Global Information Assurance Certification Paper

The Art of Web Filtering

Robert Alvey
GSEC Practical v1.4b
February 9, 2004

Abstract

It's hard to believe, but just 5 years ago the Internet was typically used by a small number of people who were curious about computers, or those in jobs that required the level of communication provided by the Internet. Since then it has grown exponentially and now most people have a connection to the Internet. A majority of the people who use the Internet do little else but view web pages. One service provided by the Internet is the ability to access web pages. One can find information on any topic known to man, anything from how to make a paper airplane to pornography. Due to the growth of traffic, the lack of central management on the Internet, and to prevent people from seeing offensive or inappropriate material on the internet, software has been developed to block communication with certain sites. This classification of software has been declared "Web Filters".

Web Filters are designed to improve the security and productivity of a network, but as with anything else, it must be implemented correctly to work properly. In order to ensure a Web Filter is implemented successfully, several factors need to be considered. Why is Web Filtering beneficial to an organization? How does a Web Filter work and interact with the network? What can be done to secure or break a Web Filter? This paper is meant to help you understand these factors without focusing on a specific product.

Contents:

Introduction:

One can filter almost anything that is used to communicate with other computers, IP filtering, E-mail filtering, etc… This is mentioned because there is a distinction between the "World Wide Web" (Web) and the Internet. The Internet is a collection of networks tied together to form one big network. The World Wide Web refers to servers on the Internet that service files viewed through the browser. Web pages are transferred using a protocol called HTTP (Hyper Text Transfer Protocol), and a browser is used to view HTTP information. However, other protocols exist, such as SMTP (Simple Mail Transfer Protocol), that are not viewable through a browser. The servers that offer services such as SMTP are part of the Internet, but not part of the World Wide Web. For more information on this subject, Webopedia has an excellent article on the difference between the Internet and the Web which can be found at http://www.webopedia.com/DidYouKnow/Internet/2002/Web_vs_Internet.asp [1].

Before reading further, remember one thing. A Web Filter is not the first step to securing and maintaining web rights. The Web Filter is meant to enforce a pre-existing policy. The first step is to always write a policy and have it approved by management. Without a policy that is adhered too, a Web Filter will never be a solution to the problem.

History of the Internet, the Web:

The origins of the Internet are traced all the way back to ARPANET (Advanced Research Project Agency NETwork), which was formed in 1969. It was originally meant as test bed for technology which would improve communication stability for the Department of Defense. The actual network itself was a series of interconnected nodes between several universities. Through the 70's and early 80's constant research was done to provide more functionality for ARPANET.

In the 80's personal computers (PC) launched into the industry and many people curious about electronics flocked to buy them. Over time modems were released for PCs which allowed them to dial into other PCs and communicate back and forth. This access to affordable personal equipment is what would eventually cause the Internet, as we know it today, to rise. When PCs gained the functionality to connect to others new research and money was thrown into the effort to develop ARPANET as something available for the general public. Eventually that effort led to the definition of the "Internet" October 24, 1995 by the FNC (Federal Networking Council). [2]

The actual World Wide Web was created by Tim Berners-Lee in 1990. While working for CERN, a particle physics laboratory, Berners-Lee developed the world's first Web browser, called WorldWideWeb but eventually renamed to Nexus to avoid name confusion. In 1994 Berners-Lee founded the W3C (World Wide Web Consortium) which develops technical specifications and standards

for the Web, most of which become industry standards. Now we have the World Wide Web as it is today.

Reasons for Web Filtering:

Although Web Filtering software has been around for several years, it has only recently become an issue to note. In the past, unwanted content (better known as spam) was a mere annoyance. Today, spam has risen from its annoyance status to a position where it can now be a legal liability.

For organizations that offer public internet access, like schools or libraries, recent legislation has made it mandatory to implement a filtering solution or lose funding. In the past all efforts to make a mandatory initiative has failed, but since the CIPA (Child Internet Protection Act) was passed, implementing a solution is mandatory. In June 23, 2003 the Supreme Court declared CIPA constitutional after a lower court has ruled against it. [3]. Following that ruling the FCC has declared all public libraries must be in compliance by July 1, 2004. [4]

Other privacy regulations also bring up the issue of Web Filtering, such as HIPAA (Health Insurance Portability and Accounting Act), GLB (Gramm-Leach-Bliley Act), or SEC (Securities and ExChange Act). Those and even more regulations all discuss some manner of filtering content affecting both the public organizations and private industry. A good example in how to get the most out of a web filter is presented through HIPAA. A requirement of HIPAA is to manage and secure all documentation as well as ensuring patient confidentiality. Web filters are capable of preventing users from accessing sites that execute malicious code on the user's computer. A computer infected with malicious code could be used by an attacker to gather patient records or other important information. Another benefit is with applications like SharePoint (Microsoft product that allows users to connect to a central server with a web browser and access documents or communicate with other users), web filters can be used to allow or deny access to SharePoint servers, which add an extra level of security beyond what is provided by the application. If a hospital were using applications like SharePoint, that extra level of security could be a saving grace in the event a attacker manages to penetrate any perimeter defenses.

However, the private sector has even more concerns than privacy regulations or protection acts. To begin with, they have a threat within their own company, employees. Hundreds of thousands of web sites exist that offer content relating to racism, pornography, gambling, or any number of other topics that employees will find offensive. If an employee finds something offensive, and a company has a policy full of holes, lawsuits start to pop up.

Another issue of liability is when an employee downloads illegal material. In this day and age people download illegal MP3's constantly, even at work. Even if the employee thinks there is nothing wrong with their actions, if illegal

material is found on a company PC, the company is liable. Without proper filtering employees can easily download MP3s, DVD movies, applications, or even pornography.

How about productivity? If an employee is looking at web sites that don't support his job function, or downloading illegal material, that employee is not being productive. Web Filtering allows companies to block web sites that don't support their mission or don't fit the requirements in their security policy.

Active & Passive Filtering:

There are four ways to set up a Web Filter on your network; a passive detection system, passive blocking system, active detection system, or active blocking system. Figure 1 shows how a typical network is connected:

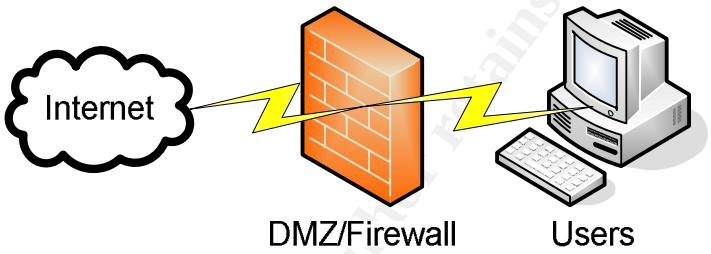

Figure 1

The internal network (Users) connects to the Internet through a De-Militarized Zone (DMZ) which contains a proxy, firewall, or some other device to help protect your internal network from the Internet. The DMZ is an area of your network that typically contains servers which people coming from the Internet need to connect too, like a Web server. Passive and active Web Filters add another device to this design.

The Internet uses TCP/IP as the protocol suite for communication, and a working knowledge of the protocol is recommended to fully understand Web Filtering. However, such understanding is not necessary, and the details of the TCP/IP protocol are outside the scope of this document. For more information on these concepts see the book "TCP/IP Illustrated Volume 1" by Richard Stevens [5]. This book is valuable if you are just interested in a brief overview, or really looking to breakdown the protocol stack.

When the user types in a URL (Uniform Resource Locator) into their browser, their computer sends a signal to the server containing the Web page saying "Give me the page at this URL." The server in turn sends the Web page to the user's computer and it's displayed by the browser. In order to filter a Web site a Web Filter must track what sites the users are going to and stop the user's

computer from sending the signal requesting the Web page or stop the server from sending the web page.

Passive Filter System:

When a user makes a request for a Web page, the request is sent out as usual to the server containing the Web page. The filter system will pick up that request at the firewall and check to see if the user is allowed to view that page as shown in Figure 2.

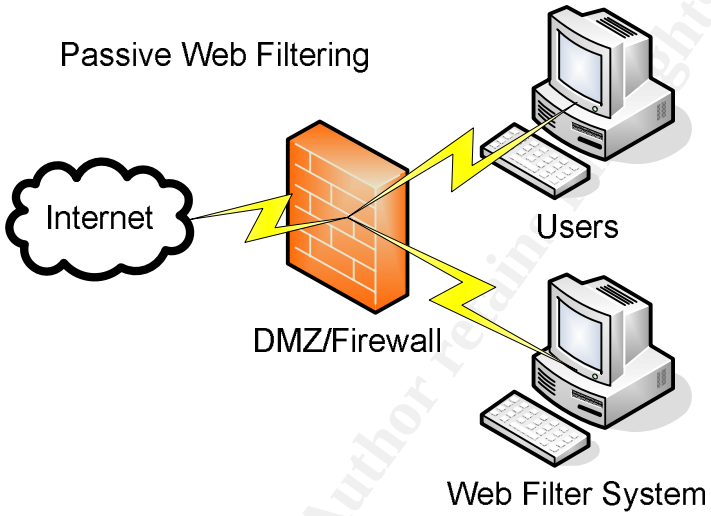

Figure 2

If there is no conflict and the user is allowed access to the site then nothing happens. The server sends the Web page back to the user without a problem. However, if the user is not allowed access to that site there are two methods used to block the Web page. Both methods work by having the firewall query the Web Filter to determine whether the user has the proper access rights to view the web page. However, with the first method, the firewall allows the request through, but it saves a copy of the information to query the Web Filter with. Then when the data comes back from the server, the firewall will block it or allow it based on the Web Filter's reply. The second method involves the firewall actually stopping the user's request from leaving the network, and then querying the Web Filter before allowing the request through or not. Neither method is better than the other; the one used is just a preference of the developers behind the Web Filter.

There are several benefits to this kind of system. One, no changes are made that could damage the physical connectivity of the network. Because the Web Filter does not control the physical connection, if the Web Filter fails network connectivity will not be affected.

However, there are some disadvantages. The firewall system will be required to do more work because it will have to ensure return traffic doesn't

make it to the user when the site is blocked, and it has to make sure all requests made by users are copied to the Web Filter system. Sometimes that is an unacceptable drain on resources. Another problem is that the Web Filter system has to have a way to block traffic, which means allowing the filter access to whatever system is used to block traffic, usually a firewall. This means a trust relationship between those two systems. The trust relationship can create a security risk because if the Web Filter were compromised the attacker would gain not only access to view your web traffic, but also gain a level of influence over your firewall. However, a properly configured network will have the Web Filter system behind the firewall, not in front of it, in that situation an attacker who compromised the Web Filter would have already defeated the firewall. Bandwidth may also be an issue, because the request still leaves the network and a reply comes back with the first method. The second method also includes a delay because the firewall must verify the request through the Web Filter before sending the request along.

Active Filter System:

The difference here is that the Web Filter system now sits, physically, between the users and the Internet, as shown below in Figure 3.

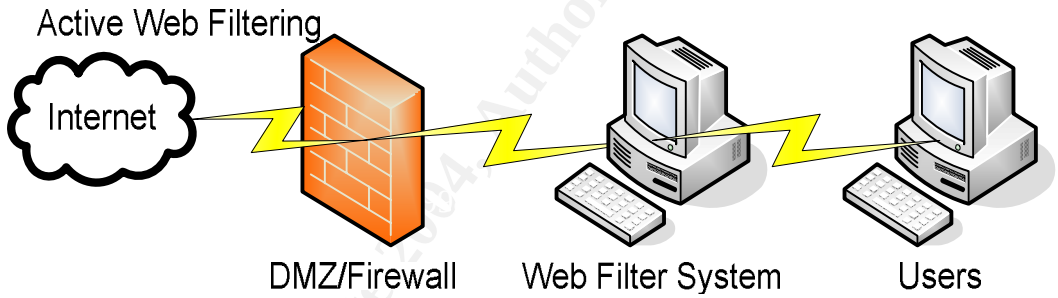

Figure 3

When a user makes a request the Web Filter will receive the request, decide whether or not the user is allowed access and stop or forward the request as needed. As with the passive system if the request is forwarded, nothing else will happen, and the server will reply with the page which the user will receive and view as normal. If the request is denied, then the Web Filter system will reply back saying the Web site can't be accessed.

Active systems are considered the better option if the network/requirements can support one. Because the Web Filter is placed along the path of communication for the user there are no devices wasting extra resources. Since internal network speeds are significantly faster than external network speeds bandwidth should never be a problem.

On the other hand, the disadvantages of an active system can cause a problem. All of the users have to send traffic through the Web Filter and should the Web Filter go down, or not be able to keep up with the traffic, network connectivity will be lost for every user/system that goes through the Web Filter.

This applies to all traffic, including e-mail, FTP, and other critical non-Web related data. Furthermore, if the Web Filter is compromised, the attacking party will have access to all data traveling between a private internal network, and the rest of the world.

There is another aspect of an active system that is entirely software based. These are Web Filter systems that work by installing them on the individual PCs in the organization. The software solutions work like just like the active filter system. The only difference is that the software makes the decision as the request is made by the user, instead of capturing the request after it has left the user's PC. The benefit is that no connectivity is lost should anything happen to the system. The disadvantage is that the software has to be installed, and maintained, on every single PC that has a connection to the Internet. That level of management can be extremely costly if there are more than a few hundred PCs that need the software.

Finally, both active and passive web filters often have a detection or blocking mode. Essentially the difference is one mode only detects which sites the users go to, while the other blocks them from going to the sites. Some organizations prefer to simply prevent their employees from accessing inappropriate sites, hence the blocking mode. Other organizations prefer to deal with employees directly as they access inappropriate sites, or simply wish to use the data as part of a metric, hence the detection mode.

How Web Filtering Works:

By now web filters probably sound like simple and straight-forward devices. Intercept the users request, see the site accessed and decide to block it or not. Well, how is it decided what sites are acceptable and what are not? That's where the contracting companies make their money, because it isn't easy. There are two methods utilized to determine whether a site should be blocked: URL database and keyword/pattern matching.

The first method, a URL database, is where the contracting company, or organization that created the Web Filter, creates a database of URLs that should be blocked. This is very effective and relatively fool-proof because the only margin for error is inputting the URL incorrectly. Already there are several databases available through software vendors that consist of thousands, if not millions, of URLs that are commonly blocked. These are usually porn, hate, gambling, chat, warez, etc… sites because they are sites that nearly every organization wants to block universally. Most of these databases will be categorized for ease of use, that way you can block sport sites, or entertainment sites individually. The only problem with a database is that it's difficult to maintain. The companies that offer those databases have to constantly search for and review Web sites for content and classify them. Given the number of

sites on the Internet, and the amount constantly being created it's a very daunting task to even think of doing.

The second method, keyword/pattern matching, is a bit more complicated. Essentially a Web site is scanned for certain keywords or patterns of words to automate the process of reviewing Web sites. There are two disadvantages to this, it's no where near perfect and if it takes context into account at all it's very poor. Since the artificial intelligence (AI) in the Web Filter application is the one determining whether the Web site is appropriate, and not a human, you can't be sure the site is judged correctly. There are some programs that have low-level AI functions that will take the keywords or patterns and look at other words near them to try and judge how the keywords are used in context. For example, if you had a keyword of "breast" and the site was about breast cancer the AI would see that the word cancer is used near breast consistently and classify the site as medical and not porn. Of course these functions, like most other low-level AI functions, are not capable of working by themselves with a high success rate, it's always best to double-check. Although one might say it defeats the purpose of the function, right now it is the only way to be sure the filter is only blocking the proper sites.

There is also another function offered by some Web Filters sometimes called a "spider". This function runs the keyword or pattern match against a site as your users go to it. This way any unchecked site a user goes too will be checked as the user is surfing the site and added to the database as needed. This provides a dynamic update service for the Web Filter should it be used, but again it comes down to the fact that despite what current vendors would argue, the effectiveness of keyword/pattern matching is limited.

URL and keyword lists come in two forms, white lists and black lists. A white list is a list of URLs or keywords that are allowed through the Web Filter system. If www.google.com was on the white list, and I tried to connect to it, I would be allowed to connect to it. A black list is the opposite. Any URL or keyboard on a black list is blocked from being viewed. If the word "breast" was on the black list, and I connected to a site about Breast Cancer, I would not be allowed to connect because the site had the word "breast" in it.

How a Web Filter detects and blocks pages isn't the only issue to consider, there is also how a Web Filter reports its findings. For some, just having the Web Filter block traffic is enough, but through reporting it's possible to track down individual users who are a consistent problem, build metrics on how effective the Web Filter is, or any number of other functions. Reporting can be done through any number of methods. Web Filters can generate alerts whenever someone goes to a blocked site or you can have reports on all the traffic passed through the system. The alerts can be saved in any number of formats like basic text, or a spreadsheet, or most Web Filters support SQL or Oracle databases for reporting.

User interface also plays an important role. Most Web Filters today use a web site, within the Web Filter system, to make changes to the configuration. Other possible interfaces are available as well, such as a command line through Telnet or SSH, uploading a configuration file to the system, or a proprietary program. It is important to consider the user interface because depending on a network configuration, some options may not be available for use. For example, using Telnet to access the command line interface can cause problems; Telnet is often blocked by firewalls and its use against company policy due to the lack of security features.

Threats to Web Filters:

Security is a very important issue to consider. Even though a Web Filter doesn't seem like a security device, it affects both ends of the spectrum as a part of a network security and a security risk itself. It improves network security by filtering access to web pages, which can contain virus or malicious code. It is a security risk because in order for a Web Filter to work properly it must have some access to the network web traffic, which makes it a greater target or a basic workstation which has little control over anything on the network.

With the passive system the Web Filter must have some method of blocking the incoming traffic. This opens a very large security hole because now the system must have access to a device that is connected to the Internet and is capable of preventing the flow of data. Depending on how a network is implemented and the placement of the Web Filter this can go from a necessary low risk high impact internal security hole, to high risk high impact vulnerability.

In respect to an active system there is another risk involved, the point of failure. Adding in an active Web Filter system introduces a chokepoint on the network where something can go wrong and cause a lot of trouble. Despite the potential problems that could arise from that most companies place the risk as much lower than found with a passive system.

The other aspect of security to consider is how the Web Filter can be defeated. It is inevitable that a user will attempt to access a site, be blocked, and decide to find an alternative method to reaching the site. The user might be someone with enough technical knowledge to make actual attempts to breach the Web Filter, or it may just be someone who stumbles across a solution. Regardless of the circumstances a Web Filter, like any other network security device, can still be defeated. However, there are some common methods to look out for.

The most common method used, although most Web Filters protect against this, is using the actual IP address of a Web site. Since Web sites are common listed by name, like www.google.com (www.google.com is the actual

DNS address, the IP address for Google is 216.239.53.99) it's easy to forget that the name is nothing more than a placeholder for an IP address. Basically when a user enters a DNS address the computer will ask a DNS server to find the IP address for the Web site the user asked for. Because users input the DNS addresses instead of the IP address Web Filters originally worked by filtering just the DNS address. In the past, people discovered that filter systems would not map the DNS address to the IP address so by using the DNS address a site could still be reached as normal. However, this became a well known vulnerability and most Web Filters stop both DNS and IP addresses for a Web site.

Another method involving DNS addresses comes in the form of a period. Computers are very straight-forward and logical, so often times software will accept extra information assuming you also include what the software is looking for. In the case of DNS software you can add a period to the end of the root address for a Web site. For an example take the address www.google.com/about.html. The www.google.com part of the address is classified as the root address, while the "about.html" signifies a specific part of the Google Web site. Instead of typing in www.google.com a user can put in www.google.com. (Notice the extra period on after ".com") No matter which format used a user can access the site. Now when a Web Filter checks the request against a database, it will register the extra period and consider it a different site from the original. If the site the user is attempting to reach is blocked, but the period makes the filter think the user is trying to access a new site that isn't blocked, it will let the user through. This vulnerability is not as well known as the other DNS issue, so there are only a few Web Filters that the vulnerability won't affect.

Encrypted traffic can cause a problem as well. The purpose behind encrypting data is to allow only the sender and receiver understands the contents of the data. Encrypting data is usually a very good idea because it is an effective security measure; however, encryption poses a problem for Web Filtering. The problem with using encryption is that the data looks scrambled when viewed by anyone other than the sender or receiver. When a Web Filter looks at the packet it would see junk traffic not related to web traffic and pass along the packet as usual. One benefit when encountering encrypted is that both the server and the client must use the encryption format and exchange keys. Web sites with offensive or inappropriate content typically don't provide that level of functionality, so if there is a good chance encrypted traffic is legitimate for the purposes of Web Filtering. That is not always the case though, so it's important to always keep track of odd traffic across your network, especially encrypted traffic. However, there are two issues that take encryption to another level, causing a bigger problem. One issue is an anonymous browser, and the other is a proxy.

An anonymous browser is an application that allows users to browse web pages like normal, but it hides the tracks. Operation Systems store tons of

information about what is done on a system to provide more functionality for the user, but this information can be used to trace where they've been on the Web. Some people see that as a breach of privacy and anonymous browsers allow people to retain that privacy. However, the main reason anonymous browsers were born was to allow people to browse the Web without fear of identity theft. Anonymous browsers work by using a proxy; however they are mentioned separately because they provide several other functions to help keep data safe in web exploration. Beyond the functions of a proxy, anonymous browsers clean and manage the information stored on your system about where you have been, ensuring that no sensitive information is saved unless you want it to be.

A proxy is a server located somewhere on the Internet that re-directs data. Data is sent to the proxy server, and the proxy server changes the header information to make it look like the data is coming from the proxy server, and not the original sender. Then when the proxy server receives the data from wherever requested, it sends the information back to the sender. This works as a work-around because if the proxy server is not blocked by a Web Filter one will be able to pass the information along to the proxy server and the proxy server will get the information requested and send it back with no problem. There is a catch here, this will only work if the sender is using encryption to encrypt the data that is sent to the proxy server, and the proxy server's IP address is not blocked. If encryption is not used then a Web Filter can still read the packet and see that one is attempting to get data from a certain site that is blocked. Even though data is sent to a proxy server first, the actual request for information is still part of the data that must be sent and therefore it can be read by the Web Filter if no encryption is used. If the proxy server's IP address is blocked the request will be blocked as well, because it must include the IP address as part of the data for it to actually reach the proxy server. Therefore the Web Filter will read the packet, see the IP address, and know to block it. Anonymous browsers use proxy servers (often with encryption) as part of the overall protection scheme; however one does not need to use an anonymous browser to use a proxy. One last note, proxy servers use ports different than standard web traffic, therefore Web Filters require special configurations in order to detect traffic sent to a proxy.

Using one of the filter solutions installed as software on individual PCs provides a specific threat as well. A user, if they have the proper rights and knowledge, can use task manager to stop the filter software process and thus disable any filtering. Another way is for a user to enter the services folder of Windows and stop or delete the filter software service from running for the same effect. Be sure to identify what rights a user has on their machine for this form of active filtering.

The threats I mentioned here are by no means all that exist. Not only are there other methods out there, but software is never perfect and can often be exploited through vulnerabilities. Those who manage Web Filtering solutions need to be aware that "the true price of peace is eternal vigilance." Always be

sure to keep up to date with solution providers to ensure the maximum level of protection for a network.

Conclusion:

Many believe Web Filtering to be simply and straightforward. Install it, select what sites you want to block, and be done with it. As with any new technology that concept no longer works. Even with a Web Filter it's important to understand and manage your system properly, not just throw it in and turn it on. Take time to review requirements and policies, explore currently available products, and learn how to use those products to their full potential.

With a paper such as this, often recommendations could be made or at the least a listing of software out there on the market. However, there is no such section/list in this document. It's more important to understand how software works, why it works, and what it can do, and then make individual decisions on what software is appropriate. When making choices, keep in mind, contracting companies want to sell their software so they will be all too happy to provide trial access of the software or hardware and any information that could possibly highlight their product. Use that as an advantage in picking out the perfect Web Filtering solution for your needs.

References

[1] “The Difference Between the Web and the Internet”. http://www.webopedia.com/DidYouKnow/Internet/2002/Web_vs_Internet.asp (October 2003)

[2] Federal Networking Council. “Definition of ‘Internet’”. Oct 30, 1995. http://www.itrd.gov/fnc/Internet_res.html (October 2003)

[3] US Supreme Court. “United States v. American Library Association Inc. http://www.supremecourtus.gov/opinions/02pdf/02-361.pdf. June 23, 2003. (January 2004)

[4] Federal Communications Commission. “FCC 03-188” http://hraunfoss.fcc.gov/edocs_public/attachmatch/FCC-03-188A1.pdf. July 24, 2003. (January 2004)

[5] Stevens, W. Richard. TCP/IP Illustrated Volume 1. Addison-Wesley Pub Co. Jan 1994.

[6] Leiner, M. Barry & Others. “A Brief History of the Internet”. Version 3.31. Aug 4, 2000. http://www.isoc.org/internet/history/brief.shtml (October 2003)

[7] N2H2. “An Introduction to Filtering: What to look for when purchasing an Internet Filtering Solution”. 2003. http://www.n2h2.com/pdf/n2h2_content_filtering_selling_guide.pdf (November 2003)

[8] Burt, David. “The Facts on Filters”. Unknown Date. http://www.n2h2.com/pdf/n2h2_content_filtering_selling_guide.pdf (November 2003)

[9] Clearswift. “A Comphrensive Approach to Web Filtering”. June 2003. http://www.us.mimesweeper.com/download/bin/Documentation/A_Documentation_314_676.pdf (November 2003)

[10] Mark E. Schreiber. “Employee E-mail and Internet Risks”. 2001. http://www.asd.net/library/sec_eMailInternetRisk.pdf (Janurary 2004)