Exhibit D

Introduction

I was asked by attorneys representing the North Central Regional Library District ("NCRL") to write this report and present testimony concerning it to the extent necessary. I agreed. Although I do not engage in litigation consulting as my primary means of earning income, I am being paid for this engagement at the rate of $400 per hour.

I was asked to explain how the NCRL filtering software works. I was also asked to assess the methods used in study of error rates in the filtering software NCRL uses as reported by Mr. Haselton. I was also asked to conduct a study of my own if I thought it would yield greater insight into whether the NCRL filters block more than the content they intended to block. I did conduct such a study, and report on the methods and results in this document.

I have also reviewed the report prepared by Bennett Haselton for purposes of the NCRL litigation. I had two main concerns with the study reported by Mr. Haselton. First, I was concerned that the set of URLs he tested might not be representative of those that NCRL library patrons actually visit. He selected a random sample of all possible domains. It seemed to me likely that library patrons would tend to visit more popular destinations than a random sample. For example, neither google.com nor yahoo.com appear in his random sample, but would almost certainly be viewed by library patrons. It also seemed to me likely that the blocking software would make fewer errors on more popular sites, because the Fortinet company that NCRL contracts with to provide the blocking service would invest more effort in correctly classifying more popular sites. Both these intuitions were born out in the results of the study I conducted.

Second, I was concerned with the reliability of Mr. Haselton's classification of URLs. He did the classification based on loosely defined criteria, without any check on the reliability of his assessments (no second rater or inter-rater reliability or test-retest reliability check). This is contrary to the accepted practice in the social sciences when trying to turn subjective human assessments into an objective, repeatable measure. In our test, using more rigorous methods, we were still not able to achieve perfect reliability in our categorization of urls, suggesting that Mr. Haselton's classification may have been even more error-prone.

Personal Background

I am a Professor at the University of Michigan School of Information. In 2002, I conducted an assessment of the error rates on health-related websites of several commercial Internet filters. That study was published in JAMA, the flagship peer-reviewed journal of the American Medical Association. Appendix 1. A subsequent paper abstracting what we had learned about methods for conducting tests of filtering software was published in the Communications of the ACM, the flagship publication distributed to all members of the Association for Computing Machinery. Appendix 2. More details of my qualifications can be found in Appendix 3.

Dr. Derek Hansen is an Assistant Professor at the University of Maryland. He served as a research assistant on the 2002 study of Internet filters, where he was responsible for rating a large number of web sites based on whether they contained health information or not, and whether they contained pornography or not. Subsequently, as part of his Ph.D. project at the University of Michigan, he again had to develop a classification system for a corpus of texts (email messages and web pages). Though unrelated to filtering or pornography, that project gave him additional experience in creating a reliable categorization scheme and instructions for raters. More details of his qualifications can be found in Appendix 4.

Michael Hess has a master's degree from the University of Michigan School of Information. He has significant experience as a database and system administrator. He wrote the scripts for processing log files and created a web-based tool that allowed the raters to look at a large number of URLs and enter their assessments of them through a web-based form. More details of his qualifications can be found in Appendix 5.

How NCRL's Filters Work

I have been told by NCRL staff that NCRL has installed a FortiGate firewall/proxy unit, sold by the Fortinet company, in each of its branch libraries. The FortiGate is a small piece of hardware, smaller than a typical laptop computer. My understanding is that all computers in all of the branches access the Internet by connecting through these FortiGate units, in the manner that I describe below.

What Happens When a Patron Fetches a Page

To understand how the FortiGate affects the Internet activity of an NCRL patron, it is helpful to consider the sequence of steps that occur behind the scenes, invisible to an NCRL patron, each time a patron tries to visit a web page. A visit may be initiated either by directly entering a URL into the toolbar, by selecting a bookmarked favorite from a menu, or by following a link from another page. Regardless of how a visit to a web site is initiated, the same sequence of events occurs in the background. Fortinet provides the following diagram on its website to explain how its filtering works. My explanation is based on the explanation Fortinet provides to accompany the diagram on its website, with significant elaboration to explain terms that may not be familiar to non-technologists.

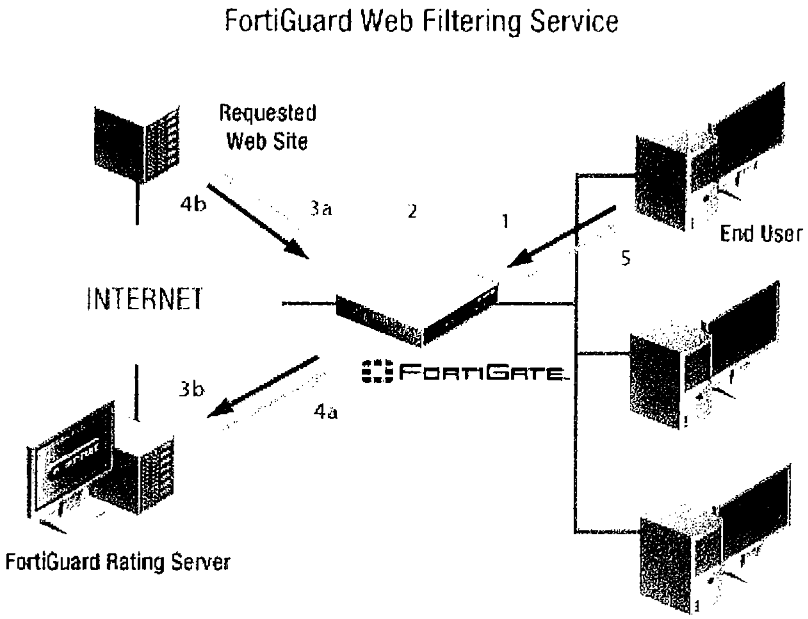


Figure 1. The sequence of events in a potentially blocked request to visit a URL.

Even before the first step shown in the diagram, the patron's computer does a little work to decode the URL that the patron has requested to view. Consider a URL such as http://www.yahoo.com/nfl. The first part, the letters "http" occurring before the //, constitute the service (or, more technically, the protocol). The http protocol is for connecting to a server to retrieve a web page. Other common protocols include https (for connecting to a server with an encrypted connection to securely retrieve a web page) and ftp (for downloading files).

The next portion, www.yahoo.com , including one or more periods and ending at the next /, is called the hostname or domain name. The Internet has a domain name system (sometimes called DNS) that allows the patron's computer to look up an address like www.yahoo.com in order to find out a corresponding IP address such as 69.147.114.210. That numeric IP address uniquely identifies the destination host (the web server).

The third portion, /nfl in this example, which begins after the domain name, is sometimes called the filename or path or urlpath.

Step 1. The patron's computer attempts to establish a connection to the IP address of the destination host. It tries to send a message that it would like to "GET" whatever the server provides in response to this urlpath, such as an HTML document or an image file. Because the patron's computer accesses the Internet through the FortiGate unit, however, a few other things happen along the way, and the patron's computer may not receive the same response it would have received had it been connected directly to the Internet, without going through a filtering proxy/firewall like a FortiGate.

Step 2. If the URL has been requested recently, by this patron or another patron in the same branch, the FortiGate may already have a cached copy of Fortinet's rating of the

URL or the server's response to that request. This may eliminate one of the requests in 3a or 3b, for efficiency reasons, but has no effect on whether the site will be blocked or not.

Step 3a. The FortiGate contacts the destination host, asking for the same urlpath that the patron's computer originally asked for. The FortiGate waits to receives a response at step 4b.

Step 3b. The FortiGate also connects another server, elsewhere on the Internet, labeled in the diagram as "Fortiguard Rating Server", and sends it the text of the requested URL. This server is maintained by Fortinet, not by NCRL. NCRL pays a subscription fee to allow its FortiGates to continue to use the Fortiguard Rating Sever.

Step 4a. The rating server takes the URL and responds to the FortiGate with two pieces of information, a "category" and a "classification". For example, http://sports.yahoo.com/nfl is in the category "Sports" and classification "Unclassified". Each URL is classified in at most one category and one class. Sometimes, the FortiGuard rating server may temporarily malfunction (e.g., a power failure), or it may not provide category and class information for particular URLs. In that case, the FortiGate receives no response or a response that it cannot interpret.

Step 4b. The FortiGate receives the requested contents from the destination host.

Policy Evaluation Step. The FortiGate runs a simple policy evaluator program that evaluates whether the URL is blocked or allowed under the filtering profile in operation. The policy evaluator first examines whether the URL is one that is specifically exempt or blocked according to explicit local rules set by NCRL staff and stored as part of the filtering profile in the FortiGate. If so, the evaluator returns the result of that local rule.

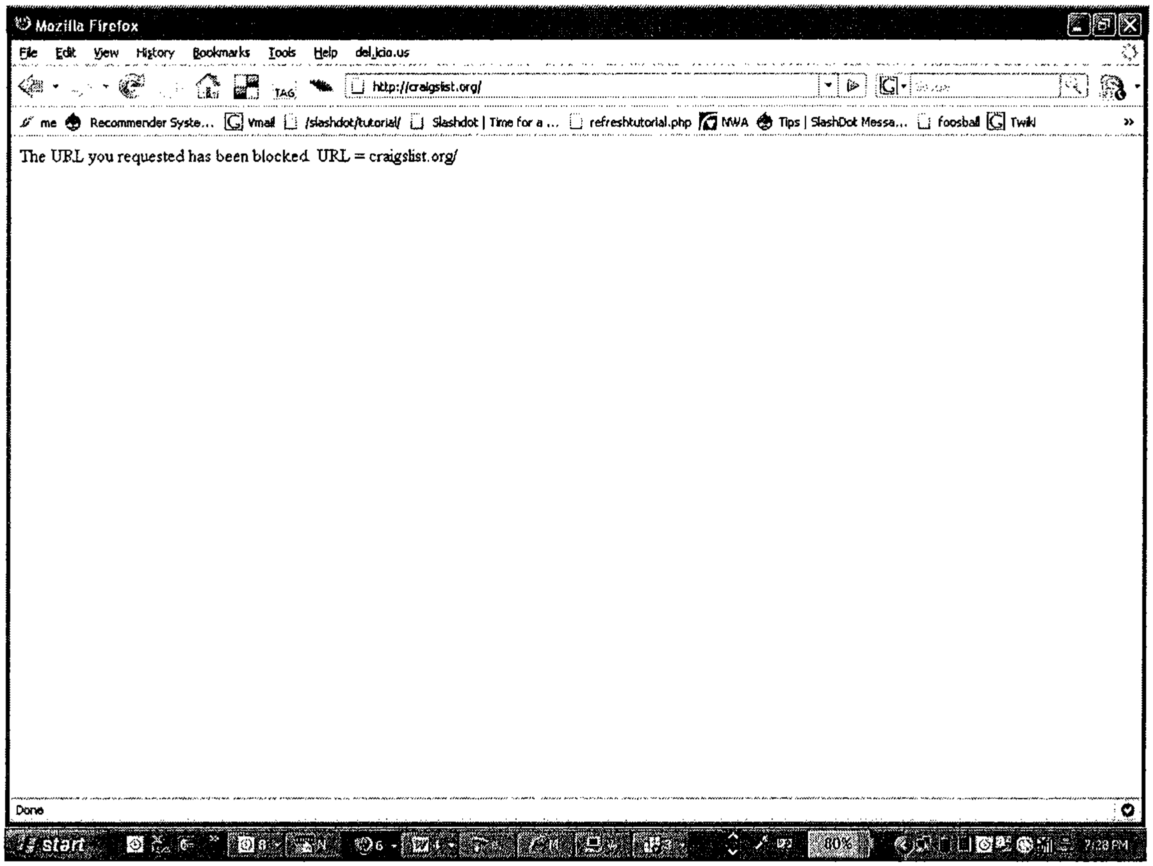
If there is no explicit local rule, then the policy evaluator compares the category returned in Step 4a to the set of blocked categories in the filtering profile. The policy evaluator also compares the classification returned in Step 4a to the set of blocked classifications in the filtering profile. If either the category or classification is blocked, then the policy evaluator determines that the requested access should be blocked; otherwise it is allowed. If the Rating Sever failed to return a category and classification, NCRL has configured the policy evaluator to allow access to the URL.

Logging Step[1]. The FortiGate then sends a line of information to a separate unit, called a FortiAnalyzer. The logfile entry includes the components of the URL (service, such as "http"; hostname, such as "www.yahoo.com"; and url path, such as "/nfl"). It also includes a timestamp for when the access occurred, and which FortiGard (and hence which NCRL branch) the request occurred in. Finally, it includes information about how the URL was classified by Fortinet (both category and class), and whether the request was allowed under the library's policy.

[1] The library had not previously maintained logs. They turned on the logging feature to facilitate this test.

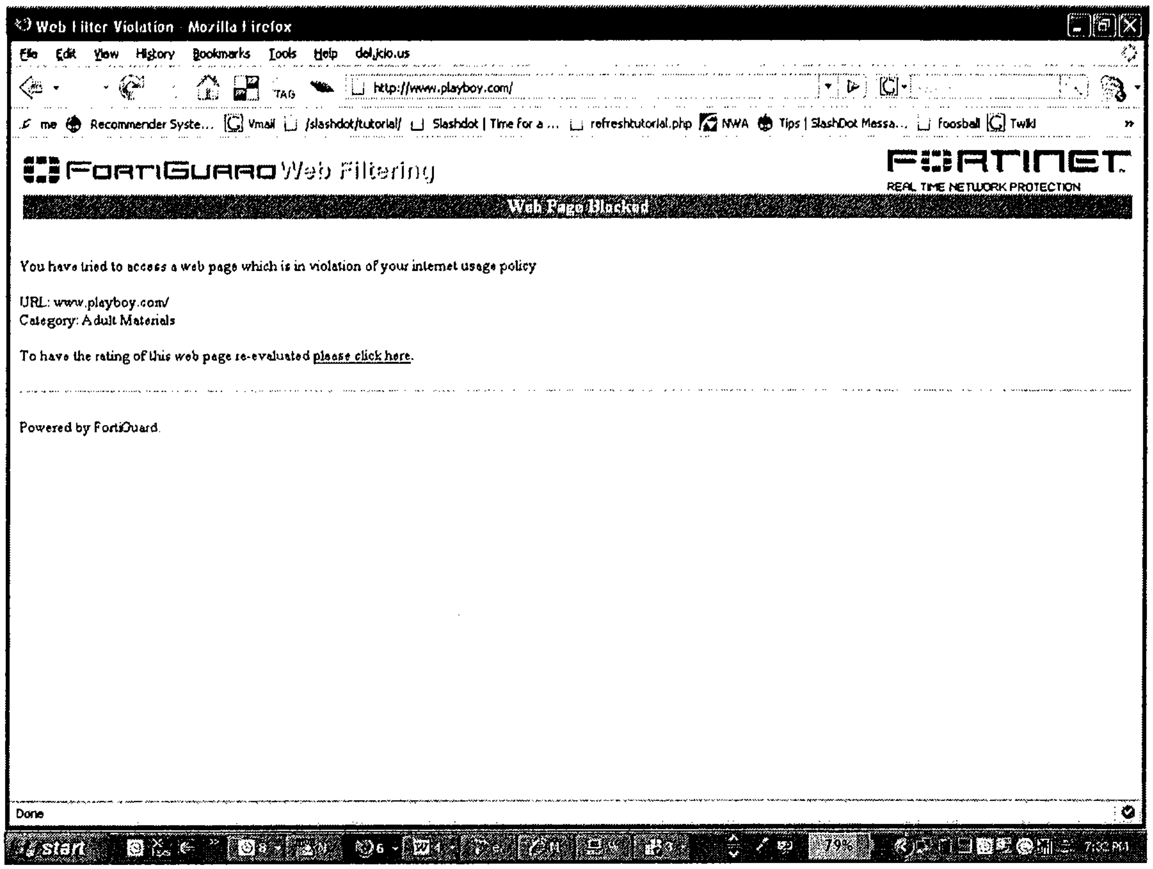
If the policy evaluator says that the access is allowed, then the data retrieved from the destination host are passed on, untouched, to the patron's computer. As far as the patron's computer (and the patron) is concerned, the content is exactly the same as it would have been had the FortiGate not been involved at all (i.e., unfiltered access).

If, on the other hand, the policy evaluator says that access is blocked, then the FortiGate does not send the contents it received from the destination host but instead sends a substitute. The exact content of the substitute depends on whether the blocked contents were an image rather than an HTML page, and on why the item was blocked. The patron's browser is not aware that it has received a substitute instead of what it would have received were the item not blocked. The patron's browser simply displays what it receives just as if that content came directly from the destination host. Figures 2-4 show three possibilities.


Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://craigslist.org/
me Recommender Syste... Vmail /slashdot/tutorial/ Slashdot | Time for a ... refreshtutorial.php NWA Tips | SlashDot Messa... foosball TWiki
Done
start

The URL you requested has been blocked. URL = craigslist.org/

Figure 2. Substitute contents displayed to a patron when a patron requests to visit http://craigslist.org. This is displayed because craigslist.org is one of the sites explicitly blocked by a local rule set by NCRL.


Web Filter Violation - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
http://www.playboy.com/
me Recommender Syste... Vmail /slashdot/tutorial/ Slashdot | Time for a ... refreshtutorial.php NWA Tips | SlashDot Messa... foosball TWiki »
FortiGuard Web Filtering
FORTINET
REAL TIME NETWORK PROTECTION
Done
start
7:32 PM

Web Page Blocked

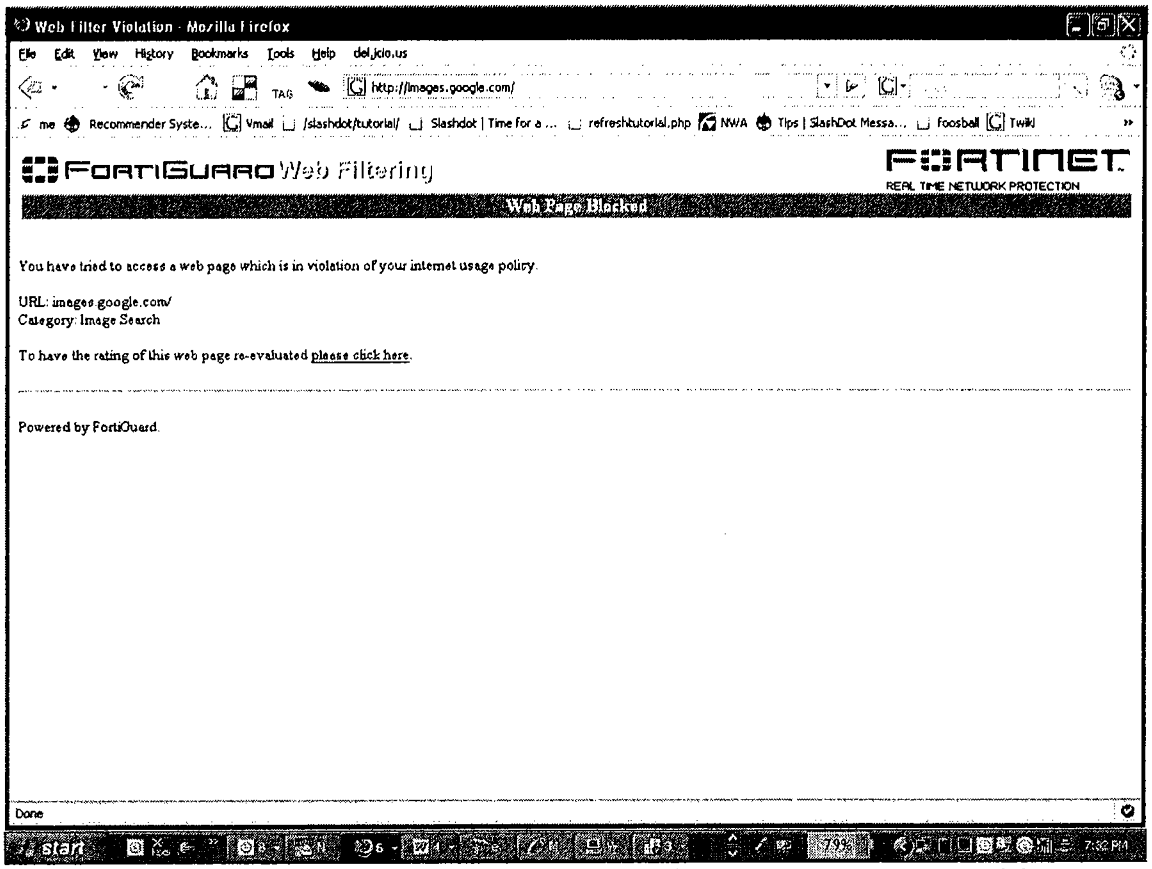
You have tried to access a web page which is in violation of your internet usage policy

URL: www.playboy.com/
Category: Adult Materials

To have the rating of this web page re-evaluated please click here.

Powered by FortiGuard.

Figure 3. Substitute contents displayed to a patron when a patron requests to visit http:://www.playboy.com. This is displayed because the FortiGuard Rating Server said that the site is in the category “Adult Materials”.


Web Filter Violation - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
TAG
http://images.google.com/
me Recommender Syste... Vmail /slashdot/tutorial/ Slashdot | Time for a ... refreshtutorial.php NWA Tips | SlashDot Messa... foosball Twiki
FortiGuard Web Filtering
FORTINET
REAL TIME NETWORK PROTECTION
Done
start
79%
7:32 PM

Web Page Blocked

You have tried to access a web page which is in violation of your internet usage policy.

URL: images.google.com/
Category: Image Search

To have the rating of this web page re-evaluated please click here.

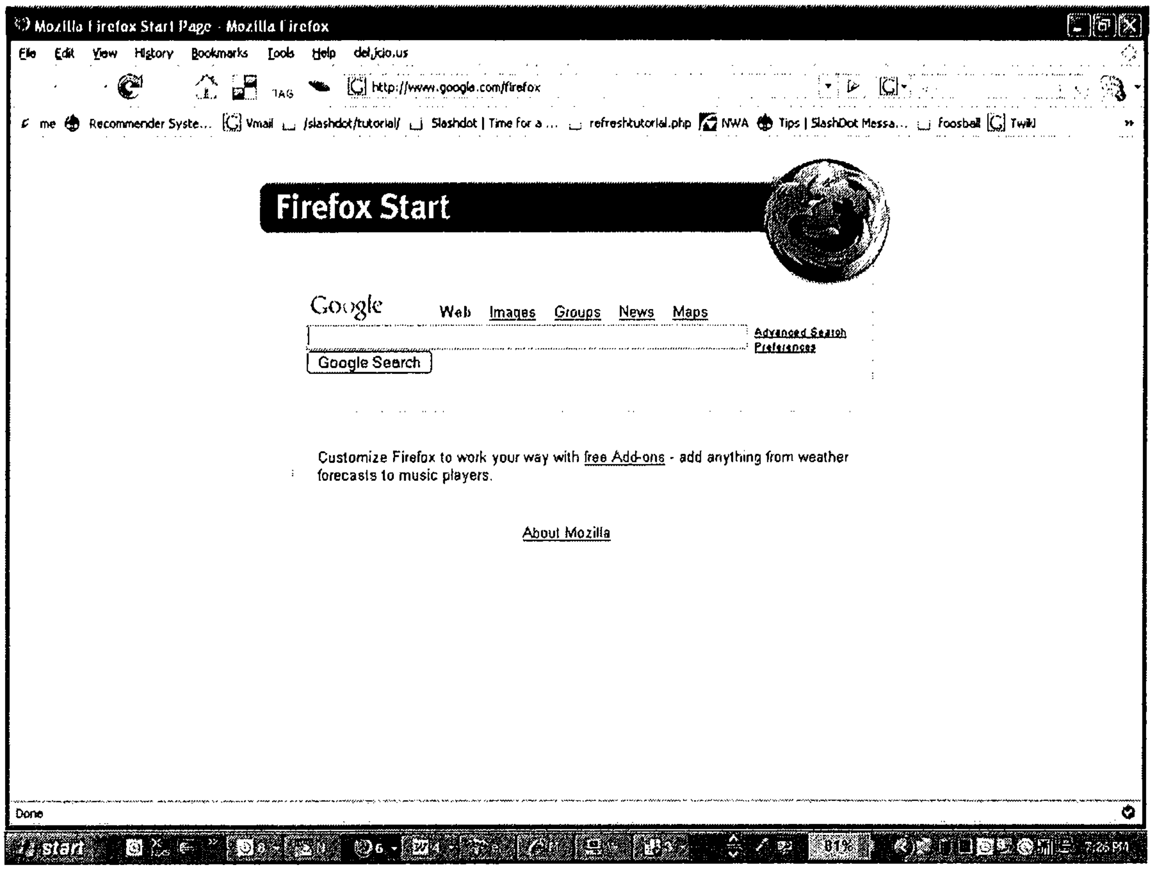
Powered by FortiGuard.

Figure 4. Substitute contents displayed to a patron when a patron requests to visit http:://images.google.com. This is displayed because the FortiGuard Rating Server said that the site is in the class "Image Search". (The information displayed to patrons does not distinguish between categories and classes, and the difference has little bearing on understanding the operation of the web filters or the assessment of error rates.)

Most modern web pages are actually combinations of many separate pieces of information, each involve the fetch of a URL. Consider, for example, the simple page shown in Figure 5. For the browser to display the page, it first requests the url http://www.google.com/firefox. A google webserver returns some text formatted in a special format called HTML. Included in that HTML text is the following line:

```
<img alt="Firefox Start" src="/images/firefox/title.gif" width=440 height=116>
```

When the web browser interprets the HTML text, it realizes that it needs to request this (and several other) images. So it generates a request for the URL http://www.google.com/images/firefox/title.gif. The server will respond to that request by sending an image file. The browser then displays that image in the appropriate place.

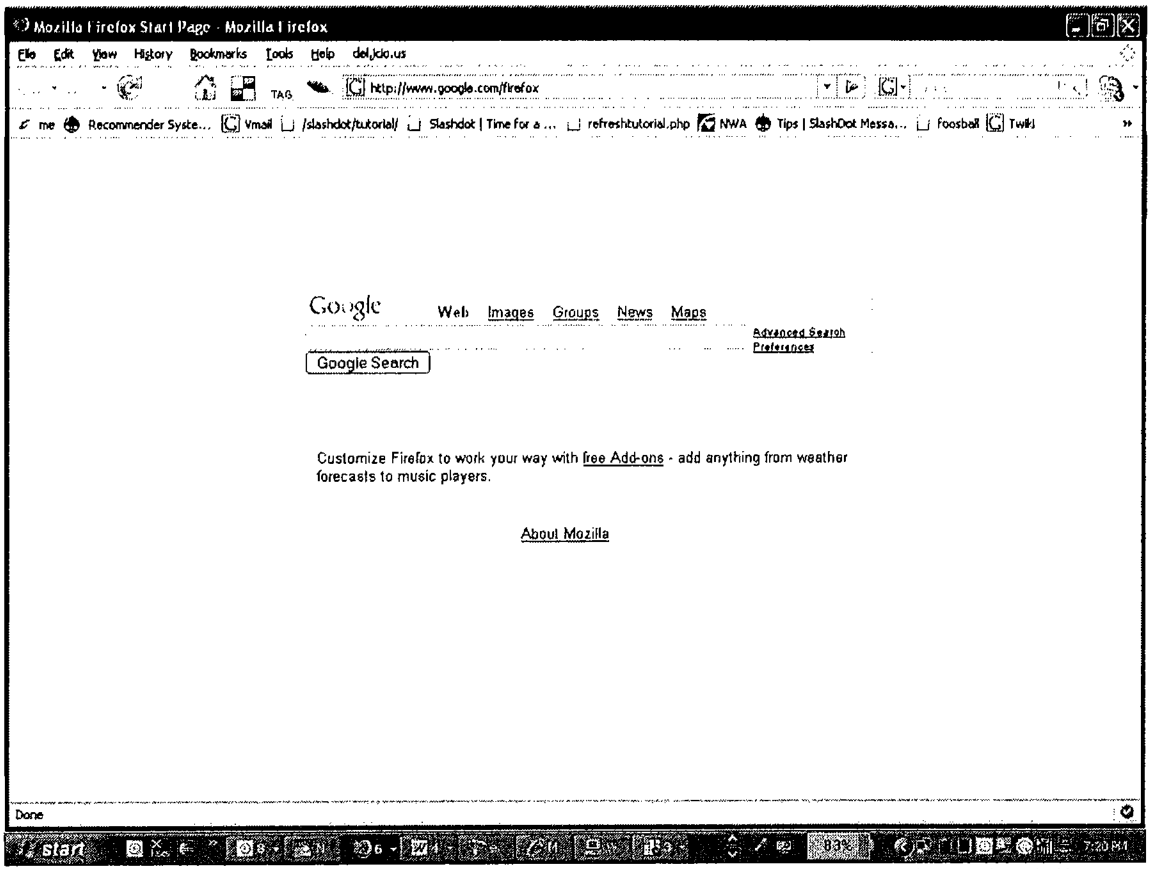

Mozilla Firefox Start Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.google.com/firefox
me Recommender Syste... Vmail /slashdot/tutorial/ Slashdot | Time for a ... refreshtutorial.php NWA Tips | SlashDot Messa... Foosball Twiki
Firefox Start
Google Web Images Groups News Maps
Advanced Search
Preferences
Google Search
About Mozilla
Done
start

Customize Firefox to work your way with free Add-ons - add anything from weather forecasts to music players.

Figure 5. A web page with embedded images that are fetched separately after the initial HTML is retrieved.

In some cases, a URL may not be blocked, but the URL for an embedded image may be blocked. This can occur if the URL for an embedded image specifies a different hostname, or a very different path on the same server. When a policy says that a URL is to be blocked and the destination host returns an image file, instead of passing a blocked message to the patron's computer, the FortiGate substitutes a small invisible image. This has the advantage of not interfering with the patron's interaction with other parts of the page from which the embedded image is blocked. The disadvantage, however, is that, unlike in Figures 2-4, a patron may not realize that anything has been blocked.

For example, we discovered in our blocking test that the Fortiguard rating server seemed to classify all URLs that begin with http://google.com/images as having the classification "Image Search." An NCRL patron visiting http://google.com/firefox would see soemthing like Figure 6, which omits several images, but still allows the patron to enter a Google search term.


Mozilla Firefox Start Page - Mozilla Firefox
File Edit View History Bookmarks Tools Help del.icio.us
http://www.google.com/firefox
me Recommender Syste... Vmail /slashdot/tutorial/ Slashdot | Time for a ... refreshtutorial.php NWA Tips | SlashDot Messa... foosball Twiki
Google
Web Images Groups News Maps
Advanced Search
Preferences
Google Search
About Mozilla
Done
start

Customize Firefox to work your way with free Add-ons - add anything from weather forecasts to music players.

Figure 6. The same web page as in figure 5, but with the FortiGate blocking the delivery of several of the embedded helper images.

NCRL's Blocking Policy

NCRL's staff use a web-based interface to configure the blocking policy and revise it as desired. The process consists of selecting, from among categories and classes defined by Fortinet, which to allow or block. For purposes of my work, I assumed the following categories and classes are being blocked by NCRL. For each, I include the description of the category or class, taken from the Fortinet website. In fact, the blocked categories and classes may now be different, as Fortinet may change the available categories and classes or NCRL may make different choices at any time.

- **Hacking** - Websites that depict illicit activities surrounding the unauthorized modification or access to programs, computers, equipment and websites.
- **Phishing** - Counterfeit web pages that duplicate legitimate business web pages for the purpose of eliciting financial, personal or other private information from the users.
- **Malware** - Sites that are infected with destructive or malicious software, specifically designed to damage, disrupt, attack or manipulate computer systems without the user's consent, such as virus or trojan horse.
- **Proxy Avoidance** - Websites that provide information or tools on how to bypass Internet access controls and browse the Web anonymously, includes anonymous proxy servers.

- **Spyware**- Sites that host software that is covertly downloaded to a user's machine, to collect information and monitor user activity, including spyware, adware, etc.
- **Pornography** - Mature content websites (18+ years and over) which present or display sexual acts with the intent to sexually arouse and excite.
- **Adult Materials** - Mature content websites (18+ years and over) that feature or promote sexuality, strip clubs, sex shops, etc. excluding sex education, without the intent to sexually arouse.
- **Nudity and Risque** - Mature content websites (18+ years and over) that depict the human body in full or partial nudity without the intent to sexually arouse.
- **Gambling** - Sites that cater to gambling activities such as betting, lotteries, casinos, including gaming information, instruction, and statistics.
- **Web Chat** - Websites that promote Web chat services.
- **Instant Messaging** - Websites that allow users to communicate in "real-time" over the Internet.
- **Image Search** – (Class) Websites providing search of images or photos, or the results of image or photo searches.
- **Video Search** – (Class) Websites providing search of video clips or the results of video searches.
- **Spam URL**- (Class) Websites or web pages whose URLs are found in spam emails. These web pages often advertise sex sites, single clubs, and other potentially nuisance or offensive materials.

In addition, NCRL's staff specify some local overrides: some URLs that will be exempt (allowed even if in a blocked category or class), and others that will be blocked (even if not in a blocked category or class). The following hostnames are explicitly prohibited:

- runescape.com
- easyriders.com
- craigslist.org
- images.google.com
- passmyass.com

The following hostnames are explicitly permitted:

- www.pongoresume.com
- Elijahlist.com
- Np.userful.com
- Userful.com
- Userful.ca
- Flickr.com
- Ringo.com
- Pubyac.org

The Overblock Test

To estimate the amount of overblocking (blocking of sites that should be allowed under the library's policy), we conducted a test based on the URLs actually visited at NCRL branch libraries during the week of August 23-29, 2007.

Each time a FortiGard unit considers a URL and decides whether to allow it or not, a line is entered in a logfile on another unit, called the FortiAnalyzer. Unfortunately, the logfile doesn't always include all the information that would be needed to recover, at a later time, what information the destination host sent in response to the request. For example, a communication session may have been established and the host's response may depend on what the previous communication was; the destination host may not respond in the same way to the same URL in the future. In addition, some requests that are sent to web servers involve sending information beyond what is in the urlpath. These pose technical difficulties for conducting a test; sometimes it will not be possible to determine whether a particular request contained in the logfile should have been blocked. Fortunately, for most of the typical requests to examine a web page, the logged URL provides sufficient information; requesting the same URL at a later time will *usually* yield the same type of information that a patron received (or would have received had it not been blocked) and it can then be rated to see if it should have been blocked.

Since the library changed its filtering policy from the time when the logfile was generated, we used the information in the logfile to determine whether the URL would be blocked under the current policy. Some URLs that were blocked at the time (such as myspace.com) would not be blocked under the current policy and thus we treated them for the purposes of this test as if they were not blocked. The purpose of the test was to determine whether the Fortinet filters would make blocking decisions in accord with the current NCRL policies.

Many of the log file entries were not for downloads of initial web pages. They were for things such as flash movies, executable files, or embedded images, stylesheets, or javascript files. Some of these, such as executable files, were not ratable, since it would not be possible for raters to evaluate them without actually executing the files, a dangerous action for the raters given the many computer viruses that are out there. Others, such as stylesheets and javascript files, it makes no sense to rate on their own, outside the context of the webpage they were initially associated with.

Some non-ratable items were identified by the raters, as described below. To reduce the number of non-ratable items they would have to consider, we employed an automated mechanism to eliminate some blocked URLs from consideration. URLs that ended in .exe, indicating an executable file, were removed. In addition, some URLs were removed based on the "mime type" that the web server returned. When a web server responds to a request, it normally sends a text header indicating the "mime type" of the information it is sending. This descriptive text is occasionally inaccurate, but is generally a good indicator. For example, most web pages are labeled with the mime type "text/html" or sometimes "text/plain". Image files in .jpg format are usually labeled with the mime type "image/jpeg". We removed items of mime types "application/json", "application/octet-stream", and "application/x-msn-messenger". The last of these types indicates instant-messenger traffic, a category that should be blocked according to NCRL policy.

The total number of distinct URLs that were requested during August 23-29 was 466,840, from all the branches combined. The number of distinct URLs[2] that would be blocked under current NCRL policies was 2,222. Of these, 42 distinct URLs were removed as executable files or as mime types indicating non-ratable items, leaving a total of 2180 potentially ratable URLs.

In addition, we included 200 distinct unblocked URLs, selected at random from the log data. The purpose was to provide some distractors so that raters would not know whether the URL they were rating was one that the Fortinet would have blocked. Without such distractors, raters might have been biased toward classifying URLs as blockable.

The complete test set consisted of 2380 distinct URLs, 2180 that would be blocked under NCRL's current filtering policy.

Popularity: Google Page Rank

The set of URLs in the test set come from a set of domains that is not a random sample from the universe of all domain names (DNS entries). One measure of the popularity of a domain is the "PageRank" that Google assigns to it. Google determines the PageRank of a page based (in part) on the number of popularity of other web pages that link to a particular page. Google uses these PageRank scores to prioritize search results—pages with higher scores are presented higher in the search results.

Google allows outsiders to request the PageRank of URLs. We developed an automated script to collect these. PageRanks returned are integers from 1 to 10. Some pages Google does not assign a rank to, presumably because they are sufficiently unpopular to not have received a rank yet—we treat those ranks as 0.

We ran the script on the set of blocked domains from Mr. Haselton's expert report. That set, reportedly, was derived by taking a random sample of all valid domain names, then running an automated script to identify all those that Fortinet classifies as Pornography. There were 1848 of these domains in all. Their median PageRank was 0 (that is, more than half were unranked by Google) with a mean of 0.69 and a maximum of 6.

We also ran the script on the 2180 blocked URLs in our test set. To facilitate comparison with Mr. Haseltons set, we checked the PageRank of the base domain rather than the individual page that was accessed (e.g., "http:// imagine-windowslive.com" instead of "http://imagine-windowslive.com/common/resources/bridge.xml.aspx?Locale"). They came from 691 different domains (fewer than the total number of URLs because there were sometimes several different URLs blocked from the same domain.) The median

[2] Our method of selecting all the distinct URLs involved an SQL query that treated URLs as identical if they differed only in use of small letters vs. capitals. In retrospect, this was a mistake on our part. In the logfiles, we found both http://mystatus.skype.com/smallicon/quitter and http://mystatus.skype.com/smallicon/Quitter. Technically, a web server may send a different file in response to requests for these. We rated only the latter. http://mystatus.skype.com/smallicon/quitter was the only URL not rated because of our over-aggressive method of de-duplication.

PageRank for these 691 domains was 2, with a mean of 2.48 and a maximum of 10. For example, "/www.aim.com/" (AOL's Instant Messenger) had a PageRank of 8.

Rating

Inter-Rater Reliability

In order to assess the reliability of a classification scheme, it is good practice to have multiple raters evaluate at least a subset of the items. The ratings of the raters are compared to assess the extent of agreement in their ratings. Researchers do not expect perfect agreement among raters, because classification is always an error-prone activity, but researchers like to see relatively high agreement in order to have confidence that the categories are well-defined. A raw measure of agreement is simply the percentage of items for which they gave the same rating classification.

In cases where one classification is much more frequent than others, however, this raw percentage can be misleading. For example, suppose there are two possible ratings, A and B. Suppose two raters each rate 100 items. Each rates A 95 times and B five times, but there is no overlap in which items they rate as B. As a raw percentage, they still agree on 90% of the cases (where they both rated A) and only disagree on 10%. However, this raw agreement rate of 90% would be misleading, as the two raters never agreed on a classification into the B category.

Social scientists have devised a measure of inter-rater reliability, called Cohen's kappa, that accounts for the "expected frequency of overlap" when one category is assigned most of the time. It measures how much more agreement the raters have than would be expected if they were just rating at random using the observed prevalence of the categories. Kappa scores are fractions between 0 and 1. A rule of thumb for interpreting kappa scores is that scores above 0.8 indicate an excellent level of agreement. Somewhat lower scores are common, however, and published papers have used classification schemes with kappa scores even below 0.5.

Training Procedures

A series of steps was taken before the actual rating of the sites. The first step was to develop a coding scheme that could be used to rate the sites.

The category scheme was developed through an iterative process. A training set of both blocked and unblocked URLs was selected from a day earlier in August, not overlapping with the set. Webpages from the training set were reviewed, classified (according to the most recent classification scheme), and discussed (especially when raters did not agree on the rating). When necessary, new categories and rules for applying categories were developed. For example, the categories for "Helper Images" and "Customer Support IM" were added to the original list. In addition, some blocked categories were grouped into one category when they covered similar content (e.g., Adult sites, Pornography, and Nudity/Risque were grouped together).

The initial coding scheme included the same categories that I was told NCRL was then blocking. We used the exact descriptions provided by Fortinet for each of these categories.
Additional categories including "inaccessible," "not for humans," and "FLAG" (which indicates that the rating should be double checked) were included based on our prior experience. This initial category scheme was refined first by me and Dr. Hansen, after discussing the appropriate rating of webpages from the training sample. The updated category scheme was again refined by the rating team (Dr. Hansen and 2 graduate student raters) after they independently rated sites and identified sites that were hard to rate or that were inconsistently rated.

The final categorization scheme and instructions to raters is documented in Appendix 6. It was discussed by the rating team to assure that each member understood how to apply it. They then practiced applying it again to the training set prior to rating of the URLs from the test site.

Classification of Sites

As part of this study, two graduate students at the University of Maryland independently rated the sample of sites during a 1-week period. Students did not to talk to one another at all about the rating during this period. They also had no idea how the Internet filter had categorized the webpages, in order to keep them from being biased. Furthermore, as mentioned previously, some non-blocked sites were included in the set so that the raters did not assume that a site should fit into one of the "blocked" categories.

The raters were encouraged to talk to Dr. Hansen if any issues arose, but not to provide specific webpages, as he would act as a third rater and did not want to be influenced by their rating of a specific site.

One of the raters asked if she should classify a site that consists of only one image (but not a helper image) based solely on the image or also on the rest of the site within which it was found. She was told to flag all such sites (to assure that they will be reviewed later by a 3rd rater) and to classify them based on the photograph and not the entire site for the time being).

A rater also contacted Dr. Hansen to tell him that she had coded some of the pages while using wireless access from a library and a coffee shop. She was concerned that there had been an Internet filter on the wireless that she had used. The result was that she classified some pages into the "No Session Data" category rather than the appropriate category, as she could not access them. To fix this problem we had her re-rate all of the webpages she had previously rated as "No Session Data", from an unfiltered location. Her updated ratings are used throughout the analysis.

Once all of the ratings were completed by the two raters, Dr. Hansen independently rated all of the webpages where there was a disagreement between the first two raters or where one of the raters had "flagged" the site as difficult to rate. He was not able to see the other rater's classifications so that his rating would be truly independent. Again, the Fortinet classification of these URLs was not provided to Dr. Hansen, so as not to bias him.

Since the main purpose of the study was to evaluate whether the URLs blocked by Fortinet actually fall into one of the categories that NCRL blocks, we abstracted each rater's classification to indicate the following:

1. Was it possible to evaluate whether the URL should have been blocked? Items classified as "No Session Data", "Not for Humans", "Need to Login", "Inaccessible", or "Download Required" were considered Not-ratable.
2. If it was ratable:
    a. Was the URL a helper element such as a small image that would normally be embedded inside a web page rather than being accessed as its own page?
    b. Was the URL in any of the blocked categories or classifications?

Where the two raters disagreed or one had "flagged" the item, Dr. Hansen's rating was used as the final arbiter to determine our classification of the item.

Finally, we revisited all URLs, except for helper images, where Fortinet's classification was in one of the blockable categories or classes and our classification was "not blockable" or where Dr. Hansen's final classification was "blockable" while the two raters initially said it was not. In these cases, Dr. Hansen and Paul Resnick reviewed the URL jointly, with awareness of the Fortinet classification, and came to a consensus. Appendix 7 documents individually all the reclassifications that were done based on this manual checking.

Results

Classifications and Reliability

Non-ratable

Subsequent to the human rating process, we realized that the raters had trouble classifying some types of items that we could automatically identify as non-ratable, as detailed below:

- 12 URLs where the web server returned the mime type "application/x-javascript", "text/css", or "text/javascript".
- 66 URLs from the site omniture.com. All of these seemed to involve passing of session data that was not apparent in the URL, so that visiting the URL did not retrieve the same information that would have been retrieved by the patron.
- 167 URLs where Fortinet classified them as "Spam URL" or "Spyware". "Spam URL" is defined as URLs that are found in spam emails—this was not an assessment that we thought our raters could make just by looking at the URL. Similarly, "Spyware" is defined as sites that covertly download software to the user's computer, something that our raters would not have been able to assess. Our raters were not even asked to identify these categories. Our test is not able to

assess whether Fortinet is correctly blocking when it blocks URLs in these categories.

- 35 URLs where Fortinet classified them as "Hacking", "Phishing", or "Malware". While we asked our raters to assess these categories, they found it difficult and were not able to do it reliably. Thus, we treat items that Fortinet classified in these categories as not ratable. Our test is not able to assess whether Fortinet is correctly blocking when it blocks URLs in these categories.

We examined the agreement of the raters on the remaining 2169 items where their human judgments of non-ratability were used. They agreed on 97.3% of these. In other words, the raters agreed 97.3% of the time whether a webpage was ratable or not. The kappa score was 0.71.

This level of agreement was sufficient to proceed. Where the two raters agreed, their assessment of ratability was used. Where Dr. Hansen independently rated the item, because the raters disagreed or one of them flagged the item, Dr. Hansen's assessment of ratability was used. A total of 96 items were excluded based on the raters' assessments (i.e., they were inaccessible or fell into one of the other non-ratable categories). Based on the manual re-check process involving Dr. Hansen and me, described above, three additional URLs were excluded as unratable. Thus, the total number of ratable items was 2070.

Helper image

It was possible to identify most of the helper images automatically. A computer program downloaded each of the image files and extracted the dimensions of the image, in pixels. Small images were defined as those meeting all of the following conditions:

- One side <150 pixels
- The other side <400 pixels
- Total pixels < 22,500 (for comparison, a 150x150 image has 22,500 pixels)

There were 1338 small images identified in this way, all of which we treated as helper images.

Not all helper images could be identified through this automated means. Some images were in a format (.bmp) that we were unable to automatically extract their dimensions. Others were delivered as the sole contents of an HTML page, presumably intended to be displayed within a frame in a larger page. Because they were not returned as images directly, we did automatically extract their dimensions. Finally, some images larger than our cutoffs were clearly intended to be embedded in a larger page (e.g., background images).

We examined the agreement of the raters on the remaining 732 items where their human judgments of non-ratability were used. They agreed on 89.2% of these. The kappa score was 0.70.

This level of agreement was sufficient to proceed. Where the two raters agreed, their assessment of "helper image" was used. Where Dr. Hansen independently rated the item, because the raters disagreed or one of them flagged the item, Dr. Hansen's assessment of was used. A total of 197 items were identified as helper images based on the raters' assessments. Based on the manual re-check process involving Dr. Hansen and me, described above, four additional URLs were identified as helper images. Thus, the total number of helper images was 1539.

Other Image

We identified other images purely by the mime type returned by the server. Of the remaining 531 URLs, 179 were of one of the following types: "image/bmp", "image/gif", "image/jpeg", "image/png", and "image/x-icon". In addition, 15 were of type "application/x-shockwave-flash", indicating that they were Flash files which, if the browser has an appropriate player installed, will launch an interactive, animated image. Flash files are often used to create ads with movement in them, as well as games. We grouped them along with the images because of their primarily visual nature.

Web Pages

The remaining 289 URLs included 280 of type "text/html" and 9 of type "text/plain". The former are what we normally think of as web pages. For example, the mime type of a page at Yahoo! such as http://www.yahoo.com/nfl is of type "text/html". Text/plain is generally used for web pages consisting purely of text, without links or formatting information.

Other

An additional 48 urls had no mime type at all. This can be caused by a web server being improperly configured or a malformed web page. For example, among these 48 I found a user page at Geocities, a website that hosts individual web pages. There were also a number of URLs ending in .jpg or .gif. We treat these urls separately in the analysis because most of them are not full web pages.

Forbidden

We classified all the ratable sites as either forbidden (i.e., should be blocked based on the library policy) or not (i.e., should be accessible based on the library policy).

As a result of the manual check process described, we identified a number of image search sites, one of the blocked categories. Since raters evaluating individual pages at these sites did not always realize that the site as a whole was an image search site, we automatically classified all the distinct URLs from these sites as forbidden. 578 of the 1539 helper images were classified this way as image search. 120 of the 194 other images were classified this way as image search. 26 of the 289 web pages were classified this way as image search. None of the other (no mime type) urls was classified as image search.

We examined the agreement of the two raters on the remaining 1346 ratable items. Overall, they agreed on 94.6% of these items. In other words, our raters agreed 94.6% of the time on whether or not a page should be forbidden based on the library policy. The kappa score was 0.84. (The automatic classification of the image search sites had little impact. The overall agreement on all 2070 items was 95.1%, with a kappa score of 0.82.)

This level of agreement was sufficient to proceed. Where the two raters agreed, their assessment of "forbidden" or not was used. Where Dr. Hansen independently rated the item, because the raters disagreed or one of them flagged the item, Dr. Hansen's assessment was used. A total of 457 items were identified as forbidden based on the raters' assessments. 613 more items were identified as forbidden based on coming from the identified image search sites. Based on the manual re-check process involving Dr. Hansen and me, described above, 41 additional URLs, from 9 sites, were identified as forbidden.

The table below shows the total number of URLs that we classified as forbidden and not, in the different groups.

	Forbidden	Not	Total
Helper image	662	877	1539
Other image	170	24	194
Web page	269	20	289
Other	5	43	48
Total	1106	964	2070

Blocked-Site Overblock rates

One way to compute error rates for filtering software is to compute the blocked-sites overblock rate.[3] This error rate is based only on the blocked sites, calculating the prevalence of errors in that set. It does not assess the prevalence of errors overall in the user experience, as it does not consider how prevalent blocking is. Where appropriate, I compute the blocked-sites overblock rate for each of the different groups of urls.

Helper images

The results for helper images are as follows.

Excluding non-ratable items, our raters classified 744 out of 1406 blocked helper images as "OK", meaning they were not in any of the forbidden categories or classifications. For helper images, our raters classified only the images themselves, without considering the rest of the site accessible at the same host. Thus, for example, a clickable image advertisement for a porn site would not have been classified as porn unless the image itself was pornographic.

[3] For a discussion of this and other error rates, see the article "Calculating Error Rates for Filtering Software" included as Appendix XX.

Some of Fortinet's classification of these helper images are errors. For example, an image of Google's logo, found at http://www.google.com/images/firefox/google.gif, was blocked as being from an "Image Search" site. It is actually not from Google's image search service. Given our rating system, it is not possible to determine what percentage of the blocks of helper images are in fact errors. There is no meaningful error rate that can be given for these helper images.

Some of the error blocks for these helper images would have a relatively benign impact on the information seeking of library patrons. For example, Figure 5 shows that a patron going to the page http://www.google.com/firefox (the default homepage after installing the Firefox browser), would miss some of the graphics, but still would be able to enter search terms. Omission of other helper images as a result of incorrect classifications by Fortinet could cause more serious degradation of the user experience. For example, if a blocked helper image was a navigation icon, its omission might leave the user with no way to navigate a site. Given the design of our study, it is not possible to determine what percentage of the erroneous blocks of helper images would have an impact on the information seeking experience of library patrons.

Appendix 8 includes a list of the URLs that were blocked in error, along with their Google PageRank.

Other image files

The results for other images are as follows.

Excluding non-ratable items, our raters classified 24 out of 194 blocked full-size images as "OK", meaning they were not in any of the forbidden categories or classifications. The estimated image-files overblock error rate was 12.4%, with a (95%) confidence interval of [8.1% - 17.9%], computed by the statistics package stata/SE version 9.1.

The mean Google PageRank for the domain names of the 170 correctly blocked items was 1.86. The mean PageRank for the domain names of the 24 incorrectly blocked items was 0.79.

Appendix 9 includes a list of the URLs that were blocked in error, along with their Google PageRank.

Web pages

The results for actual web pages are as follows.

Excluding non-ratable items, our raters classified 20/289 as "OK". The estimated blocked-sites overblock error rate was 6.9% with a (95%) confidence interval of [4.3% - 10.5%], computed by the statistics package stata/SE version 9.1.

Of these, 5 were placeholder/advertising/splog sites that contained only outbound links to other sites and may well have been porn sites previously. Leaving those out, we have a ratio of 15/284, or 5.3% of the blocks of URLs for actual web pages were errors. The

(95%) confidence interval for this estimate is [3.0% - 8.6%], computed by the statistics package stata/SE version 9.1.

The mean Google PageRank for the domain names of the 269 correctly blocked items was 3.53. The mean PageRank for the domain names of the 20 incorrectly blocked items was 2.65.

Appendix 10 includes a list of the URLs that were blocked in error, along with their Google PageRank.

Other Pages

Only 3 of the ratable "other" URLs were blocked by FortiGard. Two of these we classified as forbidden; one as "OK". Because of the small numbers, it is not possible to estimate an error rate for this type of item.

Overall Overblock Rate

Another measure of the error rate considers how frequent overblocks are among all page accesses, not just among those that are blocked. I compute this error rate for those URLs that were web pages.

We were unable to collect the mime types for all of the urls in the logfiles for August 23-29 before this report was due. However, there were at least 55,439 distinct URLs of type text/html and 5,514 of type text/plain, for a total of at least 60,000 web pages. With 20 total errors in the week, that means that only 1 out of every 3,000 page views would have resulted in a page being blocked in error. Some additional pages may have had some embedded helper images incorrectly omitted.

Summary of Opinions

The Internet offers a vast array of information resources, not all of it desirable. Any real-time filter that attempts to block access to certain kinds of items, whether it be computer viruses or pornography, will make some errors of commission and some errors of omission. It will block some items that are not truly in the categories that are meant to be blocked, and fail to block some items that.

Rather than focus on whether errors occur, I think the focus should be on their prevalence. When assessing the prevalence of errors, I think it is important to choose a sample of items to test that is representative of the appropriate universe of items for which error prevalence is being investigate. For filters installed in libraries, the most appropriate sample is a representative set of items that were actually accessed. I have found in this study that NCRL patrons tend to access more popular domains, as measured by Google PageRanks, than a random sample of all registered domain names. Moreover, I have found that blocking errors are less likely to occur on more popular domains. Thus, any study based on a random sample is likely to overstate the error rate.

In this study, I found that less than 1/3 of 1% of patron requests for web pages resulted incorrect blocks. Approximately 5-10% of actual blocks were errors, but blocks were infrequent overall. Indeed, during a week, across all the NCRL branches, only 20 pages were incorrectly blocked among more than 60,000 page requests. Similar results hold for requests for full-size images, with an additional 24 incorrect blocks during the week, across all the NCRL branches.

Potentially of greater concern is the greater prevalence of over-blocking for embedded images. 877 different URLs for embedded images were incorrectly blocked, at least according to our classification which was based only on the images themselves. The number may be inflated because a single page typically refers to many helper images and if one is blocked they may all be. Arguably, it may be appropriate to block images that are teasers for click-throughs to adult-only sites, even if the teaser images do not themselves involve nudity. Still, there is reason for some concern on this front that helper images such as navigation icons may be getting blocked incorrectly. Fortinet could probably remedy this situation relatively easily by monitoring frequently checked URLs (navigation icons are likely to be used on multiple pages and thus be accessed quite frequently) and reassessing their classifications. I recommend that NCRL, as a customer, urge Fortinet to take these measures. I do not, however, believe that the error rate on these helper images is sufficient to warrant a drastic move such as removing the filters altogether or shutting off Internet access altogether, if the filters are otherwise accurately reflecting the library's collection policy and if it is legal to attempt to apply that collection policy to the Internet.

Another potential concern with overblocking, besides prevalence, is whether it is systematically biased against certain types of information. For example, five years ago there were concerns that then-popular commercial filtering products (Fortinet was not included in that study) were blocking access to valuable health-related information along with pornographic materials. I conducted a study specifically to assess the overblocking rates specifically for health-related materials. Overall, I did not find in that study that filters were blocking a lot of health information (when configured only to block pornography and not broader categories). I did, however, find that certain types of health information, on controversial topics such as safe sex and homosexuality, were more likely to be blocked, which was a special cause for concern.

I have not been informed of any special concerns about particular kinds of information being blocked in error by NCRL, using the Fortiguard rating service. In reviewing the particular web pages and full-size images that were blocked in error, and glancing through the URLs of blocked helper images, I do not detect any pattern that would suggest to me that the errors are anything but random. Thus, I do not see any reason to believe that particular categories of information are being blocked, other than those that NCRL has explicitly chosen to block.

Exhibit E

CIPA COMPLIANCE

The Children's Internet Protection Act (CIPA), HR4577, is a federal law enacted by Congress in December 2000 to address concerns about access in schools and libraries to the Internet and other information. Fortinet's FortiGuard Web Filtering Subscription Service has been extensively developed to help our customers attain compliance with HR4577 by offering a hosted service that provides Web URL filtering for schools, libraries, government agencies, and enterprise businesses of all sizes. The FortiGuard Web Filtering solution consists of two parts, the FortiGuard Rating Server and the FortiGate Antivirus Firewall. The FortiGuard Rating Server is a master ratings database that is made up of billions of web addresses. The FortiGuard Web Filtering solution does not require any additional hardware and can be activated on all FortiGate Antivirus Firewalls to improve productivity, and reduce substantial risks and limit access to inappropriate and illegal content.

For any school or library that receives discounts for Internet access or for internal connections, CIPA imposes certain requirements. Starting in 2001, the Federal Communications Commission (FCC) issued rules to ensure that CIPA is carried out.

CIPA Requirements

Under CIPA, schools and libraries subject to CIPA do not receive the discounts offered by the "E-Rate" program (discounts that make access to the Internet affordable to schools and libraries) unless they certify that they have certain Internet safety measures in place. These include measures to block or filter pictures that: (a) are obscene, (b) contain child pornography, or (c) when computers with Internet access are used by minors, are harmful to minors; Schools subject to CIPA are required to adopt a policy to monitor online activities of minors; and Schools and libraries subject to CIPA are required to adopt a policy addressing: (a) access by minors to inappropriate matter on the Internet and World Wide Web; (b) the safety and security of minors when using electronic mail, chat rooms, and other forms of direct electronic communications; (c) unauthorized access, including so-called "hacking," and other unlawful activities by minors online; (d) unauthorized disclosure, use, and dissemination of personal information regarding minors; and (e) restricting minors' access to materials harmful to them. CIPA does not require the tracking of Internet use by minors or adults.

Schools and libraries are now required to certify that they had their safety policies and technology in place, or that they were taking the necessary actions to put them in place before receiving E-rate funding for the following school year. After a lower court found the application of CIPA to libraries unconstitutional, the United States Supreme Court upheld the statute in U.S. v. American Library Association.

For further information on Fortinet's CIPA compliant FortiGuard Web Content Filtering subscription service, or to learn more about our award-winning family of FortiGate antivirus firewalls, please contact Fortinet at www.fortinet.com or call us at 408-235-7700 or toll free 866-868-3678.

FORTINET™