REBUTTAL EXPERT REPORT AND RULE 26(E) DISCLOSURE OF
PHILIP B. STARK, PH.D.

6 JULY 2006

ACLU v. Gonzales

Civ. Action No. 98-5591 (E.D. Pa.)

1

REBUTTAL EXPERT REPORT AND RULE 26(E) DISCLOSURE OF
PHILIP B. STARK, PH.D., 6 JULY 2006

I. INTRODUCTION

1. This report discusses the methods and data Matthew A. Zook and Lorrie Faith Cranor rely upon in their expert reports. It estimates the percentage of nominally free foreign sexually explicit webpages that have commercial ties to the United States to address some of Dr. Zook's claims. It estimates the effectiveness of an additional content filter[1] to address some of Dr. Cranor's claims. Finally, it revises estimates in my 8 May 2006 report to reflect corrections to the classification of twenty of the 68,150 webpages previously classified by CRA International.

2. To prepare this report, I read the expert reports of Drs. Lorrie Faith Cranor (7 May 2006), Edward W. Felten (8 May 2006), Donna L. Hoffman (5 May 2006) and Matthew A. Zook (4 May 2006). I examined computer files containing a portion of Dr. Zook's work: databases,[2] spreadsheets[3] and program files.[4] I also

---

1 The Verizon ISP-based filter.

2 "CONFIDENTIAL Zook-database.mdb" and "CONFIDENTIAL Zook Copa-database1.mdb." They have the same internal name, "copa-work." They contain tables and queries with the same names, but they differ. It is not clear which database Dr. Zook relied on.

3 "CONFIDENTIAL Zook whois data.xls" and "CONFIDENTIAL Zook table2.xls."

4 "CONFIDENTIAL_Zook1_get_google_member_list.pl,"
"CONFIDENTIAL_Zook1_get_google_nonmember_list.pl,"
"CONFIDENTIAL_Zook1_get_sextracker_list.pl," and
"CONFIDENTIAL_Zook1_get_whois_list.pl." I understand that Plaintiffs claim Dr. Zook relied on these programs to prepare his report; however, they have internal dates of 20 June 2006, about six weeks after Dr. Zook filed his expert report of 4 May 2006.

read some of the material cited in Dr. Cranor's report. The new and revised

estimates rely on a database provided by CRA International.

II. REPORT OF MATTHEW A. ZOOK

3. Dr. Zook claims that less than half of "free"[5] pornographic websites are located

in the United States, and that about two thirds of adult membership websites are

located in the United States. (Zook, 4 May 2006, at 1, 12−18) These claims are

based on five pre-existing lists of webpages, lists he found on the Internet.[6] (Zook,

4 May 2006, at 6–8) He acknowledges that the lists are not a full census of

pornography on the Internet, nor are they scientific samples.[7]

4. There is no reason to think that the lists are representative of any larger group of

webpages, that they cover "free" and membership websites equally thoroughly, that

---

5   What Dr. Zook means by "free" is discussed in Appendix I, paragraph I.6.

6   The lists are from Adultreviews.net, Adultwebmasters.org, the Google Web Directory
    and Sextracker.com. He found them by searching for existing lists of pornography
    using Google. (Zook, 4 May 2006, at 3, 8)

7   "[I]ndices and directories of adult webpages that are readily identifiable and available
    on the Internet to an English speaking user are utilized." (Zook, 4 May 2006, at 6)

    "[T]his database (*i.e.*, the five combined datasets) represents a sample of adult oriented
    websites on the Internet rather than a complete census. This sample is based on indices
    that are readily available to an Internet user conducting Google searches in English for
    adult materials." (Zook, 4 May 2006, at 8)

    Dr. Zook took random samples from the Google lists, although he does not describe
    how. (Zook, 4 May 2006, at 7) A random sample from a sample of convenience is
    just a smaller sample of convenience.

they cover domestic and foreign websites equally thoroughly, or that their coverage has remained even over time. They comprise what is known in Statistics as a "sample of convenience."

5. Samples of convenience tend to be unrepresentative (biased), while random samples tend to be representative (unbiased). It is generally impossible to quantify the error in extrapolating from samples of convenience.[8] Dr. Zook's sample is no exception.[9]

---

8 "[A]bout the only way of examining how good [a sample of convenience] may be is to find a situation in which the results are known, either for the whole population or for a [random] sample, and make comparisons. Even if a method appears to do well in one such comparison, this does not guarantee that it will do well under different circumstances." (Cochran, W.G., 2002. *Sampling Techniques, Third Edition*, John Wiley and Sons, New York, at 10)

"The use of [random] sampling techniques maximizes both the representativeness of the survey results and the ability to assess the accuracy of estimates obtained from the survey. … [Random] sampling offers two important advantages over other types of sampling. First, the sample can provide an unbiased estimate of the responses of all persons in the population from which the sample was drawn … Second, the researcher can calculate a confidence interval that describes explicitly how reliable the sample estimate of the population is likely to be. … [Q]uantitative values computed from [samples of convenience] … should be viewed as rough indicators rather than as precise quantitative estimates. Confidence intervals should not be computed." (Diamond, S.S., 2000. *Reference Guide on Survey Research*, in *Federal Judicial Center Reference Manual on Scientific Evidence, 2nd edition,* Federal Judicial Center, at 242–244)

"[Samples of convenience have] a heavy dependence on the validity of broad assumptions about the distributions of the survey variables in the population. On the contrary, from the results of ideal [random] sampling, the inferences to the population can be made entirely by statistical methods, without assumptions regarding the population distributions." (Kish, L., 1965. *Survey Sampling*, John Wiley and Sons, New York, at 19)

9 He acknowledges that the reliability of estimates of the prevalence of pornographic material on the Internet is "difficult to determine." (Zook, 4 May 2006, at 5)

6. The study that the government commissioned used random sampling. As a result,[10] one can quantify its uncertainty, as I did using confidence intervals. (Stark, 8 May 2006, at 14–20; Appendix II below)

7. In order to check Dr. Zook's work thoroughly, I would need the five lists he started with and the computer code he used to process the lists.[11] The government requested these from the Plaintiff. The five lists were not produced,[12] and only some of his computer code was produced. Thus, it is impossible to check his programming completely or to reproduce his work.

8. However, files that were produced show that Dr. Zook's analysis went awry: for example, his databases[13] list aol.com, msn.com, yahoo.com and about.com as free pornographic websites. And they list lycos.fr, and lycos.co.uk, the Lycos websites in France and the United Kingdom, as pornographic websites.

9. Many commercial websites end in .com (e.g., www.google.com). Those websites comprise the ".com domain." There are also country-specific .com

---

10 See footnote 8.
11 Processing included extracting addresses from the lists, truncating the addresses (see paragraph 10), looking up the registrants, identifying the host countries, importing the data into a database, and forming tables that combined the processed lists. Only the final output of the processing and some of the computer instructions were produced.
12 It is impossible to reconstruct the five lists from the files that were produced.
13 See footnote 2.

domains. For example, many commercial websites registered in the United

Kingdom end in .co.uk (e.g., www.google.co.uk), and many commercial websites

registered in Argentina end in .com.ar (e.g., www.google.com.ar). Dr. Zook's

databases[14] list 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 and co.za as pornographic websites.

These are country-specific ".com" domains for Argentina, Australia, Brazil,

Hungary, Israel, Korea, Mexico, New Zealand, Poland, Portugal, Spain, Taiwan,

the Ukraine, the United Kingdom, Venezuela, Yugoslavia, and South Africa,

respectively.[15]

10. I will explain how Dr. Zook came to mislabel these websites and domains. He

truncated the addresses of the webpages[16] in the five lists to form what he calls

"websites."[17] (Zook, 4 May 2006, at 6) He shortened the addresses too far. For

example, suppose his programs[18] found the address

---

14 See footnote 2.

15 http://www.com.es is a Spanish website that contains Internet-related information, but
    .com.es is a commercial domain in Spain. Similarly, http://co.za is the website for the
    agency in South Africa that administers the co.za domain. Many non-profit
    organizations have websites ending in .org. There are also country-specific ".org"
    domains; for instance, .org.uk is the .org domain in the United Kingdom. Dr. Zook's
    databases (footnote 2) list org.uk as a free pornographic website.

16 "These webpages, e.g., http://domain.com/page1.html or
    http://www.domain.com/page2.html, were aggregated to the level of websites, e.g.,
    domain.com." (Zook, 4 May 2006, at 6)

17 When "website" is distinguished from "webpage," it usually means a collection of
    webpages whose content is controlled or maintained by a single entity, such as an
    institution, business or individual. (e.g., Zook, 4 May 2006, at 3) However, many
    entries in Dr. Zook's databases (footnote 2) are not, by that definition, websites.

18 The first three programs mentioned in footnote 4 collect and truncate addresses.

http://communities.msn.com/GayJudo/[19] on one of the lists. His programs would shorten it to msn.com, and thus list msn.com as a pornographic website. Similarly, his programs would shorten the address http://www.porn.com.br/topless-beach-pics/[20] to com.br, and thus list the ".com" domain of Brazil as a pornographic website.[21]

11. In summary, Dr. Zook's estimates are not reliable. He used a sample of convenience, so his results could not be extrapolated reliably beyond the five lists he started with even if he had analyzed them correctly. He made programming errors that led to misclassifying aol.com, msn.com, yahoo.com and the ".com" domains of at least seventeen countries as pornographic websites. He used methods biased in favor of his conclusions.[22] He did not produce his raw data and some of his programming, so his results cannot be checked completely or reproduced. And his conclusions are contradicted by those of the only study I

---

19 This address was in the Google Adult Directory on 25 June 2006. There is no way to tell from the files that were produced whether Dr. Zook processed this address. However, I am confident that something similar led Dr. Zook to conclude that msn.com is a pornographic website.
20 This address was in the Google Adult Directory on 2 July 2006.
21 Incidentally, on 25 June 2006, http://communities.msn.com/GayJudo/ automatically redirected the browser to the page http://groups.msn.com/GayJudo/_homepage.msnw?pgmarket=en-us, which did not contain pornography. And on 2 July 2006, http://www.porn.com.br/topless-beach-pics/ did not contain pornography, although it had links to other websites that might have. Dr. Zook does not report checking whether the webpages on the five lists worked or contained pornography. See Appendix I.
22 See Appendix I.

know of that has used a proper (random) statistical sample, namely, the study that the government commissioned. Appendix I discusses some other issues with his analysis.

III. REPORT OF LORRIE FAITH CRANOR

12. Dr. Cranor claims that content filters and parental supervision are effective alternatives to COPA. She cites reports, magazine articles and depositions, but apparently did not collect or analyze any data.

13. Dr. Cranor does not cite any empirical study of the effectiveness of parental supervision. She does cite empirical studies of filtering. She does not discuss their methods, but acknowledges that "test results vary and some filter evaluation reports do not completely document their methodology." (Cranor, 7 May 2006, at 14) I tried to determine the type of sample and the sources of the webpages used to test the filters in evaluations she cites. Table 1 shows the results.

Reference	Year	Sample type	Quantitative	Source of test webpages
1. eTesting Labs, 2001	2001	convenience[23]	yes	searches on Google
2. eTesting Labs, 2002	2002	convenience[24]	yes	searches on Google; DMOZ
3. NetAlert	2001	quota sample[25]	yes	unknown
4. PC Magazine	2004	unknown	no	unknown
5. Consumer Reports	2005	convenience[26]	no	unknown
6. Rulespace deposition	2006	convenience	yes[27]	unknown[28]

Table 1: References on the empirical effectiveness of filters cited by L.F. Cranor, 7 May 2006. (1) eTesting Labs, October 2001. U.S. Department of Justice: Web Content Filtering Software Comparison; (2) eTesting Labs, March 2002. Corporate Content Filtering Performance and Effectiveness Testing; (3) Greenfield, P., Rickwood, P, and Tran, H.C., September 2001. Effectiveness of Internet Filtering Software Products, CSIRO Mathematical and Information Sciences; (4) Munro, J., August 2004. Cybersitter 9.0 review, PC Magazine, http://www.pcmag.com/article2/0,1759,1618830,00.asp; (5) Consumers Union of the U.S., Inc., 2005. Filtering software: Better, but still fallible. http://www.consumerreports.org/cro/electronics-computers/internet-filtering-software-605/overview/index.htm; (6) Deposition of Alistair R. Allan, 28 February 2006. The fourth column indicates whether the reference gives a quantitative estimate.

---

23 eTesting took a random sample from a sample of convenience, which yields a smaller sample of convenience. The sample of adult content was collected by searching for "free adult sex" using Google. The sample of webpages with no adult content was collected in a similar manner. (eTesting Labs, October 2001, at 2, 6−7)

24 eTesting took a random sample from samples of convenience assembled from Google search results and the DMOZ directory. (eTesting Labs, March 2002, at 2, 3, 15−17) DMOZ (www.dmoz.org) is a volunteer-edited directory of websites called the Open Directory Project.

25 Quota samples are non-random samples (see footnote 8) with limitations similar to those of samples of convenience. In particular, quota samples tend to be unrepresentative, and generally it is not possible to assess the uncertainty in extrapolating from a quota sample. Greenfield et al. mention the word "random," but do not describe drawing a random sample from a larger collection of webpages; I conclude that by "random" they meant "haphazard" rather than "drawn using probability methods." In all, they used no more than 30 pornographic webpages. (Greenfield et al., 2001, at 23, 25–26)

26 "[W]e built a list of objectionable sites that anyone can easily find, plus informational sites ..." (Consumers Union, 2005)

27 The meaning of the number is obscure. See paragraph 23 and footnote 36.

28 The pages are "found" by the Rulespace content raters. (Allan Deposition, 28 February 2006, at 113, 115)

14. It does not appear that any of the empirical studies of filtering Dr. Cranor cites used random samples. Most used samples of convenience, discussed above in paragraph 5 and footnote 8. Statisticians would generally agree that the methodology used in the study that the government commissioned—random sampling—is more reliable. Moreover, the independent estimates Dr. Cranor cites are over four years old, a long time by Internet standards. In contrast, the data for the study the government commissioned were collected and analyzed within the last year.

15. Many of Dr. Cranor's citations do not say what she says they say. Paragraphs 16–22 give some examples.

16. Dr. Cranor says the COPA Report[29] supports her assertion that filtering, monitoring and time-limiting technologies and parental supervision are effective alternatives to COPA. (Cranor, 7 May 2006, at 15–16, 25) The COPA Report rates effectiveness on a 10-point scale.[30] Table 2 summarizes the ratings. They are not high.

---

29 Telage, D., et al., 2000. Commission on Child Online Protection (COPA) Report to Congress, 20 October. I refer to this as the COPA Report below.
30 The ratings seem to summarize the opinions of the Commission members rather than objective tests of the approaches.

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

Table 2: COPA Report ratings of the effectiveness of various methods for protecting children on the Internet, measured on a 10-point scale. (COPA Report, at 18, 19, 21, 22, 34, 36, 38)

17. Dr. Cranor attributes to the NRC Report[31] the conclusion that content filters are highly effective. (Cranor, 7 May 2006, at 16, citing NRC Report at 302–303) But the section she cites 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 Report, at 301–302) The NRC Report also points out that it is easy to defeat many filters and that filters can "lead to a false sense of security." (NRC Report, at 280–281, 302) 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 ... [I]t is not clear how directly [the finding that the method

31 Thornburgh, D., and Lin, H.S., eds., 2002. *Youth, Pornography, and the Internet.* National Academy Press, Washington, D.C., 450pp. I refer to this as the NRC Report. Dr. Cranor writes that the NRC Report was published in 2005, but it was published in 2002.

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." (NRC Report, at 420–421)

18. Dr. Cranor says page 376 of NRC Report makes it "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." (Cranor, 7 May 2006,

at 25) The only mention of software on that page is, "If technology is used to limit

access, consider the age-appropriateness of the limits you wish to impose."[32]

19. The NRC Report suggests that parental supervision can help, but it warns that:

- Parents generally do not know what their children do on the Internet. (NRC Report, at 164–165)
- It is not feasible for parents to supervise children's activity on the Internet constantly. (NRC Report, at 223)
- Supervising children's activity on the Internet competes with other parental responsibilities. (NRC Report, at 370)
- Parents need training to supervise their children's online activity effectively. (NRC Report, at 226, 228, 231–232, 257, 378)

Because apparently innocuous searches can return sexually explicit materials and

website names do not always reflect website contents, it is implausible that parents

could anticipate reliably whether a link or search leads to sexually explicit

materials.

---

32 Most of the page shows the large effort required to follow "best practices" for Internet use, which may include setting time limits. (NRC Report, 2002, at 376)

20. Dr. Cranor says that pages 18 and 36 of the COPA Report make it "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." (Cranor, 7 May 2006, at 25) But neither page 18 nor page 36 of the COPA Report mentions limiting children's access time.[33]

21. Dr. Cranor 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.'" (Cranor, 7 May 2006, at 17) The title of the review is "Filtering software: Better, but still fallible."[34] The review draws no quantitative conclusions about the effectiveness of filters. It finds—based on a 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." (Consumers Union, 2005)

33 Page 18 of the COPA Report does not mention software at all: it rates the effectiveness of "family education programs" at 5.2 points out of 10. (COPA Report, 2002, at 18) Page 36 rates the effectiveness of "acceptable use policies and family contracts" "regarding the types of materials that may be accessed" at 4.6 points out of 10. The only reference to software is implicit: "[acceptable use] policies may or may not be accompanied by monitoring." (COPA Report, at 36)
34 Dr. Cranor changed the word "software" to "products" in the title of the article. (Cranor, 7 May 2006, at 17)

22. Dr. Cranor claims that the Expert Report of Corey Finnell (30 November 2001) says the CyberPatrol content filter had an error rate of 4.69 percent to 7.99 percent and that two other filters did nearly as well. (Cranor, 7 May 2006, at 15) The report contains no such numbers. It does not make any quantitative estimates of filter accuracy.

23. Dr. Cranor relies on the filter vendors' claims that filters work well. For example, she cites a Rulespace claim that their English-language filter is 99.52 percent accurate.[35] (Cranor, 7 May 2006, at 18, citing "Rulespace Depo." [Allan Deposition, 28 February 2006], at 113, 178) It is clear that the 99.52 percent figure is based on a sample of convenience. (Allan Deposition, 28 February 2006, at 113–115) But even if the sample had been random, the 99.52 percent figure alone says nothing about how well the filter blocks pornography,[36] and the deposition has little clarifying information. And the NRC Report warns, "Filter vendors

---

35 She also cites a Rulespace filter accuracy claim of 99.48 percent. (Cranor, 7 May 2006, at 12, citing "Rulespace Depo." [Allan Deposition, 28 February 2006], at 109–110) That number refers to a product for filtering wireless (e.g., mobile phone) content, not to the general Internet. It too is based on a sample of convenience.

36 An example might help. Suppose that a filter is tested on 10,000 pages of which 48 contain pornography, and that the filter does nothing at all—it lets every page through. Then the filter does the right thing 9,952 times out of 10,000: its accuracy on the test set is $9{,}952/10{,}000 = 99.52$ percent. But it misses 100 percent of the pornography.

The NRC Report discusses filter accuracy at length. (NRC Report at 60–61, 275, 277–283, 303–304) Unlike the Allan Deposition, the NRC Report and the study the government commissioned keep overblocking and underblocking rates separate, which makes accuracy figures easier to interpret.

sometimes provide estimates of overblock and underblock rates, but without

knowing the methodology underlying these estimates, the cautious user must be

concerned that the methodology is selected to minimize these rates." (NRC Report,

at 277) The study that the government commissioned tested filters based on

Rulespace technology using random samples of sexually explicit and clean

websites. The measured rates of overblocking and underblocking are reported in

Appendix II, tables 4, 7 and 9.

24. Dr. Cranor's argument has logical gaps. For example, she claims that because

users find content through search engines and filter companies use search engines,

filter vendors find "the large majority of sites with inappropriate images that users

might actually see." (Cranor, 7 May 2006, at 12) Her argument is logically no

different from "geologists use hammers and carpenters use hammers; therefore,

geologists and carpenters do largely the same thing." Search results depend on

what is searched for. A strength of the study the government commissioned is that

it tested filters on the results of real searches.[37]

25. In summary, Dr. Cranor offers no empirical evidence to support her claim that

parental supervision is an effective alternative to COPA. Many of her citations do

not contain what she says they contain. Some material she cites for support

---

37 It seems that Symantec, at least, uses its own search terms. (Trollope Deposition, 2
    March 2006, at 19–21)

actually contradicts her position. Her quantitative claims about the accuracy of content filtering rely on tests that are out of date and on an uninterpretable number given by a filter vendor. The number cannot be interpreted because her source omits crucial information. Statisticians would generally agree that the methodology used in the government-commissioned study is more reliable than that used in the filtering studies she cites.

IV. "FREE" FOREIGN SEXUALLY EXPLICIT WEBPAGES

26. CRA International determined whether the nominally free foreign sexually explicit webpages they categorized[38] have commercial ties to the United States.[39] From their data, I estimate that roughly 90 percent do. Table 3 gives more detail.

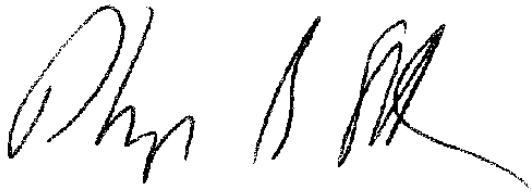| Data Source	Estimated percentage of "free" sexually explicit webpages with commercial ties to the U.S.
Google index	90.3%
MSN index	89.8%
AOL, MSN and Yahoo! queries	88.2%
Wordtracker queries	95.9%

Table 3: Estimated percentage of nominally free sexually explicit foreign webpages that have commercial ties to the United States, based on data provided by CRA International. The 6 July 2006 Rebuttal Report of Paul Mewett will explain how commercial ties were ascertained. Data sources are described in my expert report of 8 May 2006. Estimates for query results take into account query weights, as described in Appendix II of that report.

---

38 These are webpages hosted outside the United States and that are in category 5f (adult entertainment) but not in category 4b or 4c (subscription or sales). The categories are explained in the 8 May 2006 Expert Report of Paul Mewett.

39 What CRA International considered to be a commercial tie to the United States will be described more fully in the 6 July 2006 Rebuttal Report of Paul Mewett.

V. QUALITY CONTROL

27. I understand that CRA International discovered that, among the 68,150 webpages they had classified, 61 were misclassified. Of those, twenty were misclassified in a way that affected my analysis.[40] I repeated the analysis in my 8 May 2006 report using a corrected database supplied by CRA International. The results are in Appendix II.[41] Changes to the values reported in my 8 May 2006 report are minor[42] and do not affect my qualitative conclusions.

_____ Dated __6__ July 2006.

Philip B. Stark

---

40 Four webpages moved out of category 1a, no sexual content, while two moved in. Five webpages moved out of category 5f, adult entertainment, while nine moved in.

41 These results are submitted under Federal Rule of Civil Procedure 26(e).

42 Most of the changes were 0.2 percent or less. The largest changes were to the percentages of domestic webpages among sexually explicit query results that filters did not block.

APPENDIX I: SOME OTHER ISSUES IN DR. ZOOK'S ANALYSIS

I.1. It is individual webpages, rather than websites, that browsers display, that search engines retrieve, and that content filters block.[43] For the purposes of measuring the amount of sexually explicit material on the Internet or the effectiveness of filters, webpages are a more appropriate unit of analysis than websites. A website with a single page containing a few images should get less weight than, for example, playboy.com,[44] which has far more content that is explicit. But truncating the addresses as Dr. Zook does[45] to form "websites" would give them equal weight. To the extent that larger pornographic websites are in the United States, truncating addresses biases Dr. Zook's estimate of the percentage of pornographic material that is foreign in favor of his conclusion.

I.2. There is nothing to suggest that the lists Dr. Zook used are kept up-to-date by removing webpages that no longer work and re-categorizing webpages whose content has changed.[46] He does not report checking whether the webpages on the lists were working at all, nor whether they contained sexually explicit material.[47]

---

43 Some filters "blacklist" entire websites. That means that the filter blocks access to every webpage in the website. Some filters sometimes block portions of a webpage.
44 Playboy.com is one of the websites in Dr. Zook's databases (footnote 2).
45 See paragraph 10.
46 Indeed, some of the websites Dr. Zook relied on in 2003 for lists of pornographic websites are no longer available. (Zook, 4 May 2006, at 6)
47 See paragraph 10 and footnotes 19, 20 and 21.

Obsolete webpages on the lists could bias his results.

I.3. Dr. Zook assumed that every webpage listed in Adultreviews.net belongs to a membership pornographic website and that the classification of pornographic webpages in the Google Adult Directory as "free" or requiring paid membership is perfectly accurate. (Zook, 4 May 2006, at 6–7) He does not report checking the accuracy of that assumption. Misclassifications could bias his results.

I.4. Domains registered to foreign owners are often hosted on servers in the United States.[48] Dr. Zook considered a website to be foreign if the postal address of the registrant was foreign. (Zook, 4 May 2006, at 10)   That inflates his estimate of the percentage of pornographic websites that are foreign—a bias in favor of his conclusion.[49]

I.5. Plaintiffs' other experts confirm that users generally or increasingly find webpages through searches. (Cranor, 7 May 2006, at 7; Felten, 8 May 2006, at 26–28; Hoffman, 5 May 2006, at 4) The study commissioned by the government

---

48 Of the sexually explicit webpages (category 5f) CRA International examined that were registered to foreign addresses, 56 percent were hosted in the United States. (Paul Mewett, personal communication, 16 June 2006)

49 He was unable to find the registrants of some "websites" in his databases because they are not actually websites. Examples include co.il, co.kr, co.nz, co.uk, co.yu, co.za, com.ar, com.au, com.br, com.es, com.pt, com.tw and com.ve. See paragraphs 9 and 10.

shows that AOL, MSN, Yahoo! and Wordtracker searches retrieve domestic

sexually explicit material preferentially. (Stark, 8 May 2006, at 9–11; Appendix II

below) Thus, even if Dr. Zook had started with a representative sample of

pornographic webpages, his method of analysis would tend to overestimate the

percentage of pornographic material users encounter that is foreign.

I.6. Dr. Zook's use of the terms "free" and "membership" does not match my

understanding of the COPA distinction between non-commercial and commercial

websites.[50] For example, it is my understanding that a website that does not require

paid membership but that collects advertising revenue is commercial according to

COPA. Thus, Dr. Zook's analysis does not address the right question.[51]

---

50 Dr. Zook writes, "The majority of [adult-oriented] websites are commercially driven."
    (Zook, 4 May 2006, at 3). But, "free adult websites comprise the largest number of
    adult websites on the Internet." (Zook, 4 May 2006, at 4) Unless he is contradicting
    himself, what Dr. Zook means by "free" includes some commercial sites.
51 Dr. Zook assumed that the coverage of "free" and membership adult websites in the
    Google lists is comparable simply because the lists come from Google: "Due to the
    fact that Google is a source for data on both free and membership websites, it is
    possible to compare these two groups directly." (Zook, 4 May 2006, at 17) He
    assumed that Sextracker's coverage of adult websites in 2001 and 2006 is comparable.
    (Zook, 4 May 2006, at 16) And he assumed that coverage of "free" and membership
    adult websites in the Google catalogs in 2001 and 2006 is comparable. (Zook, 4 May
    2006, at 14, 15, 17) He does not report testing any of these assumptions, none of
    which is plausible.

20

APPENDIX II: ESTIMATES

II.1 This Appendix revises estimates given in my 8 May 2006 report to reflect

corrections to the data by CRA International, and estimates the effectiveness of an

additional content filter. It also estimates the percentage of nominally free sexually

explicit foreign webpages that have commercial ties to the United States.

II.2 After corrections to the data, the estimated percentages of sexually explicit

webpages in the Google and MSN indexes both remain 1.1 percent. The estimated

percentage of sexually explicit webpages in the Google index that are domestic

remains 44.2 percent, and the estimated percentage of sexually explicit webpages

in the MSN index that are domestic increases from 56.6 percent to 56.7 percent.

II.3 The estimated percentage of searches that retrieve at least one sexually explicit

webpage remains 6 percent after the corrections to the data, and the estimated

percentage that retrieve at least one domestic sexually explicit webpage remains

5.7 percent. The estimated percentage of search results that are sexually explicit

remains 1.7 percent. The estimated percentage of sexually explicit search results

that are domestic increases from 87 percent to 88.4 percent.

II.4 The lower confidence bounds in Table 1 of my 8 May 2006 report are

unchanged by the corrections to the data.

II.5 The percentage of Wordtracker queries that retrieve at least one sexually

explicit webpage decreases from 37.3 percent to 37.1 percent after corrections to

the data, and the percentage of Wordtracker queries that retrieve at least one

domestic sexually explicit webpage decreases from 37.2 percent to 37.0 percent.

The percentage of Wordtracker query results that are sexually explicit increases

from 13.9 percent to 14.1 percent. The percentage of sexually explicit Wordtracker

search results that are domestic remains 87.4 percent.

II.6 Table 4 gives estimates of underblocking and overblocking and Table 5 gives

lower confidence limits, for the Google and MSN indexes. Table 6 gives estimates

of the percentage of domestic sexually explicit webpages among the sexually

explicit webpages in the Google and MSN indexes that filters do not block. Table

7 gives estimates of overblocking and underblocking for webpages retrieved by

AOL, MSN and Yahoo! queries. Table 8 gives lower confidence limits for the

percentage of AOL, MSN and Yahoo! queries that return at least one sexually

explicit webpage that would not be blocked by filters. Table 9 shows

underblocking and overblocking for Wordtracker queries. Differences between

these tables and the corresponding tables in my 8 May 2006 report are small.

Filter[52]	Underblocking		Overblocking	
	Google	MSN	Google	MSN
1	8.9%	*8.6%	22.6%	23.6%
2a	16.8%	*18.7%	19.6%	10.3%
2b	17.7%	20.5%	21.9%	18.9%
3a	38.3%	*45.4%	2.8%	3.0%
3b	28.3%	*46.7%	1.4%	0.7%
4	31.0%	*33.5%	1.4%	0.9%
5a	12.7%	*16.5%	3.6%	4.1%
5b	12.4%	*18.9%	4.0%	3.7%
6	16.1%	*26.0%	12.4%	13.2%
7	44.0%	*46.1%	3.3%	2.2%
8a	60.2%	*54.9%	1.4%	0.7%
8b	58.4%	*54.2%	0.9%	0.4%
9	41.8%	40.3%	9.4%	5.7%

Table 4: Estimated underblocking and overblocking of webpages in the Google and MSN indexes. Among sexually explicit webpages, the percentage that are not blocked by a filter is the rate of underblocking. Among clean webpages, the percentage that are blocked by a filter is the rate of overblocking. The filter settings and testing protocol are explained in the 8 May 2006 Expert Report of Paul Mewett. Asterisks signify values that differ from those reported in Table 2 of my 8 May 2006 report.

---

52 The filters are as follows; settings are described more fully in the 8 May 2006 Expert Report of Paul Mewett and the 6 July 2006 Rebuttal Report of Paul Mewett. 1: AOL Mature Teen. 2a: MSN Pornography. 2b: MSN Teen. 3a: ContentProtect Default setting. 3b: ContentProtect Custom setting. 4: CyberPatrol Custom setting. 5a: CyberSitter Default setting. 5b: CyberSitter Custom setting. 6: McAfee Young Teen. 7: Net Nanny Level 2. 8a: Norton Default setting. 8b: Norton Custom setting. 9: Verizon.

Filter	Underblocking		Overblocking	
	Google	MSN	Google	MSN
1	*5.6%	*6.5%	18.4%	21.0%
2a	*12.1%	*15.7%	15.8%	8.5%
2b	12.8%	*17.4%	17.8%	16.6%
3a	31.3%	*41.3%	1.5%	2.1%
3b	22.2%	*42.6%	0.6%	0.4%
4	24.6%	*29.7%	0.6%	0.5%
5a	8.6%	*13.6%	2.1%	3.1%
5b	8.4%	*15.9%	2.4%	2.7%
6	11.4%	*22.5%	9.3%	11.3%
7	36.8%	*41.9%	1.9%	1.5%
8a	52.9%	*50.7%	0.6%	0.4%
8b	51.1%	*50.1%	0.4%	0.2%
9	34.7%	36.2%	6.7%	4.4%

Table 5: 95% lower confidence limits for the entries in Table 4. For illustration, at 95% confidence, filter 2b fails to block at least 12.8% of the sexually explicit webpages in the Google index. Similarly, at 95% confidence, filter 2b blocks at least 16.6% of the clean webpages in the MSN index. Asterisks signify values that differ from those in Table 3 of my 8 May 2006 report.

Filter	Estimated Domestic Underblocking	
	Google	MSN
1	40.0%	*40.6%
2a	31.6%	*42.9%
2b	40.0%	37.7%
3a	39.0%	45.8%
3b	40.6%	47.1%
4	48.6%	*44.0%
5a	50.0%	*32.8%
5b	57.1%	*36.2%
6	44.4%	*37.5%
7	41.7%	*48.1%
8a	35.3%	*49.3%
8b	36.4%	*49.7%
9	37.0%	42.4%

Table 6: Of the sexually explicit webpages in the Google and MSN indexes that filters do not block, the estimated percentage that are domestic webpages. Asterisks signify values that differ from those in Table 4 of my 8 May 2006 report.

Filter	Underblocking for results	Overblocking for results	Domestic Underblocking	Underblocking for queries
1	6.2%	12.5%	57.0%	15.6%
2a	21.4%	4.4%	86.1%	32.3%
2b	*20.8%	5.8%	91.9%	28.1%
3a	18.4%	6.4%	70.1%	46.2%
3b	20.4%	0.0%	62.1%	42.2%
4	34.6%	0.4%	*94.9%	65.6%
5a	11.2%	4.6%	33.8%	23.2%
5b	10.0%	5.3%	44.1%	20.1%
6	14.2%	20.7%	80.7%	30.9%
7	28.1%	3.7%	*79.4%	36.6%
8a	42.1%	0.8%	*85.3%	51.6%
8b	43.4%	0.0%	*85.6%	56.1%
9	23.1%	1.3%	80.9%	41.6%

Table 7: Estimated underblocking and overblocking of the results of AOL, MSN and Yahoo! searches. "Underblocking for results" is the percentage of sexually explicit search results that are not blocked. "Overblocking for results" is the percentage of clean search results that are blocked. "Domestic underblocking" is the percentage of domestic webpages among sexually explicit webpages the filters do not block. "Underblocking for queries" is, among queries that retrieve any sexually explicit webpages, the percentage that retrieve at least one sexually explicit webpage that is not blocked. Asterisks signify values that differ from those in Table 5 of my 8 May 2006 report.

Filter	Underblocking for queries
1	5.3%
2a	20.9%
2b	18.8%
3a	10.0%
3b	25.4%
4	24.4%
5a	11.2%
5b	8.1%
6	10.4%
7	20.8%
8a	49.3%
8b	54.3%
9	31.4%

Table 8: Lower 95% confidence limits for the rightmost column in Table 7. Corrections to the data by CRA International did not change any values in Table 6 of my 8 May 2006 report.

Filter	Underblocking for results	Overblocking for results	Domestic Underblocking	Underblocking for queries
1	*1.3%	*19.6%	*69.2%	*4.3%
2a	2.7%	13.3%	*86.1%	*8.2%
2b	2.6%	*13.7%	*83.1%	*8.3%
3a	*7.5%	*12.4%	*84.1%	*23.1%
3b	*8.1%	7.8%	*84.9%	*25.3%
4	*3.9%	9.2%	*86.4%	*10.1%
5a	*1.4%	*19.9%	*69.3%	*5.1%
5b	*2.9%	*18.2%	*84.0%	*9.4%
6	*2.8%	32.8%	*70.7%	*9.3%
7	*12.6%	*9.5%	*82.9%	*34.4%
8a	*9.9%	*4.8%	*79.4%	*25.2%
8b	*10.2%	2.9%	*79.4%	*25.9%
9	4.4%	16.1%	67.9%	15.0%

Table 9: Underblocking and estimated overblocking for the results of Wordtracker queries. "Underblocking for results" is the percentage of sexually explicit search results that are not blocked. "Overblocking for results" is the percentage of clean search results the filter blocks. "Domestic underblocking" is the percentage of domestic webpages among the sexually explicit webpages the filters do not block. "Underblocking for queries" is, among the queries that retrieve any sexually explicit webpages, the percentage that retrieve at least one sexually explicit webpage that is not blocked. Overblocking was estimated from a random sample of clean search results. Underblocking was determined from all the sexually explicit search results. Asterisks signify values that differ from those in Table 7 of my 8 May 2006 report.