VICTORIA UNIVERSITY OF WELLINGTON
Te Whare Wananga o te Upoko o te Ika a Maui

School of Mathematical and Computing Sciences
Computer Science

PO Box 600
Wellington
New Zealand

Tel: +64 4 463 5341, Fax: +64 4 463 5045
Email: office@mcs.vuw.ac.nz
http://www.mcs.vuw.ac.nz/research

Internet NZ Study - Stage 1 Report

Christian Seifert, David Stirling, Ian Welch, Peter Komisarczuk
{cseifert, david.stirling, ian.welch, peter.komisarczuk}@mcs.vuw.ac.nz

February 2008

Executive Summary

Broadband connectivity and the great variety of services offered over the Internet have made it an important source of information and entertainment and a major means of communication. In 2008, users are more connected than ever.

With connectivity to the Internet, however, come security threats. Because the Internet is a global network, an attack can be delivered anonymously from any location in the world. In 2003, a study was published that compared vulnerabilities in web servers from various countries. It showed that web sites in the New Zealand domain were one of the most vulnerable to Google hacking attacks [4].

Vulnerable web servers lead to compromised web servers, which do not only pose a risk to operators of these web servers, but also to their users or customers, because the attackers might insert client-side attacks that trigger upon interaction with such a compromised server [8].

As part of work commissioned by InternetNZ, Victoria University of Wellington has gathered intelligence on the threat from malicious servers across different country domains (.au, .com, .nz, and .uk) In this study, it was assessed whether the web servers located in the nz domains would show an elevated percentage of malicious web servers in line with their vulnerability independent by Nehring [4].

Results of this study show malicious URLs in all domains. Inspecting 664,000 URLs, 38 malicious URLs were identified. The majority of these malicious URLs were unknown to providers of security products/ services therefore unable to protect potential end users that visit those URLs.

No indications were detected that show an elevated percentage of malicious web servers in the nz domain. Rather, a statistically significant elevated percentage was detected in the au domain compared to any other domain inspected.

In the nz domain, three malicious URLs were detected in this study that scanned part of the nz domain. Assuming approximately 4 million web servers exist in the nz domain, the number of malicious web servers is estimated to be in the order of 640 or 0.00016%.

In the next stage of this work, all web servers in the nz domain will be inspected to provide a more complete picture of the threat of client-side attacks to New Zealand.

1 Introduction

Broadband connectivity and the great variety of services offered over the Internet have made it an important source of information and entertainment and a major means of communication. In 2008, users are more connected than ever.

With connectivity to the Internet, however, come security threats. Because the Internet is a global network, an attack can be delivered anonymously from any location in the world. In 2003, a study was published that compared vulnerabilities in web servers from various countries. It showed that web sites in the New Zealand domain were one of the most vulnerable to Google hacking attacks [4].

It is suspected that the number of servers with exposed vulnerabilities directly correlate to the number of compromised servers. This is likely because attackers are oblivious to the physical location of the servers. Wide ranges of scans are performed and servers are attacked as soon as a vulnerable service is found [9].

Compromised web servers not only pose a risk to the operators of these web servers, but also to users, because the attackers might insert client-side attacks that trigger upon interaction with such a compromised server [8]. For example, a browser could send a request to the server and this server responds with a malicious web page that attacks a specific browser vulnerability that allows the server to execute malware on the victim's machine. This usually happens without the user's notice or consent. Traditional defenses, such as firewalls, pose no barrier against these attacks and antivirus detection is currently poor [12].

Client-side attacks are prevalent in many areas of the Internet [11]. As such, any user might be affected and if this problem is not tackled more effectively it might result in a loss of trust in usage of the Internet and therefore negatively impact cyber commerce, banking and e-government efforts.

In order to tackle the problem, first, intelligence needs to be gathered to assess the magnitude of the problem. Internet NZ has commissioned Victoria University of Wellington to gather such intelligence on the web servers in the New Zealand domain. Stage 1 of this project is tasked with comparing the threat of malicious servers across different countries. If more vulnerable servers are observed in the NZ domain [4], it is likely that more malicious servers would be found in the NZ domain as well. This is important information as one could devise defensive strategies with the information gathered. This report summarizes the results of stage 1.

2 Methodology

Using the hybrid system, 664,000 web pages from the Australia, New Zealand, UK, and .com domain were inspected. The hybrid system is composed of low- and high-interaction client honeypot Weta v1.0 (currently not publicly available) and Capture-HPC v2.1 [10]. Each malicious URL was recorded and the number of malicious URLs per domain were analyzed to determine statistically significant differences across domains. The study design is described in this section; results are presented in Section 3.

2.1 Client Honeypots

Client honeypots are an emerging technology for detecting malicious servers on a network. Client honeypots interact with potentially malicious servers and assess whether the web page returned by the server contains an attack. There are two types of client honeypots used in this study: low- and high-interaction client honeypots.

2.1.1 Low-Interaction Client Honeypots

Low-interaction client honeypots are client honeypots that use a simulated browser to interact with a web server and then examines the structure and elements on the web page to determine whether it is malicious. For instance, a low-interaction client honeypot could use a signature to check for the existence of a particular attack. The advantages of this technology is that it is fast, but on the contrary is known to miss attacks and may produce many false alerts.

The low-interaction client honeypot used for this study is Weta v1.0. It retrieves web pages at high speed and inspects the underlying static attributes of the initial HTTP response and embedded HTML code, such as the existence of IFrames, Script tags, etc. Because of the common attributes used by malicious web pages, it is possible to identify malicious pages using this method. Some attacks will be missed with this method.

2.1.2 High-Interaction Client Honeypots

High-interaction client honeypots use a dedicated, vulnerable computer system to interact with potentially malicious servers. As responses from these servers are consumed by the client honeypot, it monitors the operating system for any unauthorized state changes (excluding authorized changes to the system state associated with background system processes, such as the creation of temporary cache files). For instance, if a new file appears in the start-up folder after the vulnerable browser interacted with a server, the client-honeypot can conclude that the server it just interacted with must have launched a successful attack and placed it there. The advantages of high-interaction client honeypots are their ability to detect unknown attacks and their low false positive rate; on the flipside they are slow and require a lot of resources.

The high-interaction client honeypot used for this study is Capture-HPC v2.0 configured with Windows XP SP2 and Internet Explorer 6 SP2. It is an open-source high-interaction client honeypot that has been developed at Victoria University of Wellington. It can drive a number of clients and therefore is able to detect a wide range of attacks. It monitors the registry, file system, and processes on the kernel-level and therefore is resilient against obfuscation attempts by attacks and malware.

Since the low-interaction client honeypot produces false positives, the low- and high-interaction client honeypot were combined into a hybrid system. The high-level view of the system is shown in Figure 1. At the front stands the low-interaction client honeypot that initially retrieves and classifies the web pages denoted by the input URLs. Because the false positive rate is high, all URLs that have been classified as malicious are forwarded to the second part of the hybrid system. The high-interaction client honeypot retrieve the page once again and make a final classification. Since the high-interaction client honeypot has a low false positive rate, the false positives will be filtered out.

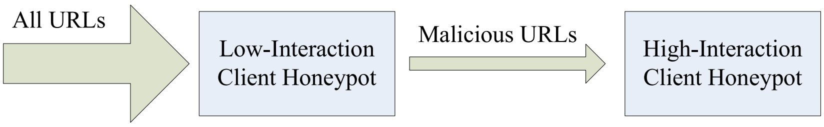

All URLs
Low-Interaction Client Honeypot
Malicious URLs
High-Interaction Client Honeypot

Figure 1: Hybrid System

This hybrid system will miss about a quarter of the attacks. This is acceptable for the purpose of this comparative study as malicious pages are likely to be missed with equal levels across all domains. As such, a comparison is still possible, although statistical assumptions are made on the size of populations.

2.2 Data Collection

In January and February 2007, approximately 168,000 pseudo-randomly chosen web pages for each of the following domains: au, com, nz, and uk were inspected; approximately 664,000 web pages in total. The URLs that denote these web pages were obtained by submitting English 5 N-gram queries randomly selected from the corpus of web pages linked by the DMOZ Open Directory Project [5] to the Yahoo! search engine [1]. The first 1000 URLs on the results page were used to build the list of 664,000 URLs. In case less than 1000 URLs were shown on the results page, identical number of URLs were taken to ensure that bias from one particular query did not result in a bias within the collected data set favoring one or the other domain. It also ensured that identical number of URLs were collected across the domain.

However, due to the number of hosts in each of the domains, the URLs returned by the search engine result are hosted on different number of unique hosts. The number of unique hosts names of the input URLs per domain is shown in Figure 2. These numbers reflect the general notion that there are more servers in the, for instance, .com than the .uk domain.

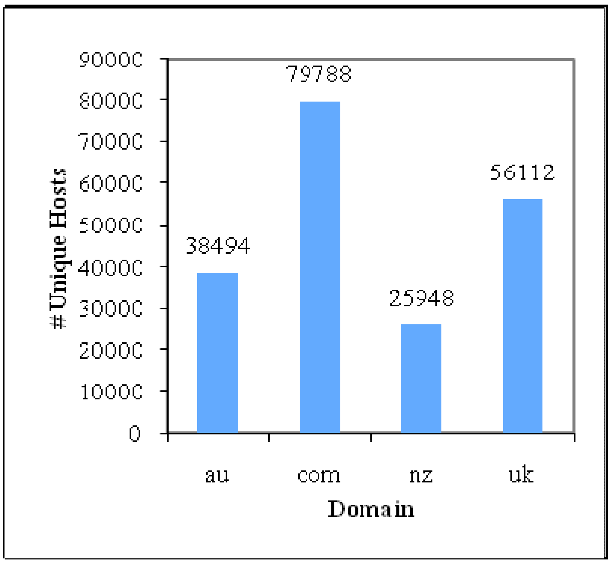


Figure 2: Unique Hosts of Input URLs per Domain

All URLs were first inspected by the low-interaction client honeypot Weta v1.0. The client honeypot retrieved the initial HTTP response and performed a static assessment on whether the page is potentially malicious. This resulted in a greatly reduced set of URLs - many of which contained false positives. These were forwarded to the high-interaction client honeypot Capture-HPC v2.0 for a second inspection. In the case the low-interaction client honeypot produced a false positive, Capture-HPC would successfully correct this assessment; in the case a true positive was produced, Capture-HPC would confirm this assessment. With each malicious page detected, Capture would report the malicious classification, the URL, and all the unauthorized state changes that occurred on the machine.

To assess whether one domain contains generally more malicious URLs than others, a simple Chi-Square test was performed on the number of malicious URLs. Because some malicious URLs might have originated from the same host, a second Chi-Square test was performed on the number of unique host names that host at least one malicious URL.

All malicious URLs were cross-checked against the URL assessment service of Google's Safe Browsing API [2] and McAfee SiteAdvisor [3].

3 Results

Inspecting the 664,000 URLs, the low-interaction client honeypot Weta produced a total of 11,095 URLs that it considers may be malicious. Since the majority of these URLs were false positives, they were inspected once again with the high-interaction client honeypot Capture-HPC. A total of 38 malicious URLs from 27 unique hosts were detected. The URLs point to web pages that attack Internet Explorer 6 SP2. A server will be able to gain complete control of such a system if the user merely navigates to such a page. The user will not notice that an attack has been launched.

Figure 3 shows the number of malicious URLs and hosts per domain. For example, of the 168,000 URLs per domain, 26 unique malicious URLs from 16 unique hosts were identified for the au domain, whereas only three URLs from three unique hosts were identified for the nz domain. The statistical Chi-Square test shows that the difference between the malicious URLs and hosts identified in the au domain and any of the other domains is statistically very significant (URLs: $p < 0.0036$; Hosts: $p < 0.0092$).

All malicious URLs are listed in Tables 1,2,3, and 4. The three malicious URLs for the NZ domain are listed in Table 3. The first URL is hosted on a site that primarily is concerned with UFO watch in NZ. The URL itself, which seem to point to pornographic content, is not related to the overall content of the site and might have been placed there by an unauthorized party. The second URL points to a business page of AWL Professional Housewashing Services Ltd dealing in "Housewashing and General Waterblasting" and the third URL points to a dating site NZ Singles. Considering the domain name and the high ranking in the Google NZ Search Engine on the term "singles" suggests that this is a site that is able to attract a lot of traffic. (Note that due to the dynamic nature of malicious networks, the URLs might or might not exhibit malicious behavior today. The time these URLs exhibited malicious behavior was in January 2008.)

Cross-checking all URLs against known bad URLs of Google's search index [2] and McAfee SiteAdvisor [3] showed only 21% of URLs were known by Google to be bad URLs whereas McAfee SiteAdvisor was not knowledgeable of any. The three NZ URLs were known to

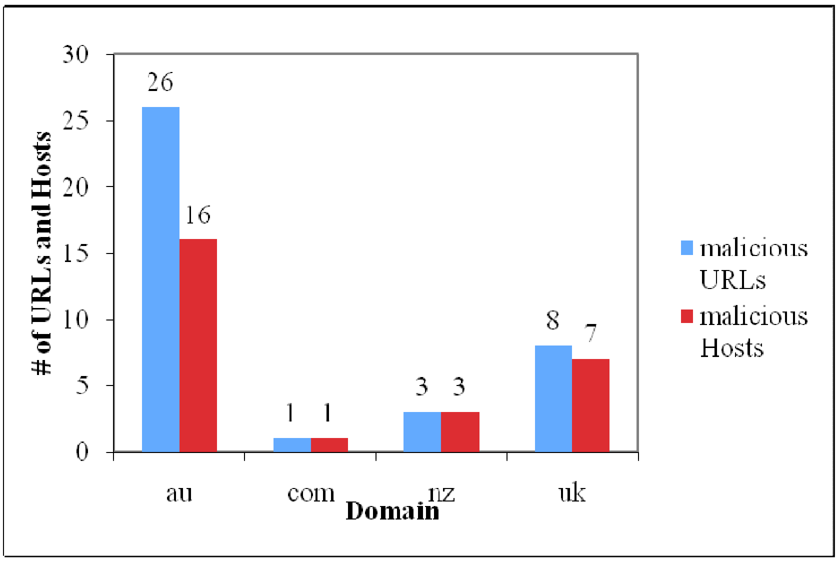


Figure 3: Malicious URLs and Hosts per Domain

neither Google nor McAfee SiteAdvisor.

4 Related Work

As part of this study, adult URLs were also inspected further using high-interaction client honeypots. While the results of this analysis showed that the majority of the malicious URLs reside in the .com domain, results were omitted in this report due to a skew in the data set. Since adult URLs were obtained submitting adult keywords to the Yahoo! Search Engine, the resulting URLs not necessarily all pointed to sites hosting adult content. Rather the percentage of actual adult sites varied across domain, which had a direct impact on the number of malicious URLs. As such, the results were discarded.

A recent technical report by Google Inc. shows China, US, and Russia at the top serving malicious content [7]. No data is provided for Australia, New Zealand, and UK. Also, the study compares relative number of URLs rather than equal number of URLs resulting in countries being ranked higher if more servers exist in those countries.

McAfee SiteAdvisor performed a comparative study of malicious URLs by top level domain in 2007 in which 265 country and several generic top level domains were compared [6]. This study not only assessed URLs for drive-by-download attacks, but rather utilized the SiteAdvisor classification which also takes into account malicious binaries, pop-ups, spam factor resulting from submitted email addresses, etc. The riskiest country top level domain was .ro (Romania). The results from this study ranks the .au domain the lowest with 0.2%, followed by .uk domain (0.5%), .nz domain (0.6%) and finally the .com domain (5.5%).

5 Conclusion

In this study, malicious URLs across the domains au, com, nz, and uk were inspected. Inspection of 664,000 pseudo-randomly URLs resulted in identification of 38 malicious URLs from 25 unique hosts. Malicious URLs were found in all domains, but a statistical significant difference was found between the au and any other domain for this date over this period. As such, more malicious URLs that attack Internet Explorer 6 SP2 seem to exist in the au domain.

The methodology used introduced some bias in the input URLs. Using queries to search engines produced URLs that are ranked highly on the search engine. 5-N gram English queries were used to reduce the bias toward popular pages. While some bias remains, it is likely to be consistent across domain, and as such should not interfere with this comparative study. Further, the methodology resulted in missed attacks. The low-interaction client honeypot component is likely to miss approximately a quarter of the attacks. However, again, since the false negative rate is applied independent of the domain, it does not impair the basis of this comparative study.

Significantly more malicious URLs and hosts were identified from the au domain. Considering the study on vulnerable web servers, the authors would have expected nz domain to also show high number of malicious URLs. However, such behavior was not observed. This might be due to the fact that nz pages attract less traffic and therefore are not attractive to attackers or the fact that the delivery mechanism of exploits to web sites might not be direct attacks on the web servers, but rather other mechanisms, such as cross-side scripting, arp spoofing, imported scripts (e.g. counters) or advertisements. Further work needs to be performed to provide more insight to this observation.

All domains contain malicious pages though. No top level domain is immune to the threat of client-side attacks. While only three URLs were identified in the nz domain, one of the URLs points to the host www.nzsingles.co.nz. Considering the domain name and high ranking in the Google's search index suggest that this site might attract a lot of traffic. A popular site hosting a client-side attack code might actually post a larger risk than a multitude of unpopular sites hosting client-side attacks, because the overall number of attacks launched on end-users might be greater.

Malicious URLs were cross-checked against known bad site of the Google index and McAfee SiteAdvisor. Very few URLs were known by these two sources. This is an indication how difficult it is to identify malicious pages on the Internet and provide defensive intelligence to end-users. Malicious pages might appear and disappear on pages within short period of times and considering the number of malicious pages on the Internet might make this like finding a moving needle in the Internet haystack.

How many needles does one expected to find in the NZ haystack? To estimate this number, it is assumed that the sample maps to the NZ domain. Three malicious hosts were identified out of approximately 25000 malicious hosts. If a false negative rate of 25% is assumed, actually 4 malicious hosts are contained in the 25000 malicious hosts. Assuming approximately 4 million hosts are registered in the NZ domain, this results in an estimate of 640 malicious hosts may exist in the NZ domain. Considering that the sample is likely to have selected URLs that are less infectious overall (e.g. adult URLs are more infectious), the number is likely to be higher.

6 Future Work

Stage 2 of the study touches on the last point: How many hosts are likely to exist in the nz domain that serve malicious web pages that attack Internet Explorer 6 SP2. A comprehensive list will allow one to perform additional analysis to more completely assess the risk of client-side attacks in the nz domain. In stage 2, all index pages of all hosts of the nz domain will be inspected. Because popularity represents a factor in the risk assessment, it is proposed to use available usage data of the URLs to perform a comprehensive risk assessment.

Acknowlegement

This work was funded by InternetNZ.

References

[1] Filo, D., and Wang, J. Yahoo! search engine, 1994.

[2] Google Inc. Google safe browsing api, 2007.

[3] McAfee, Inc. Mcafee siteadvisor, 2005.

[4] Nehring, N. *The Nature of Vulnerability Detectable by Google Hacking: Whos at Risk?* PhD thesis, Massey University, 2005.

[5] Netscape Communications Corporation. Dmoz open directory project, 1998.

[6] Nunes, D., and Keats, S. Mapping the mal web, 2007.

[7] Provos, N., Mavrommatis, P., Rajab, M. A., and Monrose, F. All your iframes point to us, 2008.

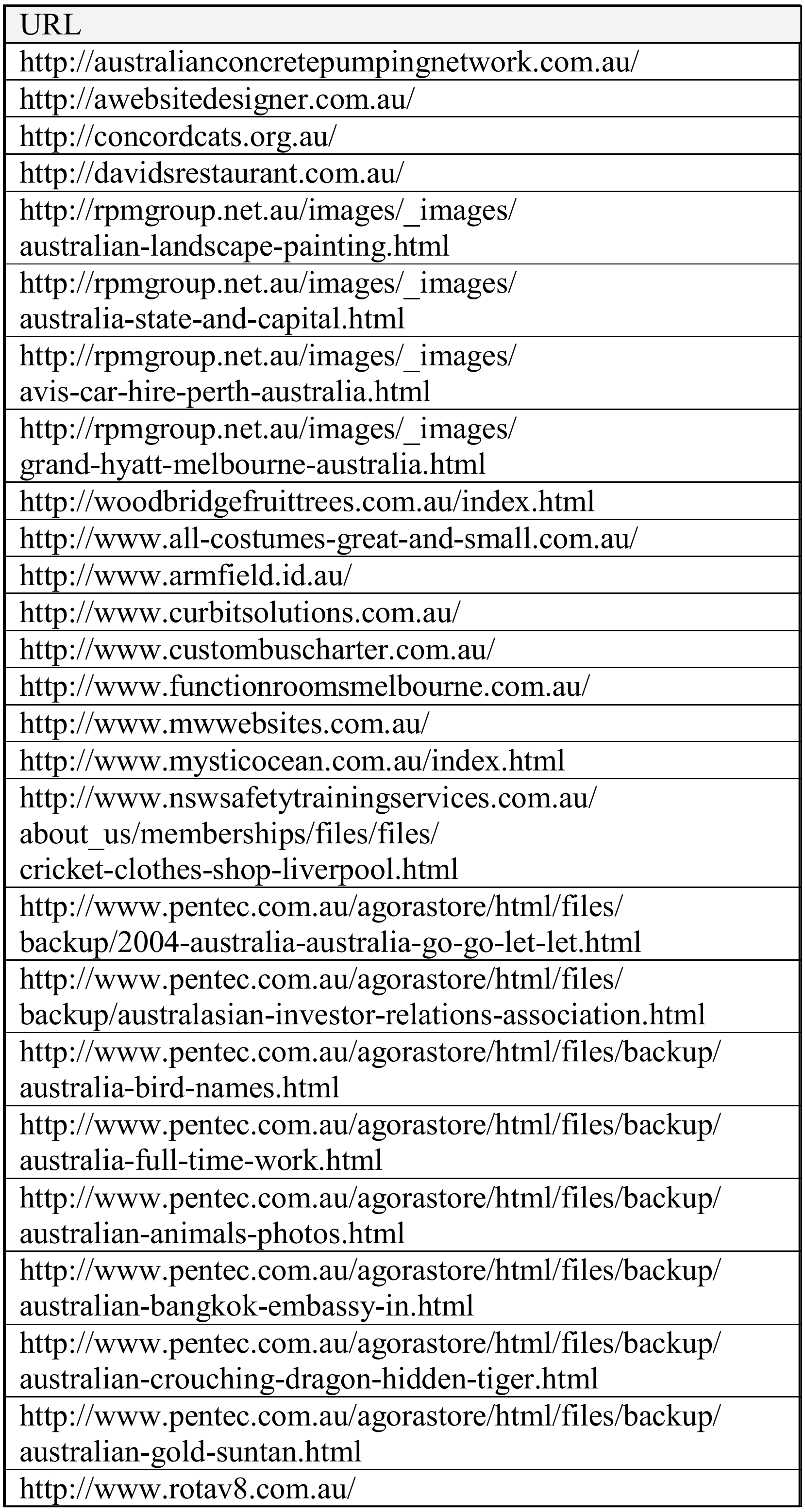
[8] Provos, N., McNamee, D., Mavrommatis, P., Wang, K., and Modadugu, N. The ghost in the browser: Analysis of web-based malware. In *HotBots'07* (Cambridge, 2007), Usenix.

[9] Seifert, C. Analyzing malicious ssh login attempts, 2006.

[10] Seifert, C., and Steenson, R. Capture - honeypot client, 2006.

[11] Seifert, C., Steenson, R., Holz, T., Bing, Y., and Davis, M. A. Know your enemy: Malicious web servers, 2007.

[12] Seifert, C., Welch, I., Komisarczuk, P., and Narvaez, J. Drive-by-downloads, February 2008 2008.



URL
http://australianconcretepumpingnetwork.com.au/
http://awebsitedesigner.com.au/
http://concordcats.org.au/
http://davidsrestaurant.com.au/
http://rpmgroup.net.au/images/_images/australian-landscape-painting.html
http://rpmgroup.net.au/images/_images/australia-state-and-capital.html
http://rpmgroup.net.au/images/_images/avis-car-hire-perth-australia.html
http://rpmgroup.net.au/images/_images/grand-hyatt-melbourne-australia.html
http://woodbridgefruittrees.com.au/index.html
http://www.all-costumes-great-and-small.com.au/
http://www.armfield.id.au/
http://www.curbitsolutions.com.au/
http://www.custombuscharter.com.au/
http://www.functionroomsmelbourne.com.au/
http://www.mwwebsites.com.au/
http://www.mysticocean.com.au/index.html
http://www.nswsafetytrainingservices.com.au/about_us/memberships/files/files/cricket-clothes-shop-liverpool.html
http://www.pentec.com.au/agorastore/html/files/backup/2004-australia-australia-go-go-let-let.html
http://www.pentec.com.au/agorastore/html/files/backup/australasian-investor-relations-association.html
http://www.pentec.com.au/agorastore/html/files/backup/australia-bird-names.html
http://www.pentec.com.au/agorastore/html/files/backup/australia-full-time-work.html
http://www.pentec.com.au/agorastore/html/files/backup/australian-animals-photos.html
http://www.pentec.com.au/agorastore/html/files/backup/australian-bangkok-embassy-in.html
http://www.pentec.com.au/agorastore/html/files/backup/australian-crouching-dragon-hidden-tiger.html
http://www.pentec.com.au/agorastore/html/files/backup/australian-gold-suntan.html
http://www.rotav8.com.au/

Table 1: Malicious URLs of the au Domain

URL
http://loversinart.com/

Table 2: Malicious URL of the com Domain

URL
http://moofish.co.nz/ecotours/inc/2/fucking-milf-trailer.html
http://awlhousewash.co.nz/
http://www.nzsingles.co.nz/

Table 3: Malicious URLs of the nz Domain

URL
http://anotherserver.co.uk/
http://property-bulgaria.org.uk/
http://tom-gibson.co.uk/
http://www.databasebackup.co.uk/
http://www.emberreptiles.co.uk/index.htm
http://www.info-sales.co.uk/
http://www.newsmakers.co.uk/2006/08/31/ati-delivers-blazing-fast-integrated-performance-for-intel-based-pcs-and-laptops
http://www.newsmakers.co.uk/page/23

Table 4: Malicious URLs of the uk Domain