and its increasing trend to the modern e-commerce environment keeps increasing and bringing some related outcomes and influences proportionally to our public and even in the broad cyber security industrial sectors and governmental agencies in New Zealand and around the world, according to the related-contents and comprehensive processes of previous discussion (i.e. Chapter Five), literature review (i.e. Chapter Two), research methodology (i.e. Chapter Three), introduction (i.e. Chapter One) and overall data-mining analysis (i.e. cluster analysis) through WEKA 3.6 Explorer.

Furthermore, the following further work can be carried out for the future expansion in this particular research:

- While this research of the thesis has been designed as a preliminary study only looking at the Top 20 rogue websites which basically implies that these account for the majority of complaints, a different statistical outcome or significant indication in terms of statistics may emerge if we analysed more various rogue websites rather than the Top 20 websites in terms of the future work.
- While this research has been done based on the timeframe of "All Available" which is suggested by Google Transparency Report (2015), a future research work can be improved and implemented effectively in terms of the following different timeframes: (a) Past Week, (b) Past Month and (c) Past Year etc.
- Many different significant pictures would be possible to emerge if various rogue websites are analysed in terms of primarily geographic location (i.e. several continents or regions) or variety types of content category (i.e. software vs movies, software vs music etc.) in more detail from the standpoint of future work.
- While this research of the thesis has been analysed mainly based on the following data mining technique (i.e. cluster analysis), a further work can be designed to generate the different significant outcomes or results if we analyse the primary data sets by utilising the implementations of using different techniques such as regression analysis, decision tree analysis or neural networks if these analytics are appropriate.

References

Abbott, J. (Ed.). (2004). *The political economy of the Internet in Asia and the Pacific: digital divides, economic competitiveness, and security challenges*. Praeger Publishers.

Adermon, A., & Liang, C. Y. (2014). Piracy and music sales: The effects of an anti-piracy law. *Journal of Economic Behavior & Organization*, *105*, 90-106.

Agarwal, A., Hosanagar, K., & Smith, M. D. (2011). Location, location, location: An analysis of profitability of position in online advertising markets. *Journal of marketing research*, *48*(6), 1057-1073.

Ahlemeyer-Stubbe, A., & Coleman, S. (2014). *A Practical Guide to Data Mining for Business and Industry*. John Wiley & Sons.

Ajzen, I., & Fishbein, M. (1980). Understanding attitudes and predicting social behaviour. *Englewood Cliffs, NJ: Prentice-Hall.*

Alazab, M., Venkatraman, S., Watters, P., Alazab, M., & Alazab, A. (2012). Cybercrime: the case of obfuscated malware. In *Global Security, Safety and Sustainability & e-Democracy* (pp. 204-211). Springer Berlin Heidelberg.

Al-Rafee, S., & Rouibah, K. (2010). The fight against digital piracy: An experiment. *Telematics and Informatics*, *27*(3), 283-292.

An, S., Jin, H. S., & Park, E. H. (2014). Children's advertising literacy for advergames: perception of the game as advertising. *Journal of Advertising*,*43*(1), 63-72.

An, S., & Kang, H. (2013). Do online ad breaks clearly tell kids that advergames are advertisements that intend to sell things?. *International Journal of Advertising*, *32*(4), 655-678.

Armstrong, S. (2001). *Advertising on the Internet: how to get your message across on the World Wide Web*. Kogan Page Publishers.

Arthur, C. (2011). Sony suffers second data breach with theft of 25m more user details. *Guardian UK, May 3, 2011*. Retrieved July 27, 2015 from the WorldWideWeb: http:// www.guardian.co.uk/technology/blog/2011/may/03/sony-data-breach-online-entertainment

Asdemir, K., Kumar, N., & Jacob, V. S. (2012). Pricing models for online advertising: CPM vs. CPC. *Information Systems Research*, *23*(3-part-1), 804-822.

Auger, N., Daniel, M., Knäuper, B., Dourian, T., & Raynault, M. F. (2015). Unintended messages in online advertising to youth: illicit drug imagery in a Canadian sports marketing campaign. *Journal of Adolescent Health*, *56*(4), 429-432.

Austin, J. M., & Reed, M. L. (1999). Targeting children online: Internet advertising ethics issues. *Journal of Consumer Marketing*, *16*(6), 590-602.

Azoulay, J. (1998). Is online on the line? Kid-based websites. *Children's Business*, *13*(6), 3-9.

Bae, S. H., & Choi, J. P. (2006). A model of piracy. *Information Economics and Policy*, *18*(3), 303-320.

Berenson, M., Levine, D., & Krehbiel, T. C. (2009). *Basic business statistics: Concepts and applications*. Pearson Education International.

Berenson, M., Levine, D., Szabat, K. A., & Krehbiel, T. C. (2012). *Basic business statistics: Concepts and applications*. Pearson Higher Education AU.

Blades, M., Oates, C., & Li, S. (2013). Children's recognition of advertisements on television and on Web pages. *Appetite*, *62*, 190-193.

Bogdanoski, M., & Petreski, D. (2013). Cyber terrorism–global security threat.*Contemporary Macedonian Defense-International Scientific Defense, Security and Peace Journal*, *13*(24), 59-73.

Boyer, W., & McQueen, M. (2008). Ideal based cyber security technical metrics for control systems. In *Critical Information Infrastructures Security* (pp. 246-260). Springer Berlin Heidelberg.

Bozdogan, H. (Ed.). (2003). *Statistical data mining and knowledge discovery*. CRC Press.

Broadhurst, R., Grabosky, P., Alazab, M., Bouhours, B., & Chon, S. (2014). An Analysis of the Nature of Groups Engaged in Cyber Crime. *An Analysis of the Nature of Groups engaged in Cyber Crime, International Journal of Cyber Criminology*, *8*(1), 1-20.

Cai, X., & Zhao, X. (2013). Online advertising on popular children's websites: Structural features and privacy issues. *Computers in Human Behavior*, *29*(4), 1510-1518.

Carter, O. B., Patterson, L. J., Donovan, R. J., Ewing, M. T., & Roberts, C. M. (2011). Children's understanding of the selling versus persuasive intent of junk food advertising: Implications for regulation. *Social Science & Medicine*, *72*(6), 962-968.

Cavelty, M. D. (2008). Cyber-terror—looming threat or phantom menace? The framing of the US cyber-threat debate. *Journal of Information Technology & Politics*, *4*(1), 19-36.

Chen, J., & Stallaert, J. (2014). An economic analysis of online advertising using behavioral targeting. *Mis Quarterly*, *38*(2), 429-449.

Cherkassky, V., & Mulier, F. M. (2007). *Learning from data: concepts, theory, and methods*. John Wiley & Sons.

Chilling Effects. (2015). About chilling effects. *Chilling Effects Electronic Frontier Foundation*. Retrieved July 27, 2015 from the WorldWideWeb: https://chillingeffects.org/pages/about

Choo, K. K. R. (2011). The cyber threat landscape: Challenges and future research directions. *Computers & Security*, *30*(8), 719-731.

Collica, R. (2011). *Customer segmentation and clustering using SAS Enterprise Miner*. SAS Institute.

Conklin, A., White, G., Cothren, C., Williams, D., & Davis, R. L. (2004). *Principles of computer security: security+ and beyond*. McGraw-Hill, Inc..

Coopers, P. (2013). IAB Internet Advertising Revenue Report: 2013 First Six Months' Results.

Coopers, P. W. H. (2014). IAB internet advertising revenue report. *URL http://www. iab. net/insights_research/industry_data_ and_landscape/adrevenuereport*.

CuteStat.com. (2015). Website Statistics and Website Valuation. *CuteStat.com*. Retrieved August 11, 2014 from the WorldWideWeb: http://www.cutestat.com/

Dehghani, M., & Tumer, M. (2015). A research on effectiveness of Facebook advertising on enhancing purchase intention of consumers. *Computers in Human Behavior*, *49*, 597-600.

Denes, S. (2001). Thumbs down for pop-ups. *Rural Telecommunications 20*, 4 (July/August., 2001), 9.

Department of Computer Science of The University of Waikato. (2008). Attribute-Relation File Format (ARFF). *Department of Computer Science of The University of Waikato*. Retrieved August 11, 2014 from the WorldWideWeb: http://www.cs.waikato.ac.nz/ml/weka/arff.html

Dou, W., Lim, K. H., Su, C., Zhou, N., & Cui, N. (2010). Brand positioning strategy using search engine marketing. *Mis Quarterly*, *34*(2), 261-279.

Douglas, T., & Loader, B. (2000). *Cybercrime: Law enforcement, security and surveillance in the information age*. Psychology Press.

Eastman, S. T. (Ed.). (2000). *Research in media promotion*. Routledge.

Economides, N. (1993). Network economics with application to finance. *Financial markets, institutions & instruments*, *2*(5), 89-97.

Eisenmann, T. (2007). The Economics of Internet Advertising: Implications for the Google-DoubleClick Merger. *Presentation for AEI-Brookings Joint Center*, July.

Ericsson, G. N. (2010). Cyber security and power system communication—essential parts of a smart grid infrastructure. *Power Delivery, IEEE Transactions on*, *25*(3), 1501-1507.

Evans, D. S. (2008). The economics of the online advertising industry. *Review of network economics*, *7*(3).

Evans, D. S. (2009). The online advertising industry: Economics, evolution, and privacy. *Journal of Economic Perspectives, Forthcoming*.

Everitt, B. (1980). *Cluster Analysis*, (Second Edition). Halsted Press.

Everitt, B. (1993). *Cluster Analysis*, (Third Edition). Halsted Press.

Fernandez, G. (2010). *Statistical data mining using SAS applications*. CRC Press.

Fishbein, M., & Ajzen, I. (1975). Belief, attitude, intention and behavior: An introduction to theory and research. *Reading, Massachusetts: Addison-Wesley*.

Francis, L. (2005). Data Mining: Concepts, Models, Methods and Algorithms, Mehmed Kantarzic, Paperback, IEEE Press/Wiley, 2001, xii+ 345 pages.

Furnell, S. (2008). *Securing information and communications systems: Principles, technologies, and applications*. Artech House.

Gainsbury, S. M., Russell, A., Hing, N., Wood, R., & Blaszczynski, A. (2013). The impact of internet gambling on gambling problems: A comparison of moderate-risk and problem Internet and non-Internet gamblers. *Psychology of Addictive Behaviors*, *27*(4), 1092.

Garbade, K. D., & Silber, W. L. (1979). Structural organization of secondary markets: Clearing frequency, dealer activity and liquidity risk. *Journal of Finance*, 577-593.

Garson, G. D. (1995). *Computer technology and social issues*. IGI Global.

Goldfarb, A., & Tucker, C. (2011). Online display advertising: Targeting and obtrusiveness. *Marketing Science*, *30*(3), 389-404.

Goldfarb, A., & Tucker, C. E. (2011). Privacy regulation and online advertising. *Management Science*, *57*(1), 57-71.

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

Good, P. (2011). *A practitioner's guide to resampling for data analysis, data mining, and modeling*. Chapman & Hall/CRC.

Goodrich, K. (2010). What's up? Exploring upper and lower visual field advertising effects. *Journal of Advertising Research*.

Google Transparency Report. (2015). Requests to remove content (Due to copyright). *Google Inc*. Retrieved July 27, 2015 from the WorldWideWeb: http:// http://www.google.com/transparencyreport/removals/copyright/

Google Transparency Report. (2015). Requests to remove content (Due to copyright), FAQ. *Google Inc*. Retrieved July 27, 2015 from the WorldWideWeb: http://www.google.com/transparencyreport/removals/copyright/faq/

Google Transparency Report. (2015). Requests to remove content (From governments). *Google Inc*. Retrieved July 27, 2015 from the WorldWideWeb: https://www.google.com/transparencyreport/removals/government/

Google Transparency Report. (2015). Specified domains. *Google Inc*. Retrieved July 27, 2015 from the WorldWideWeb: http://www.google.com/transparencyreport/removals/copyright/domains/?r=all-time

Google Transparency Report. (2013). Transparency Report. *Google Inc*.

Google Transparency Report. (2013). Transparency Report. *Google Inc*. Retrieved from the WorldWideWeb: http://www.google.com/transparencyreport/

Google Transparency Report. (2015). Transparency Report (Access to information). *Google Inc*. Retrieved July 27, 2015 from the WorldWideWeb: https://www.google.com/transparencyreport/?hl=en

Guha, S., Cheng, B., & Francis, P. (2010). Challenges in measuring online advertising systems. In *Proceedings of the 10th ACM SIGCOMM conference on Internet measurement* (pp. 81-87). ACM.

Guha, S., Cheng, B., & Francis, P. (2011, March). Privad: Practical Privacy in Online Advertising. In *NSDI*.

Gup, B. E. (2003). *Investing online*. Blackwell Publishing.

Ha, L., & McCann, K. (2008). An integrated model of advertising clutter in offline and online media. *International Journal of Advertising*, *27*(4), 569-592.

Hansen, L., & Nissenbaum, H. (2009). Digital disaster, cyber security, and the Copenhagen School. *International Studies Quarterly*, *53*(4), 1155-1175.

Haugtvedt, C. P., Machleit, K. A., & Yalch, R. (Eds.). (2005). *Online consumer psychology: understanding and influencing consumer behavior in the virtual world*. Psychology Press.

Herath, T., & Rao, H. R. (2009). Encouraging information security behaviors in organizations: Role of penalties, pressures and perceived effectiveness.*Decision Support Systems*, *47*(2), 154-165.

Hibbard, J. D., Kumar, N., & Stern, L. W. (2001). Examining the impact of destructive acts in marketing channel relationships. *Journal of Marketing Research*, *38*(1), 45-61.

Hiller, J. S., & Russell, R. S. (2013). The challenge and imperative of private sector cybersecurity: An international comparison. *Computer Law & Security Review*, *29*(3), 236-245.

Hua, J., & Bapna, S. (2013). The economic impact of cyber terrorism. *The Journal of Strategic Information Systems*, *22*(2), 175-186.

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.

Ilves, L. S. (2011). *Internet advertising and sales: Internet theory, technology and applications*. Nova Science Publishers, Inc.

InformationWeek. (2007). Google Launches Test of Pay-Per-Action Ads. *UBM Tech (2015).* Retrieved July 27, 2015 from the WorldWideWeb: http://www.informationweek.com/news/internet/search/showArticle.jhtml?articleID=198500474

Jackson, D. W. (2013). *Cybersecurity: Shared Risks, Shared Responsibilities*. Shane, P. M., & Hunker, J. A. (Eds.). Carolina Academic Press.

Jansen, J. (2011). *Understanding sponsored search: Core elements of keyword advertising*. Cambridge University Press.

Johnson, R. A., & Wichern, D. W. (1992). *Applied multivariate statistical analysis* (Vol. 4). Englewood Cliffs, NJ: Prentice hall.

Johnson, R. A., & Wichern, D. W. (1998). *Applied multivariate statistical analysis*. Englewood Cliffs, NJ: Prentice hall.

Johnson, R. A. (81). WICHERN, DW-1998-Applied multivariate statistical analysis. *Prentice-Hall, Inc., Englewood Cliffs, New Jersey*, *7632*, 594.

Johnson, V. R. (2005). Cybersecurity, Identity Theft, and the Limits of Tort Liability. *ScL REv.*, *57*, 255.

Kairouz, S., Paradis, C., & Nadeau, L. (2012). Are online gamblers more at risk than offline gamblers?. *Cyberpsychology, Behavior, and Social Networking*,*15*(3), 175-180.

Kantarzic, M. (2011). *Data Mining: Concepts, Models, Methods and Algorithms*. IEEE Press/A John Wiley & Sons, Inc., Publication

Kelsen, K. (2012). *Unleashing the power of digital signage: content strategies for the 5th screen*. CRC Press.

Kiema, I. (2008). Commercial piracy and intellectual property policy. *Journal of Economic Behavior & Organization*, *68*(1), 304-318.

Klapdor, S. (2013). *Effectiveness of Online Marketing Campaigns: An Investigation Into Online Multichannel and Search Engine Advertising*. Springer Science & Business Media.

Kritzinger, E., & von Solms, S. H. (2010). Cyber security for home users: A new way of protection through awareness enforcement. *Computers & Security*,*29*(8), 840-847.

Kshetri, N. (2005). Pattern of global cyber war and crime: A conceptual framework. *Journal of International Management*, *11*(4), 541-562.

Kunkel, D., Wilcox, B. L., Cantor, J., Palmer, E., Linn, S., & Dowrick, P. (2004). Report of the APA task force on advertising and children. *Washington, DC: American Psychological Association*.

Layton, R., Watters, P., & Dazeley, R. (2012, October). Unsupervised authorship analysis of phishing webpages. In *Communications and Information Technologies (ISCIT), 2012 International Symposium on* (pp. 1104-1109). IEEE.

Layton, R., Watters, P., & Ureche, O. (2013, November). Identifying Faked Hotel Reviews Using Authorship Analysis. In *Cybercrime and Trustworthy Computing Workshop (CTC), 2013 Fourth* (pp. 1-6). IEEE.

Lee, D. T., Shieh, S. P., & Tygar, D. (Eds.). (2005). *Computer security in the 21st century*. Springer Science & Business Media.

Lee, J., & Lee, M. (2011). Factors influencing the intention to watch online video advertising. *Cyberpsychology, Behavior, and Social Networking*, *14*(10), 619-624.

Lee, T. Y., & Bradlow, E. T. (2011). Automated marketing research using online customer reviews. *Journal of Marketing Research*, *48*(5), 881-894.

Lehr, W. H., & Pupillo, L. (Eds.). (2009). *Internet Policy and Economics: Challenges and Perspectives*. Springer Science & Business Media.

Lewis, T. G. (2014). *Critical infrastructure protection in homeland security: defending a networked nation*. John Wiley & Sons.

Lu, C. S. (Ed.). (2004). *Multimedia Security: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property: Steganography and Digital Watermarking Techniques for Protection of Intellectual Property*. Igi Global.

Lysonski, S., & Durvasula, S. (2008). Digital piracy of MP3s: consumer and ethical predispositions. *Journal of Consumer Marketing*, *25*(3), 167-178.

Ma, L., Yearwood, J., & Watters, P. (2009, September). Establishing phishing provenance using orthographic features. In *eCrime Researchers Summit, 2009. eCRIME'09.* (pp. 1-10). IEEE.

Maimon, O., & Rokach, L. (Eds.). (2005). *Data mining and knowledge discovery handbook* (Vol. 2). New York: Springer.

Marti-Pellón, D., & Saunders-Uchôa-Craveiro, P. (2015). Children's Exposure to Advertising on Games Sites in Brazil and Spain. *Comunicar*, *23*(45).

MathWorks. (2015). Statistics and Machine Learning Toolbox: Cluster Analysis. *The MathWorks, Inc*. Retrieved July 27, 2015 from the WorldWideWeb: http://au.mathworks.com/products/statistics/features.html

McCombie, S., Pieprzyk, J., & Watters, P. (2009). Cybercrime attribution: an Eastern European case study.

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

McStay, A. (2009). *Digital advertising*. Palgrave Macmillan.

McStay, A. (2011). *The mood of information: a critique of online behavioural advertising*. A&C Black.

Miner, G., Nisbet, R., & Elder IV, J. (2009). *Handbook of statistical analysis and data mining applications*. Academic Press.

Miyazaki, A. D., Stanaland, A. J., & Lwin, M. O. (2009). Self-regulatory safeguards and the online privacy of preteen children. *Journal of Advertising*,*38*(4), 79-91.

Mo, Y., Kim, T. H. J., 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.

Nandedkar, A., & Midha, V. (2012). It won't happen to me: An assessment of optimism bias in music piracy. *Computers in Human Behavior*, *28*(1), 41-48.

Navarro, J. N., Marcum, C. D., Higgins, G. E., & Ricketts, M. L. (2014). Addicted to pillaging in cyberspace: Investigating the role of internet addiction in digital piracy. *Computers in Human Behavior*, *37*, 101-106.

Newman, R. C. (2009). *Computer security: protecting digital resources*. Jones & Bartlett Publishers.

Ngai, E. W. T. (2003). Selection of web sites for online advertising using the AHP. *Information & Management*, *40*(4), 233-242.

Oger, M., Olmez, I., Inci, E., Kücükbay, S., & Emekci, F. (2015). Privacy Preserving Secure Online Advertising. *Procedia-Social and Behavioral Sciences*, *195*, 1840-1845.

Ohsawa, Y., & Yada, K. (Eds.). (2009). *Data mining for design and marketing*. CRC Press.

Olson, D. L., & Shi, Y. (2007). *Introduction to business data mining* (Vol. 10, pp. 2250-2254). Englewood Cliffs: McGraw-Hill/Irwin.

Parker, T., Sachs, M., Shaw, E., & Stroz, E. (2004). *Cyber adversary characterization: Auditing the hacker mind*. Syngress.

Peitz, M., & Waelbroeck, P. (2006). Piracy of digital products: A critical review of the theoretical literature. *Information Economics and Policy*, *18*(4), 449-476.

Previc, F. H. (1990). Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, *13*(03), 519-542.

Price, L. L., Arnould, E. J., & Tierney, P. (1995). Going to extremes: Managing service encounters and assessing provider performance. *The Journal of Marketing*, 83-97.

PRS for Music & Google. (2012). The six business models for copyright infringement. *BAE Systems plc*. Retrieved July 27, 2015 from the WorldWideWeb: http://www.prsformusic.com/aboutus/policyandresearch/researchandeconomics/Documents/TheSixBusinessModelsofCopyrightInfringement.pdf

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

Refaat, M. (2010). *Data preparation for data mining using SAS*. Morgan Kaufmann.

Rhee, M. Y. (2003). *Internet security: cryptographic principles, algorithms and protocols*. John Wiley & Sons.

Rinaldi, S. M., Peerenboom, J. P., & Kelly, T. K. (2001). Identifying, understanding, and analyzing critical infrastructure interdependencies. *Control Systems, IEEE*, *21*(6), 11-25.

Rowe, D. C., Lunt, B. M., & Ekstrom, J. J. (2011, October). The role of cyber-security in information technology education. In *Proceedings of the 2011 conference on Information technology education* (pp. 113-122). ACM.

Roy, A., Kim, D. S., & Trivedi, K. S. (2010, April). Cyber security analysis using attack countermeasure trees. In *Proceedings of the Sixth Annual Workshop on Cyber Security and Information Intelligence Research* (p. 28). ACM.

Russell Indexes. (2015). Russell 2000 Index. *Russell Investments*. Retrieved July 27, 2015 from the WorldWideWeb: http://www.russell.com/indexes/americas/indexes/fact-sheet.page?ic=US2000

Sarma, K. S. (2013). *Predictive modeling with SAS Enterprise Miner: Practical solutions for business applications*. SAS Institute.

Schumann, D. W., & Thorson, E. (Eds.). (1999). *Advertising and the world wide web*. Psychology Press.

Schumann, D. W., & Thorson, E. (2007). *Internet advertising: theory and research*. L. Erlbaum Associates Inc..

Shane, P. M., & Hunker, J. A. (2013). *Cybersecurity: Shared Risks, Shared Responsibilities*. Carolina Academic Press.

Shatz, A. (2009). Regulators rethink approach to online privacy. *Wall Street Journal*, (August 5).

Shin, W., Huh, J., & Faber, R. J. (2012). Developmental antecedents to children's responses to online advertising. *International Journal of Advertising*,*31*(4), 719-740.

Shoemaker, D., & Sigler, K. (2015). *Cybersecurity: Engineering a secure information technology organization*. Cengage Learning.

Song, P., Xu, H., Techatassanasoontorn, A., & Zhang, C. (2011). The influence of product integration on online advertising effectiveness. *Electronic Commerce Research and Applications*, *10*(3), 288-303.

Speed, T., & Ellis, J. (2003). *Internet security: a jumpstart for systems administrators and IT managers*. Digital Press.

StackExchange. (2015). K-Means Clustering – Calculating Euclidean distances in a multiple variable dataset. *Stack Exchange Inc*. Retrieved July 27, 2015 from the WorldWideWeb: http://stats.stackexchange.com/questions/69353/k-means-clustering-calculating-euclidean-distances-in-a-multiple-variable-data

Stamp, M. (2011). *Information security: principles and practice*. John Wiley & Sons.

Suh, S. (2012). *Practical applications of data mining*. Jones & Bartlett Learning Publishers.

Team, C. I. T. (2014). Unintentional Insider Threats: A Review of Phishing and Malware Incidents by Economic Sector.

Tehrani, P. M., Manap, N. A., & Taji, H. (2013). Cyber terrorism challenges: The need for a global response to a multi-jurisdictional crime. *Computer Law & Security Review*, *29*(3), 207-215.

Ten, C. W., Govindarasu, M., & Liu, C. C. (2007, October). Cybersecurity for electric power control and automation systems. In *Systems, Man and Cybernetics, 2007. ISIC. IEEE International Conference on* (pp. 29-34). IEEE.

Thomes, T. P. (2013). An economic analysis of online streaming music services. *Information Economics and Policy*, *25*(2), 81-91.

TriPHP. (2015). Website Statistics and Website Valuation. *TriPHP*. Retrieved August 11, 2014 from the WorldWideWeb: http://spy.triphp.com/website-worth

Tufféry, S. (2011). *Data mining and statistics for decision making*. John Wiley & Sons.

Tuten, T. L. (2008). *Advertising 2.0: social media marketing in a web 2.0 world*. Greenwood Publishing Group.

USC Annenberg Innovation Lab. (2013). USC Annenberg Lab Ad Transparency Report (January 2013). *USC Annenberg Innovation Lab*. Retrieved July 27, 2015 from the

WorldWideWeb: http://www.annenberglab.com/sites/default/files/uploads/USCAnnenbergLab_AdReport_Jan2013.pdf

USC Annenberg Innovation Lab. (2013). USC Annenberg Lab Ad Transparency Report (February 2013). *USC Annenberg Innovation Lab*. Retrieved July 27, 2015 from the WorldWideWeb: http://www.annenberglab.com/sites/default/files/uploads/USCAnnenbergLab_AdReport_Feb2013.pdf

USC Annenberg Innovation Lab. (2013). USC Annenberg Lab Ad Transparency Report (March 2013). *USC Annenberg Innovation Lab*. Retrieved July 27, 2015 from the WorldWideWeb: http://www.annenberglab.com/sites/default/files/uploads/USCAnnenbergLab_AdReport_Mar2013.pdf

USC Annenberg Innovation Lab. (2013). USC Annenberg Lab Ad Transparency Report (April 2013). *USC Annenberg Innovation Lab*. Retrieved July 27, 2015 from the WorldWideWeb: http://www.annenberglab.com/sites/default/files/uploads/USCAnnenbergLab_AdReport_Apr2013.pdf

USC Annenberg Innovation Lab. (2013). USC Annenberg Lab Ad Transparency Report (May 2013). *USC Annenberg Innovation Lab*. Retrieved July 27, 2015 from the WorldWideWeb:http://www.annenberglab.com/sites/default/files/uploads/USCAnnenbergLab_AdReport_May2013.pdf

Von Solms, R., & Van Niekerk, J. (2013). From information security to cyber security. *computers & security*, *38*, 97-102.

Wang, W., & Lu, Z. (2013). Cyber security in the Smart Grid: Survey and challenges. *Computer Networks*, *57*(5), 1344-1371.

Watters, P. A. (2012). *Cyber Security: Concepts and Cases*. CreateSpace Independent Publishing Platform.

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. (2013). Measuring Online Advertising Transparency in Singapore: An Investigation of Threats to Users. *Available at SSRN 2362626*.

Watters, P., Watters, M. F., & Ziegler, J. (2015, January). Maximising Eyeballs but Facilitating Cybercrime? Ethical Challenges for Online Advertising in New Zealand. In *System Sciences (HICSS), 2015 48th Hawaii International Conference on* (pp. 1742-1749). IEEE.

Watters, P. (2014). The Prevalence of High-Risk and Mainstream Advertisements Targeting Canadians on Rogue Websites. *Available at SSRN 2389850*.

Watters, P. (2015). Censorship is ~~futile~~ possible but difficult: A study in algorithmic ethnography. *First Monday*, *20*(1).

WEKA 3.6 Explorer. (2014). Waikato Environment for Knowledge Analysis Version 3.6.12. *Department of Computer Science of The University of Waikato, New Zealand*.

Whitman, M., & Mattord, H. (2010). *Management of information security*. Course Technology, Cengage Learning.

Wikipedia. (2015). Chilling Effects. *Wikimedia Foundation, Inc*. Retrieved July 27, 2015 from the WorldWideWeb: https://en.wikipedia.org/wiki/Chilling_Effects

Wikipedia. (2015). Digital Millennium Copyright Act. *Wikimedia Foundation, Inc*. Retrieved July 27, 2015 from the WorldWideWeb: https://en.wikipedia.org/wiki/Digital_Millennium_Copyright_Act

Wilson, C. (2003). Computer attack and Cyberterrorism: Vulnerabilities and Policy issues for Congress. *Focus on Terrorism*, *9*, 1-42.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Witten, I. H., Frank, E., Hall, M. A. (2011). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann Publishers.

Yan, J., Liu, N., Wang, G., Zhang, W., Jiang, Y., & Chen, Z. (2009, April). How much can behavioral targeting help online advertising?. In *Proceedings of the 18th international conference on World wide web* (pp. 261-270). ACM.

Yin, Y., Kaku, I., Tang, J., & Zhu, J. (2011). *Data mining: Concepts, methods and applications in management and engineering design*. Springer Science & Business Media.

Yong, Y., Ikou, K., Jiafu, T., & JianMing, Z. (2011). Data Mining: Concepts, Methods and Applications in Management and Engineering Design (Decision Engineering), UK.

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.

Zhang, L., Yu, S., Wu, D., & Watters, P. (2011, November). A survey on latest botnet attack and defense. In *Trust, Security and Privacy in Computing and Communications (TrustCom), 2011 IEEE 10th International Conference on* (pp. 53-60). IEEE.

Zhu, Z. A., Chen, W., Minka, T., Zhu, C., & Chen, Z. (2010, February). A novel click model and its applications to online advertising. In *Proceedings of the third ACM international conference on Web search and data mining* (pp. 321-330). ACM.