Internet Filter Effectiveness: Testing Over and Underinclusive Blocking Decisions of Four Popular Filters

In the wake of the Littleton, Colorado shooting tragedy, public attention has again been focused on the problem of potentially harmful Internet content. Many parents and legislators have proposed that commercially available filtering software is the best way to keep children away from the "red light districts of cyberspace," while also protecting the First Amendment. Civil libertarians and others however, have noted that Internet content filters do not work as advertised, failing to block much dangerous material, while also unjustly blocking benign content. The aim of this paper is to asses these competing claims by rigorously testing the effectiveness of four popular filtering programs: CYBERsitter, Cyber Patrol, Net Nanny, and SurfWatch. The findings of this study suggest that current support for filtering software should be reconsidered.

Keywords: software filters, Internet pornography, Communications Decency Act, public policy

Introduction

Since the Internet came to the fore of public attention around 1994, Americans have been obsessed with the scourge of easily accessed online pornography, violence, and hate speech. Newspaper and magazine articles have fed this fear with titillating stories about pornographic web sites, hate groups, and online sexual predators (Turow, 1999). This perceived abundance of harmful material, has led Congress to pass two laws, the Communications Decency Act (CDA), and the Child On-line Protection Act (COPA) aimed at criminalizing Internet content deemed harmful to minors. In conjunction with these legislative solutions, the software industry has developed its own technological solution, namely content filtering software.

Over the past three years, courts have rejected both the CDA and COPA as unconstitutional restraints of First Amendment protected speech. In overturning these legislative solutions, the courts pointed to the supposedly "equally effective" but "less restrictive alternative" of Internet filtering software as the best way to keep the Internet a safe place for children (Volokh, 1997). As a result, filter technologies have been championed as *the* solution for keeping inappropriate content at the edge of cyberspace, and away from children. These self regulatory, market driven technologies are seen as First Amendment friendly, and far preferable to direct government regulation. No less than the White House has endorsed this idea, noting that "Advanced blocking and filtering technology is doing a far more effective job of shielding children from inappropriate material than could any law (Clinton, 1997)." In keeping with this statement, the White House has aggressively pushed the development and implementation of content blocking software. This push has only intensified in the wake of the Littleton, Colorado shooting tragedy.

In the days following the massacre, the news media uncovered the fact that the shooters frequently used the Internet to access Neo-Nazi and bomb making web sites. In the rush to blame something for the inexplicable killing spree, both the public and politicians cast a collective pointing finger at the Internet. A CNN/USA Today poll conducted shortly after the killings found that 64 percent of respondents said the net contributed to the tragedy (cited in McCullagh, 1999). Responding to this perceived problem, Congress and the White House drafted a flurry of new laws and proposals to curb access to "dangerous" Internet content. Several legislators are aggressively pushing the Childrens' Internet Protection Act (McCain, 1999) which will require all schools and libraries receiving federal funds for Internet access to install blocking software. Another proposed law would require any Internet Service Provider (ISP) with more than 50,000 subscribers to distribute content blocking software (Bloomberg, 1999). Similarly, the executive branch has fully endorsed filters. Speaking about Littleton at a recent conference, FCC chairman William Kennard noted "We need filtering software for families to use on their PC's. Just as you wouldn't send a child off alone in a big city, you wouldn't -- and shouldn't -- let them explore the vast landscape of the Internet without a chaperone (1999)." In a similar speech announcing a joint industry - White House "Parents Protection Site", Vice President Gore noted that filters were the best tool parents could use to protect children from the "free-fire zones and red light districts of cyberspace (1999)."

The Internet content industry has also thrown its support behind filter use. In September 1999, the Bertelsmann Foundation released a major self-regulation proposal which seeks to "protect children online as well as guarantee free speech (1999: 8)." To achieve this end, the proposal calls for the development of a voluntary international content rating and filtering system.

While the public, Congress, White House, and Internet industry may accept that content filters are the way to go, a number of scholars, civil libertarians, and journalists have asked whether these technologies are indeed the best solution to inappropriate Internet content. They point to the fact that content filtering software tends to block a great deal more speech than even government regulation would deem off limits. Further, blocking decisions can be based upon nearly any criteria, and are not open to public or institutional review. Finally, many filters do not even work as advertised, failing to block many objectionable web sites and thus giving parents a false sense of security. In short, Internet software filters championed as effective and First Amendment friendly, would seem to be anything but (Beeson and Hansen, 1997).

Are Filters First Amendment Friendly?

The majority of reports of Internet content filters being both underinclusive (failing to block the worst pornography), and overinclusive (blocking non-sexual, non-violent content), have come from journalists and anti-censorship groups who have used largely unscientific methods to arrive at the conclusion that filters are deeply flawed. A common method used by such groups has been to select a purposive sample of interesting sites and simply see if they are blocked or not by a particular filtering product. For example, the Censorware Project has used this method to expose unjustified blocking of benign web sites by Cyber Patrol (1998) and X-Stop (1998). Similarly, the Center for Media Education tested several filtering programs against a sample of 45 alcohol and tobacco related web sites. Their study found underinclusive performance, and concluded that "stand-alone filters do not effectively screen promotional alcohol and tobacco content (1999: 3)." While such studies are informative, they are limited to narrow areas of the web, and generally suffer from a lack of methodological rigor. The goal of this paper is to improve upon the above studies by applying social science methods of randomization and content analysis to examine the effectiveness of Internet software filters.

Hypotheses

Based on the assertions made by anti-censorship groups and journalists who claim that filters fail to block many "dangerous" sites, while conversely frequently blocking benign content, the following hypotheses are put forward for analysis:

1. Internet content blocking software will be underinclusive. They will fail to block access to sites with "objectionable material."

2. Internet content blocking software will be overinclusive. They will block access to sites with no "objectionable material."

Methods

The hypotheses above beg the question of what is "objectionable" Internet content? To answer this question I used the Recreational Software Advisory Council's Internet rating system or RSACi. RSAC was originally developed by Stanford Communication professor, Donald F. Roberts, to rate the content of video games, and provide parents with a way to protect their children from excessive violence. However, with the advent of the Internet, the system was adapted to allow web site owners to self rate their content. Currently, RSACi is the most popular system for rating content on the Internet, with more than 100,000 web sites using it to self rate (RSAC, 1999).

RSACi contains four content categories (language, nudity, sex, and violence) each with five levels of severity (0, 1, 2, 3, 4). So for example, within the language category, a site may be rated 0 if it contains no objectionable language, 1 if it contains mild expletives, 2 if it has profanity, 3 with strong language, and 4 if it contains crude, vulgar language. Table 1 gives a summary of RSACi's rating categories.

Table 1: The RSACi Rating System

	Violence	**Nudity**	**Sex**	**Language**
Level 4	rape or wanton, gratuitous violence	provocative frontal nudity	explicit sexual acts or sex crimes	crude, vulgar language or extreme hate speech
Level 3	aggressive violence or death to humans	frontal nudity	non-explicit sexual acts	strong language or hate speech
Level 2	destruction of realistic objects	partial nudity	clothed sexual touching	moderate expletives or profanity
Level 1	injury to human beings	revealing attire	passionate kissing	mild expletives
Level 0	none of the above or sports related	none of the above	none of the above or innocent kissing; romance	none of the above

This system was used to rate the content of 200 web sites drawn from three web page samples described below. Only the first page of all sites was rated. Links were not followed to subsidiary pages. The only exception to this rule was on pages that had no other content than an "Enter this site" link, in which case the link was followed, and the first fully developed page was rated.

RSACi rating decisions were then compared to the actual filter performance -- i.e. site blocked, site not blocked -- of CYBERsitter, Cyber Patrol, Net Nanny, and SurfWatch. A site was considered blocked if the filter programs completely denied access to it. Partial blocks, such as word masking were not considered, as they still allow access to the majority of a page.

Each of these filter products was purchased or downloaded, and all were left with their default settings on. Default settings were used due to the theory that few parents customize filter software. The only change made to these programs was to download the most recent blocked sites list from each company. Filters were tested against selected web sites in June 1999.

A site was deemed to contain "objectionable" material if any of its content received an RSACi rating of 2, 3, or 4. Such sites should theoretically be blocked by filter software. Conversely, a site was deemed "not objectionable" if the highest score in all content categories was either 0 or 1. Such sites should theoretically not be blocked. For example, a site with an RSACi score of 0 - language, 4 - nudity, 3 - sex, and 1 - violence, would be deemed "objectionable" because its highest rating was a 4, and it should therefore be blocked.

Using these RSACi-based definitions of "objectionable - not objectionable" our inclusiveness hypotheses can be clarified. A filter was deemed underinclusive if it failed to block sites with a 2, 3, or 4 RSACi rating. A filter was deemed overinclusive if it blocked sites with only a 0 or 1 as its highest RSACi rating.

Web Page Samples

Internet users, including children, come across content through numerous surfing techniques. People stumble across pages through serendipitous surfing, by using search engines, and by using indexes such as Yahoo. As such, I choose to select three different samples of web content to rate for objectionable content, and to test filters against.

The first sample, roughly analogous to serendipitous surfing is a set of 50 randomly generated web pages. On April 15th and 16th, 1999, I used the Webcrawler search engine's, random links feature to produce a sample of 50 English language web sites. Although these sites were randomly provided by Webcrawler, this does not mean they are a random sample of all web content. Because of the web's vast size, currently estimated at some 800 million individual pages, even the most powerful individual search engine only indexes about 16 per cent of the web's content (Lawrence and Giles, 1999). As such, the random sample produced by WebCrawler is only representative of the percentage of the web indexed by the search engine (about 50 million pages).

The second sample, roughly analogous to typical search engine use, is a set of 50 popular search term results. In April 1999, Searchterms.com, a site which tracks the most frequently searched for terms on major search engines, listed yahoo, warez, hotmail, sex, and MP3 as the five most searched for terms. I took each of these terms and entered them into the AltaVista search engine. For each search result, I took only the first ten links generated by AltaVista, thus producing an overall sample of 50 sites.

The final sample, roughly analogous to using a web index, is a set of 100 purposively selected web sites. I intentionally choose a number of web content categories that filters have been shown to have problems with. First I selected the 36 web sites of organizations who filed amicus briefs in the ACLU's challenge of the CDA and COPA. These organizations argued that Congressional legislation would place their content off-limits. However, seeing as some of these sites deal with touchy issues such as homosexuality and safe sex, filters may also deem them inappropriate and thus accomplish the same end as legislation. In addition to the ACLU litigants, I used the Yahoo web site to select content in the following areas: Internet portals, political web sites, feminist web sites, hate speech sites, gambling sites, religious sites, gay pride/homosexual sites, alcohol, tobacco, and drug sites, pornography sites, news sites, violent game sites, safe sex sites, and pro and anti-abortion sites. Five links were selected in each category except pro and anti-abortion sites, where I only selected four to round out the overall sample to 100 sites.

Reliability

I tested the reliability of my use of the RSACi rating system by having four colleagues rate a 21 site subset of the larger 200 site sample. Overall, use of the RSACi rating system was found to be highly reliable. Coders rated sites with perfect reliability seventy-three percent of the time. Additionally, coders only differed by one rating point 12 percent of the time.

Intercoder reliability for each individual RSACi content category; language (Alpha = .92), nudity (Alpha = .98), sex (Alpha = .96), and violence (Alpha = .82), was also extremely high. Finally, intercoder reliability across all RSACi content categories and web sites was found to be excellent (Alpha = .94).

While these results point to a highly reliable coding scheme they may be artificially high due to the large amount of non-objectionable content in the sample. In other words, since the vast majority of sites were rated 0 for all categories by coders, there was little variation in the amount of objectionable content across the sites. This reduces the room for error among coders.

Combined Results

Combining all three of the web site samples described above into one 200 site sample gives us an excellent overview of filter performance. As mentioned earlier, each sample is meant to represent one way in which average Internet users find information. As such, combining all three web site samples represents a rough approximation of the types of information a typical Internet user might come across in the process of surfing the web. While this combined sample is roughly reflective of average Internet use, the generalizability of results is limited due to low sample size, and a lack of completely randomized sites tested.

Among all 200 sites, a relatively high percentage had some form of objectionable content. Again, this is due to the presence of search term, and category areas relating to sex, hate sites, and violent games (see Table 2).

Table 2: All Samples Combined (N=200) Objectionable Content

	objectionable	**not objectionable**
language	27 (13.5%)	173 (86.5%)
nudity	19 (9.5%)	181 (90.5%)
sex	17 (8.5%)	183 (91.5%)
violence	7 (3.5%)	193 (96.5%)
any objectionable (all categories)	36 (18%)	164 (82%)

As shown in Table 8, 18 percent of sites contained some form of objectionable material. Based on this number, a perfectly operating filter should block the 18 percent of objectionable sites within the sample. In terms of overall percentage of sites blocked, Cyber Patrol achieves this goal, by blocking 18 percent of content. However, as we shall see, among this 18 percent, Cyber Patrol blocked a substantial proportion of non-objectionable sites. In other words, its 18 percent of blocked sites were not the 18 percent of objectionable sites in the sample. CYBERsitter was by far the most restrictive filter, blocking 25 percent of all content. SurfWatch and Net Nanny would seem to be underinclusive, blocking 14 and 6 percent of content respectively. Finally, 31 percent of sites were blocked by at least one filter.

Underinclusive and Overinclusive Blocking

Using highest RSACi score obtained as the independent variable, we can see how under and overinclusive each filter was in blocking content in the 200 site sample.

CYBERsitter did the best job of all filters by properly blocking 69 percent of objectionable material. Still this falls well short of its 90-95 percent objectionable content block product claim (Solid Oak, 1999). While CYBERsitter may do the best job blocking inappropriate content, it carries with it the worst record for overinclusive blocks. It blocked 15 percent of sites with no RSACi rated objectionable material (see Table 3).

Table 3: CYBERsitter Over - Underinclusive

	not objectionable	**objectionable**	**total**
not blocked	140 (85.4%)	11 (30.6%)	151 (75.5%)
blocked	***24*** ***(14.6%)***	***25*** ***(69.4%)***	49 (24.5%)
total	164 (100%)	36 (100%)	200 (100%)

Cyber Patrol placed second in correctly blocking objectionable material 56 percent of the time. However, it also overinclusively blocked 9 percent of content with no objectionable material (see Table 4).

Table 4: Cyber Patrol Over - Underinclusive

	not objectionable	**objectionable**	**total**
not blocked	149 (90.9%)	16 (44.4%)	165 (82.5%)

blocked	***15*** ***(9.1%)***	***20*** ***(55.6%)***	35 (17.5%)
total	164 (100%)	36 (100%)	200 (100%)

SurfWatch failed to block 56 percent of objectionable content (see Table 5). A woeful score considering SurfWatch's product literature claims to block 90-95 percent of objectionable material (SurfWatch, 1999). On the flip side, SurfWatch improperly blocked 7 percent of non-objectionable web sites.

Table 5: SurfWatch Over - Underinclusive

	not objectionable	**objectionable**	**total**
not blocked	152 (92.7%)	20 (55.6%)	172 (86%)
blocked	***12*** ***(7.3%)***	***16*** ***(44.4%)***	28 (14%)
total	164 (100%)	36 (100%)	200 (100%)

Finally, Net Nanny performed horrendously in blocking a measly 17 percent of objectionable content. However, of all filters, it blocked the least appropriate material, only blocking 3 percent of non-objectionable content (see Table 6).

Table 6: Net Nanny Over - Underinclusive

	not objectionable	**objectionable**	**total**
not blocked	159 (97%)	30 (83.3%)	189 (94.5%)

blocked	***5*** ***(3%)***	***6*** ***(16.7%)***	11 (5.5%)
total	164 (100%)	36 (100%)	200 (100%)

With all blocking decisions combined, filters correctly blocked objectionable material 75 percent of the time. On the other hand, they also overinclusively blocked 21 percent of non-objectionable material (see Table 7).

Table 7: All Filter Combined Over - Underinclusive

	not objectionable	**objectionable**	**total**
not blocked	129 (78.7%)	9 (25%)	138 (69%)
blocked	***35*** ***(21.3%)***	***27*** ***(75%)***	62 (31%)
total	164 (100%)	36 (100%)	200 (100%)

Discussion

Support for both of my under and overinclusive hypotheses was clearly found. Put simply, taken all together, filters failed to block objectionable content 25 percent of the time, while on the other hand, they improperly blocked 21 percent of benign content. If we assume the web has 800 million unique documents, filters would incorrectly block approximately 168 million pages (note: this inference has limited validity due to the lack of a truly random sample). Just imagine the outrage if your local library incorrectly removed 21 percent of its books, and then gave no explanation for their removal, nor made public the book titles removed! This is exactly the reality created by the filters reviewed above.

These results point to a profound conflict for parents and policy makers considering the adoption of content blocking filters. They can purchase filters such as CYBERsitter and Cyber Patrol which correctly block large percentages of objectionable web content, but at the same time also block significant amounts of appropriate Internet material. These overinclusive filters cause particular damage to any content dealing with gays, safe sex material, and left leaning political groups.

If parents and policy makers are unhappy with overinclusive blocking they could go with SurfWatch and Net Nanny. Unfortunately, these products let through a tremendous amount of objectionable material.

This catch-22 situation brings current support for Internet content filtering into question. In Reno v. ACLU (1997) the Supreme Court noted that content filters were an effective, and less restrictive means for shielding children from objectionable content, while maintaining access to other non-dangerous content. Yet, as the results above show, filters are (1. not effective, and (2. not less restrictive. They fail to block access to significant amounts of pornography, hate speech, violence, etc., but at the same time make indefensible blocks of political sites such as the White House (blocked by Net Nanny).

Based on these tremendous problems, governmental support for Internet content filters should be seriously reconsidered. Similarly, parents should think twice about the benefit of spending $30, plus update fees, for products which will not protect children from significant portions of "dangerous" Internet material.

Methodological Improvements

To my knowledge, no other study has attempted to combine a content analysis with the blocking performance of Internet filters. As such, the results presented above represent a first attempt at using such a methodology. Future uses of this methodological framework would greatly benefit from several improvements.

First, a larger random sample of web pages (say 1,000+) would improve the generalizability of results. While this sounds straight forward, future researchers must develop a better way of achieving a truly random sample of the vast universe of web pages. A promising methodology for future studies may be to use a random sample of possible Internet Protocol (IP) addresses (Lawrence and Giles, 1999).

Also associated with the idea of a larger random sample, is the fact that such a sample would likely contain little "objectionable" material that parents would want filtered. Thus, it would fail to test filters against the pornographic and violent content that filters are meant to block. This would point to the use of a second purposive sample, perhaps derived from usage statistics of what sites adolescents attempt to access. This test, although less representative of the overall universe of web pages, would be representative of the universe of pages that adolescents -- the group that filters are meant to protect -- typically attempt to view.

With regards to rating web content for "objectionable" material, it is possible that RSACi is not the best system. It only covers four categories, and some of its definitions are not particularly clear. This problem becomes evident when evaluating blocking decisions about alcohol and gambling related sites by Cyber Patrol and SurfWatch. Both types of sites received non-objectionable RSACi ratings, but were blocked due to internal off-limits categories in both filters. Basically, RSACi failed to capture the fact that alcohol and gambling sites may be dangerous to children. To remedy this flaw, a more inclusive system for rating Internet content should be developed. Similarly, future studies using RSACi as the coding system, should have better controls for additional internal blocked content categories used by filters. Such controls would allow for a better assessment of overinclusive blocking.

Conclusion

This study sought to provide objective evidence of Internet software filter performance. As many journalists and civil libertarians have speculated, filters are not a particularly effective technology for protecting children from objectionable Internet content. Further, such programs also block a substantial percentage of web pages with no objectionable material. Overall, filters failed to block objectionable content 25 percent of the time, while on the other hand, they improperly blocked 21 percent of benign content. Given these problematic results, parents and legislators should rethink their current support for the use of Internet filtering technology.

References

Beeson, A. & Hansen, C. (1997). Fahrenheit 451.2: Is cyberspace burning? ACLU. Available via the World Wide Web at http://www.aclu.org/issues/cyber/burning.html .

Bertelsmann Foundation. (1999, September). Self-regulation of internet content. Available via the World Wide Web at http://www.stiftung.bertelsmann.de/internetcontent/english/download/Memorandum.pdf .

Bloomberg News. (1999, 13 May). Senate unanimously passes filtering bill. Bloomberg News. Available via the World Wide Web at http://www.news.com/News/Item/0,4,36540,00.html .

Censorware. (1998). Cyber Patrol and Deja News. The Censorware Project. Available via the World Wide Web at http://www.censorware.org/reports/dejanews.html .

Censorware. (1998). The X-Stop files: Deja voodo. The Censorware Project. Available via the World Wide Web at http://www.censorware.org/reports/x-stop.html .

Center for Media Education. (1999). Youth access to alcohol and tobacco web marketing: The Filtering and rating debate. Available via the World Wide Web at http://www.cme.org/ .

Clinton, W. (1997, 16 July). Remarks by the President at event on the e-chip for the internet. The White House Office of the Press Secretary. Available via the World Wide Web at http://www.whitehouse.gov/WH/News/Ratings/remarks.html .

Gore, A. (1999, 5 May). Remarks on the Internet. White House Office of the Press Secretary. Available via the World Wide Web at http://www.whitehouse.gov/WH/New/html/19990505-4219.html .

Kennard, W. (1999, May 4). Remarks of William Kennard at the Annenberg Public Policy Center conference on Internet and the family. Available via the World Wide Web at http://www.fcc.gov/Speeches/Kennard/spwek916.html .

Lawrence, S. and Giles, C.L. (1999, July). Accessibility of information on the web. Nature.

McCain, J. (1999). Childrens' Internet protection act. Senate Bill 97. Available via the World Wide Web at http://www.thomas.loc.gov/ .

McCullagh, D. (1999, 23 April). Looking for something to blame. Wired News. Available via the World Wide Web at http://www.wired.com/.

Recreational Software Advisory Council. (1999). About RSACi. Available via the World Wide Web at http://www.rsac.org/ .

Reno v. ACLU. (1997). 117 S.Ct. 2329.

Solid Oak. (1999). Product literature. Available via the World Wide Web at http://www.cybersitter.com/ .

SurfWatch. (1999). Product literature. Available via the World Wide Web at http://www.surfwatch.com/ .

Turow, J. (1999, May). The Internet and the family: The View from parents the view from the press. Annenberg Public Policy Center of the University of Pennsylvania, Report No. 27.

Volokh, E. (1997). Freedom of speech, shielding children, and transcending balancing. Supreme Court Review, 31.