Assignment 2.3

Analyse Website Data

Name: Beau Lebens

Student Number: 09918322

Unit Name: NET 26: Cyberanalytics

Email Address: beau@dentedreality.com.au

Date Submitted: 26 October 2003

Word Count: 828

By submitting this assignment, I declare that I have retained a suitable copy of this assignment, have not previously submitted this work for assessment and have ensured that it complies with university and school regulations, especially concerning plagiarism and copyright.

Web Log Traffic Analysis and Interpretation

Quality of Collected Data

The data collected via web server logs is prone to errors in relation to areas such as the following;

- Duplicate visits from the same user, with a different IP address due to dynamic IP allocation, causing the user to appear unique.
- Automated systems such as search spiders/robots, viruses and site downloaders can bump up traffic while not really reflecting true visits (although they will reflect data volume etc accurately).
- User Agent, country IP lookup and operating system reports can all be faulty due to 'spoofed' or incomplete data being reported from the user's end.

Upon initial analysis of the log file supplied, it was apparent that there was a certain amount of traffic which was obviously attempting to exploit well-known vulnerabilities in the Microsoft Internet Information Services server using a buffer overflow in its handling of Chunked Encoding (See Microsoft site for more details on how to patch this if it hasn't been done already: http://www.microsoft.com/technet/treeview/default.asp?url=/technet/security/bulletin/MS02-018.asp). Traffic that appeared to be attempting to exploit this vulnerability and another known one (http://www.apacheweek.com/features/codered) was stripped from the log before analysis took place to avoid false results. The following command was used on a RedHat Linux machine to modify the log file:

```
cat sarulog.txt | grep -v " HEAD " | grep -v "default.ida" > modified_sarulog.txt
```

See the attached output file (HTML) for the resulting statistical analysis and the log file (TXT) for the modified version.

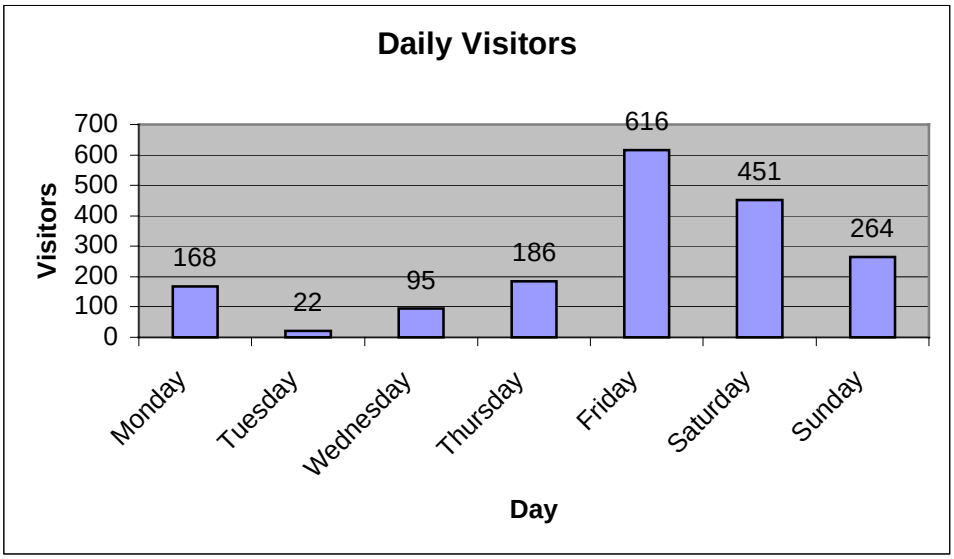
General Traffic Review

A summary of some statistical details of the entire range of data available is below;

	Hits	Uniques
Total	1802	335
Mean	120	22
Median	99	20
Maximum	473	39
Minimum	22	4
Range	451	35

This shows us a common trend, where there are a large number of requests from a smaller number of unique visitors (on a technical level, each image counts as a request, as well as each page, JavaScript file, animation etc). There is clearly a much smaller range in values, as well as a lower mean and median value, when analysing the unique visitors to the site.

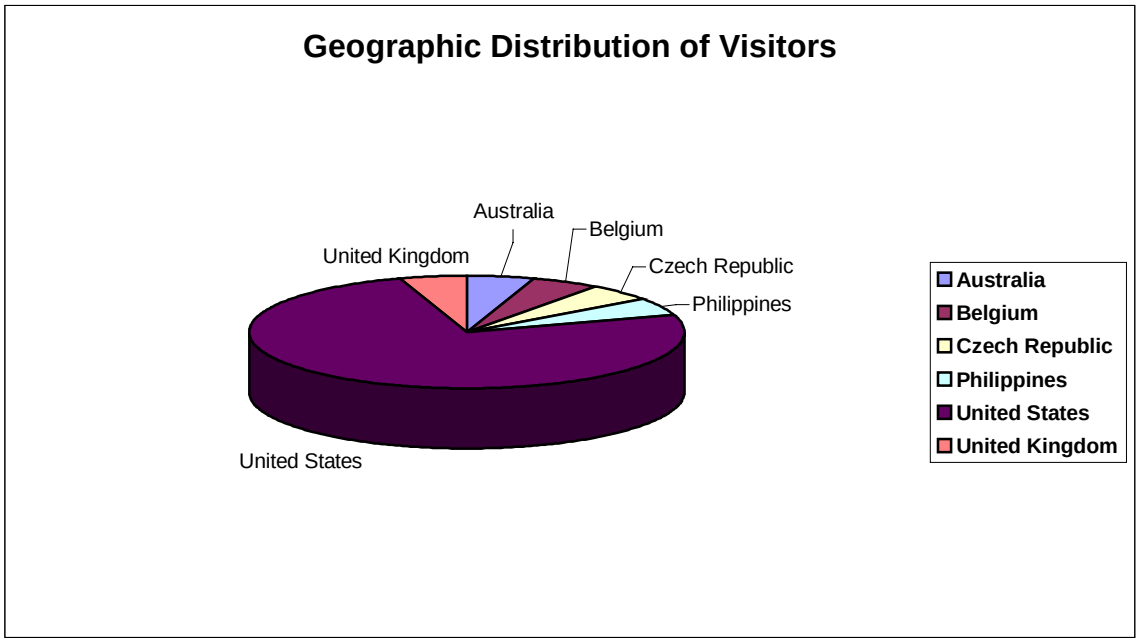
Looking at a daily spread of traffic shows a spike towards the end of the week, moving in to the weekend, as shown below. This suggests that the end of the week will be the heaviest time of load for the servers, and thus the most potential for failure. This can be used to determine technical staffing requirements for the department hosting the server.



Geographical Distribution of Visitors

Taking the most recent 20 IP addresses of visitors (as listed by OpenWebScope), the following spread of geographic locations can be determined (country information gathered via MaxMind – http://www.maxmind.com/)

	Count	Percentage
Australia	1	**5%**
Belgium	1	**5%**
Czech Republic	1	**5%**
Philippines	1	**5%**
United States	15	**75%**
United Kingdom	1	**5%**
Total	20	100%



This spread indicates a large international contingent of visitors, and may also help to explain the apparent "off-centre" weekend traffic (See: General Traffic Patterns for more detail).

Technical Support/Server Load

With a large number of visitors being from overseas (more importantly, in different time zones), it could be recommended that technical support will need to be available all hours to cater for

potential support calls when the server is under its main load (local night time). Reviewing the traffic volumes per hour reveals that two thirds of the daily traffic for the site occurs 'after hours', which means that the servers are under their peak load when it is *least likely* that technical staff will be available.

Total Hits	1802	
After Hours	1211	**67%**
Business Hours	591	**33%**

Search Engine Spidering

The logs indicate that this website has been spidered a number of times during the 3 weeks of data supplied, by Google (Googlebot), Teoma/Ask Jeeves, FAST and the 'Grub' client. One possible technique which may assist these search engine spiders would be to use the '/robots.txt; file to tell them what not to bother indexing (i.e. dynamic content). The 'Broken Links' list indicates that this file is not on the server, and all robots should attempt to read it when first accessing a site to see what to ignore. Simply add the file in the root of the web server's files and they will find it (See http://www.searchtools.com/robots/robots-txt.html for more details).

Browser Dominance

From the output provided by OpenWebScope, it appears that Microsoft's Internet Explorer represents a solid majority of the requests to the site (Internet Explorer 6.0 accounts for 73% of traffic). When versions 5.0 to 6.0 of Internet Explorer are combined, they account for 87% of logged traffic to the site. This information could be useful for web developers so that they know what level of compliance they should target across different browser versions. It should be noted however, that browser detection is not entirely accurate because it relies on a value passed from the web browser to the server. In some cases this value can be modified or withheld entirely, which will obviously affect final results and accuracy.

Appendices

Log Data

I have included a copy of the log file that I actually used for analysis, since it is a modified version of the raw log file (to remove apparent virus activity – See Quality of Collected Data for more information).

OpenWebScope Output

Please see the attached HTML file (net26a2.3beaulebens.html) for the output produced by OpenWebScope. Please note that this output is the result from the log file once apparent virus/security exploit traffic was removed.

Statistical Analysis

Please see the attached Microsoft Excel file (net26a2.3beaulebens.xls) for the raw data and further statistical analysis performed in support of this report. Note that the further analysis is performed on separate worksheets within the one workbook.

References

Cox, M. J. (2001). *Code Red requests for /default.ida*. Retrieved 26 October, 2003, from http://www.apacheweek.com/features/codered

Microsoft Security Bulletin MS02-018. Retrieved 26 October, 2003, from http://www.microsoft.com/technet/treeview/default.asp?url=/technet/security/bulletin/MS02-018.asp

Search Indexing Robots and Robots.txt. Retrieved 26 October, 2003, from http://www.searchtools.com/robots/robots-txt.html