The Effectiveness of Internet Content Filters

Philip B. Stark

Department of Statistics
University of California, Berkeley

Distinguished Lecture
Center for Security, Reliability, and Trust
University of Luxembourg
13 July 2012

Background

http://youtu.be/cNARJPNz2CA

http://youtu.be/1bQo05WkHyc

http://www.nytimes.com/2012/05/22/science/big-data-troves-stay-forbidden-to-social-scientists.html?_r=1&nl=todaysheadlines&emc=edit_th_20120522

- Study commissioned by USDoJ re Child Online Protection Act of 1998 (COPA).
- Apologies: stale data. 2005–2006. Required subpoenas of Google, AOL, MSN, Yahoo!
- Think it's still only statistical study of prevalence of pornography & effectiveness of filters.
- Attempts to legislate protection of minors: CDA, CIPA, COPA.
- I worked primarily on COPA; a little on CIPA.
- Team at CRAI led by Paul Mewett collected and categorized webpages and ran filter tests.
- I designed experiments, drew random samples, analyzed data.

COPA

- 2nd attempt to legislate protection from commercial "harmful-to-minors" content
- NOT ABOUT CHILD PORNOGRAPHY
- Exemptions for literary, artistic, and educational content, ISPs, search engines.
- Requires age screen for commercial porn.
- Credit card number deemed adequate proof of age.

Supreme Court

- COPA made several trips to the Supreme Court.
- Feds have legitimate interest in protecting children.
- COPA potentially “chilling” of free speech.
- DoJ had to show that COPA is “least restrictive alternative.”
- How well do filters work?

My job was to figure out:

- How much pornography is there on the Internet?
- How often do people come across it?
- How effective are filters at blocking it?
- How much “clean stuff” do filters block?

How to test filters?

- Underblocking, over blocking (Type I and Type II errors; “precision” and “recall”)—but of what universe of pages?
- Population of pages matters. What’s relevant?
- How do you draw a representative sample of those pages?

Subsets of the Internet

- What's there (including the "dark" web)
- What's readily found (search indices)
- What's actually found (search results)
- What people see most often (results for popular queries)

Data sources

Browsers speak other protocols, but `http` and `https` dominate.

Moreover, Internet largely mediated by search engines.

People use the search box instead of the URL box:
Most popular search term was “google.”

Suggests the indexed web comprises most of what people see, and that what queries retrieve likely reflects how often people see it.

Different for illegal content like child pornography—it tries to be invisible.
But commercial adult sites try to be highly visible to be profitable.

Acquiring data

Initially tried to use Neilson NetRatings data; useless for this.

Considered doing our own crawl

Purchased list of popular search terms from Wordtracker

Subpoena of AOL, Google, MSN, Yahoo!
Explicitly said *no* PII: no IP addresses, usernames, cookie data, etc.

Surprise!

- Google fought subpoena; led to lots of publicity
- Phenomenal amount of hate mail:
  Attacks on my motives, ethics, and physical appearance
  Many expletives.
- Claims of conspiracy by DoJ to spy on search behavior
  Fears of law enforcement from data mining of searches
  Apparently unaware of NSA . . .
- Lack of public understanding of privacy (and lack of privacy) on the Internet
- Queries are not private, nor are visited URLs
- Google was giving data to Chinese government at the same time
- The law was passed under Clinton, although Bush was in office during trial

Tame email 1: Dorothy

well now good for you – instead of teaching parents/caregivers of minors how to block unwanted porn sites you have given this administration an EXCUSE to peruse search engine data bases.

enough erosion of civil liberties

Dorothy [last name deleted to protect her privacy]

[DELETED]@comcast.net

Tame email 2: Paul

The Google user is an actual person, not just a statistic, and your attempt to expose my personal information (even buried in a large quantity of data) is at best short sighted on your part. It is also annoying. It is absolutely NONE OF YOUR BUSINESS what I search for on Google.

I am aware of the fact that some people (especially the young) seem to place no value on privacy. But this is not the case for everyone. Do you think for a minute that the government will be satisfied with ”anonymous” data if it sees ”suspicious” patterns? Using statistical methods to identify criminals has enormous potential for misuse. Look at the early use of genetics that produced eugenics. Before you accept your next consulting fee, stop and talk with someone about the ethics of your work.

Even if you do not value your personal privacy in this matter, ask yourself if you would want the public or the government examining all of your communication or internet use. When the government gains the right to watch our private non-criminal lives, this power will not exist only for the current well meaning Bush administration but will be available for the next Bush, Clinton or Nixon as well.

It is absolutely NONE OF YOUR BUSINESS what I search for on Google. It is none of my business whether the baseball cap just looks cute or is hiding thinning hair. Some things are private.

The Subpoenas

Judge ruled DoJ could get sample of Google's index but no queries.

Long process of specifying/negotiating how to draw the samples from AOL, MSN, Google, Yahoo!

Ultimately received simple random samples from indices and a "representative" week of queries.

- Random sample of 50,000 webpages from Google search index in 2006. (Pages users might find.)
- Random sample of 1 million webpages from MSN search index in 2005. (Pages users might find.)
- Week of search queries from AOL, MSN and Yahoo! by subpoena, about 1.3 billion (Pages users do find.)
- 685 most popular queries from Wordtracker 11/12/05–2/20/06. (Pages users find most often.)

Categorization of Pages

Team at CRA International attempted to view and categorize

- 39,999 random webpages from MSN index
- 11,000 random the webpages from Google index
- first 10 results of each of a stratified random sample of 7,541 queries (total weight 15,461)
- first 10 results of the 685 Wordtracker searches

Raw results

- 68,150 webpages of which 63,105 worked.
- 60,833 Category 1a: no reference to sex and no nudity.
- 1,382 Category 5f: adult entertainment.
- 890 in other categories, e.g., show genitalia in an artistic or educational context.

Drew random samples of the Category 1a pages to test filters.

Nonparametric lower confidence limits

Sample is without replacement (hypergeometric distribution of counts), but population so large that binomial is good approximation.

Moreover, Binomial more dispersed than the hypergeometric, so confidence intervals are conservative.

Find lower confidence limit by inverting binomial hypothesis tests: Smallest p such that chance of observing at least as many as actually observed is at least 5%.

$$\min \left\{ p : \sum_{j=j_0}^{n} \binom{n}{j} p^j (1-p)^{n-j} \geq 0.05 \right\}. \tag{1}$$

n is sample size; j_0 is number of pages in the sample with property.

Confidence bounds based on weighted queries

Pages with adult content (5f) had higher mean weights than others.

Ignoring weights biases estimated prevalence downwards.

To find lower confidence limits, treat queries as if all had weight 1; then use binomial.

Result is conservative.

Sizes of populations and samples. Searches weighted by frequency.

	Google index	MSN index	AOL, MSN & Yahoo! searches	Wordtracker searches
pages in sample	11,100	39,999	22,405	206 million
working pages in sample	10,009	36,557	21,870	195 million
queries in population			1.3 billion	20.6 million
queries in sample			2,345	20.6 million

Estimated prevalence of adult pages

Source	Google index	MSN index	AOL, MSN & Yahoo! searches	Wordtracker searches
adult webpages	1.1%	1.1%	1.7%	14.1%
domestic adult webpages	44.2%	56.7%	88.4%	87.4%
searches with adult results			6.0%	37.1%
searches with domestic adult results			5.7%	37.0%

Conservative 95% lower confidence limits

	Google index	MSN index	AOL, MSN & Yahoo! searches
adult	1.0%	1.0%	2.5%
domestic adult	0.4%	0.5%	2.2%

Estimated underblocking & overblocking rates

Filter	Underblocking		Overblocking	
	Google	MSN	Google	MSN
AOL Mature Teen	8.9%	8.6%	22.6%	23.6%
MSN Pornography	16.8%	18.7%	19.6%	10.3%
MSN Teen	17.7%	20.5%	21.9%	18.9%
ContentProtect Default	38.3%	45.4%	2.8%	3.0%
ContentProtect Custom	28.3%	46.7%	1.4%	0.7%
CyberPatrol Custom	31.0%	33.5%	1.4%	0.9%
CyberSitter Default	12.7%	16.5%	3.6%	4.1%
CyberSitter Custom	12.4%	18.9%	4.0%	3.7%
McAfee Young Teen	16.1%	26.0%	12.4%	13.2%
Net Nanny Level 2	44.0%	46.1%	3.3%	2.2%
Norton Default	60.2%	54.9%	1.4%	0.7%
Norton Custom	58.4%	54.2%	0.9%	0.4%
Verizon	41.8%	40.3%	9.4%	5.7%
8e6	18.3%	23.0%	9.4%	7.5%
SafeEyes	16.2%	15.2%	3.3%	3.2%

Conservative 95% lower confidence limits

Filter	underblocking		overblocking	
	Google	MSN	Google	MSN
AOL Mature Teen	5.6%	6.5%	18.4%	21.0%
MSN Pornography	12.1%	15.7%	15.8%	8.5%
MSN Teen	12.8%	17.4%	17.8%	16.6%
ContentProtect Default	31.3%	41.3%	1.5%	2.1%
ContentProtect Custom	22.2%	42.6%	0.6%	0.4%
CyberPatrol Custom	24.6%	29.7%	0.6%	0.5%
CyberSitter Default	8.6%	13.6%	2.1%	3.1%
CyberSitter Custom	8.4%	15.9%	2.4%	2.7%
McAfee Young Teen	11.4%	22.5%	9.3%	11.3%
Net Nanny Level 2	36.8%	41.9%	1.9%	1.5%
Norton Default	52.9%	50.7%	0.6%	0.4%
Norton Custom	51.1%	50.1%	0.4%	0.2%
Verizon	34.7%	36.2%	6.7%	4.4%
8e6	13.1%	19.6%	6.7%	6.0%
SafeEyes	11.4%	12.3%	1.9%	2.3%

Of adult pages not blocked, estimated percentage that are domestic

Filter	Google	MSN
AOL Mature Teen	40.0%	40.6%
MSN Pornography	31.6%	42.9%
MSN Teen	40.0%	37.7%
ContentProtect Default	39.0%	45.8%
ContentProtect Custom	40.6%	47.1%
CyberPatrol Custom	48.6%	44.0%
CyberSitter Default	50.0%	32.8%
CyberSitter Custom	57.1%	36.2%
McAfee Young Teen	44.4%	37.5%
Net Nanny Level 2	41.7%	48.1%
Norton Default	35.3%	49.3%
Norton Custom	36.4%	49.7%
Verizon	37.0%	42.4%
8e6	42.1%	46.8%
SafeEyes	35.3%	40.4%

Estimated underblocking & overblocking
AOL, MSN, & Yahoo! search results

filter	underblocking for results	overblocking for results	domestic underblocking	underblocking for queries	95% confidence limit
AOL Mature Teen	6.2%	12.5%	57.0%	15.6%	5.3%
MSN Pornography	21.4%	4.4%	86.1%	32.3%	20.9%
MSN Teen	20.8%	5.8%	91.9%	28.1%	18.8%
ContentProtect Default	18.4%	6.4%	70.1%	46.2%	10.0%
ContentProtect Custom	20.4%	0.0%	62.1%	42.2%	25.4%
CyberPatrol Custom	34.6%	0.4%	94.9%	65.6%	24.4%
CyberSitter Default	11.2%	4.6%	33.8%	23.2%	11.2%
CyberSitter Custom	10.0%	5.3%	44.1%	20.1%	8.1%
McAfee Young Teen	14.2%	20.7%	80.7%	30.9%	10.4%
Net Nanny Level 2	28.1%	3.7%	79.4%	36.6%	20.8%
Norton Default	42.1%	0.8%	85.3%	51.6%	49.3%
Norton Custom	43.4%	0.0%	85.6%	56.1%	54.3%
Verizon	23.1%	1.3%	80.9%	41.6%	31.4%
8e6	7.3%	7.5%	78.0%	23.4%	11.7%
SafeEyes	13.7%	1.9%	87.8%	29.8%	14.9%

Underblocking & estimated overblocking for Wordtracker query results

filter	underblocking for results	overblocking for results	domestic underblocking	underblocking for queries
AOL Mature Teen	1.3%	19.6%	69.2%	4.3%
MSN Pornography	2.7%	13.3%	86.1%	8.2%
MSN Teen	2.6%	13.7%	83.1%	8.3%
ContentProtect Default	7.5%	12.4%	84.1%	23.1%
ContentProtect Custom	8.1%	7.8%	84.9%	25.3%
CyberPatrol Custom	3.9%	9.2%	86.4%	10.1%
CyberSitter Default	1.4%	19.9%	69.3%	5.1%
CyberSitter Custom	2.9%	18.2%	84.0%	9.4%
McAfee Young Teen	2.8%	32.8%	70.7%	9.3%
Net Nanny Level 2	12.6%	9.5%	82.9%	34.4%
Norton Default	9.9%	4.8%	79.4%	25.2%
Norton Custom	10.2%	2.9%	79.4%	25.9%
Verizon	4.4%	16.1%	67.9%	15.0%
8e6	3.4%	25.1%	93.0%	10.3%
SafeEyes	2.0%	16.5%	96.6%	6.4%

Summary of Filtering

- Most restrictive filter blocked 91% of adult pages; also blocked about 23-24% of the clean webpages in the indexes.
- Would block 22–23 clean webpages for each adult page it blocks in Google or MSN search index
- Less restrictive filters blocked as little as 40% of the adult pages.
- The most restrictive filter blocked about 94% of the adult pages among search results; also blocked about 13% of clean search results.
- On average, it would block about 7.6 clean results for every adult result it blocks.
- For the most popular queries, the most restrictive filter blocks over 98% of adult results; also blocked $\approx$20% of clean results.
- Would block $\approx$1.1 clean results of popular searches for each adult result it blocks.

Foreign Adult Websites with Commercial Ties to the US

Data Source	Percentage
Google index	90.3%
MSN index	89.8%
AOL, MSN & Yahoo! queries	88.2%
Wordtracker queries	95.9%

Estimated percentage of nominally free adult foreign webpages that have commercial ties to the United States, based on data provided by CRA International. Estimates for query results take into account query weights.

Filtering studies cited by Plaintiffs' Expert

Reference	Year	Sample type	Quantitative	Source of pages
eTesting Labs	2001	convenience	yes	searches on Google
eTesting Labs	2002	convenience	yes	searches on Google; DMOZ
NetAlert	2001	quota	yes	unknown
PC Magazine	2004	unknown	no	unknown
Consumer Reports	2005	convenience	no	unknown
Rulespace depo	2006	convenience	yes	unknown

eTesting 1: Google search for "free adult sex." eTesting 2: Added DMOZ; took sample of results. NetAlert: at most 30 webpages.

Not science.

Plaintiff's expert's testimony on filtering effectiveness

- Cites documents (e.g., COPA study, NRC report, expert declarations, product reviews) as saying things that plainly are not in the document.
- Claims documents say *the opposite* of what they actually say.

Example 1

Cites COPA study as saying filtering, monitoring and time-limiting technologies and parental supervision are effective alternatives to COPA.

Actual ratings on 10-point scale (not clear whether empirical):

Method	Effectiveness
Family education programs	5.2
Server-side filtering using URL lists	7.4
Client-side filtering using URL lists	6.5
Filtering using text-based content analysis	5.4
Monitoring and time-limiting technologies	5.5
Acceptable use policies/family contracts	4.6
Real time content monitoring/blocking	5.6

Example 2

Says NRC report concludes filters are highly effective.

Section cited says: "Today's filters cannot be the sole element of any approach to protecting children from inappropriate sexually explicit material on the Internet (or any other inappropriate material), and it is highly unlikely that tomorrow's filters will be able to serve this role either."

NRC also says easy to defeat many filters and that filters can "lead to a false sense of security."

And the report says of the primary technology used for content filtering, automatic text categorization, "The effectiveness of these methods is far from perfect—there is always a high error rate . . . not clear how directly [the finding that the method is sometimes nearly as accurate as a human rater] applies to, for example, pornography. . . . Substantially improved methods are not expected in the next 10 to 20 years."

Example 3

Cites NRC report to imply "clear [that] non-content filtering tools such as [software to limit access time] are very valuable and effective in helping parents control their children's Internet activities."

In fact, only mention of software on page cited is:
"If technology is used to limit access, consider the age-appropriateness of the limits you wish to impose."

The NRC Report suggests parental supervision can help, but warns:
Parents generally do not know what their children do on the Internet. Not feasible for parents to supervise children's activity on the Internet constantly. Supervising children's activity on the Internet competes with other parental responsibilities. Parents need training to supervise their children's online activity effectively.

Example 4

Cites a 2005 product review by the Consumers Union as saying that "all of the products tested [in 2005] were very good or excellent at blocking pornography."
The title of the review is "Filtering software: Better, but still fallible."

The review draws no quantitative conclusions about the effectiveness. Finds—using sample of convenience—that "[f]ilters keep most, but not all, porn out. . . . Informative sites are snubbed, too. The best porn blockers were heavy-handed against sites about health issues, sex education, civil rights, and politics. . . . These programs may impede older children doing research for school reports. Seven [of eleven products] block the entire results page of a Google or Yahoo! search if some links have objectionable words in them."

Example 5

Claims another expert's report says the CyberPatrol content filter had an error rate of 4.69–7.99% and that two other filters did nearly as well.

That report contains no such numbers: does not make *any* quantitative estimates of filter accuracy.

Plaintiffs' Geography Study

- Claim: less than half of "free" porn sites are in US, and about 2/3 of adult membership websites are in US
- Universe: Adultreviews.net, Adultwebmasters.org, Google Web Directory, Sextracker.com.
- Sample of convenience, not census or random sample.
- According to his database, the following are porn sites: aol.com, msn.com, yahoo.com, about.com lycos.fr, lycos.co.uk com.ar, com.au, com.br, co.hu, co.il, co.kr, com.mx, co.nz, com.pl, com.pt, com.tw, com.ua, co.uk, com.ve, co.yu, co.za
- Claims entire commercial domains of at least 17 countries are porn sites:
  Argentina, Australia, Brazil, Hungary, Israel, Korea, Mexico, New Zealand, Poland, Portugal, Spain, Taiwan, the Ukraine, the United Kingdom, Venezuela, Yugoslavia, and South Africa, respectively.

Not science. Judge took his results at face value nonetheless.