APPENDIX A: METHODOLOGY

I. SAMPLE SELECTION

The Survey was based on two target populations – a random sample of all Web sites with at least 39,000 monthly visitors (the "Random Sample") and the 100 most popular U.S. commercial sites (the "Most Popular Group"). Both were drawn from the January 2000 Nielsen//NetRatings data, described below. In the case of the Random Sample, the full data set was used as a "sampling frame," from which a "sampling pool" was created using a systematic sampling procedure. The Random Sample was drawn from this "sampling pool" as described below. In the case of the Most Popular Group, the Survey involved data collection for the top 100 sites, subject to the exclusions discussed below, and thus did not involve a sample.

A. CREATION OF SAMPLING FRAME

Nielsen//NetRatings provided the data from which the Random Sample and the Most Popular Group were drawn.[1] It provided a list of all ".com" domains with a projected audience of at least 39,000 unique visitors in January 2000.[2] The domains were ranked according to traffic, and each domain's unique audience was provided. A "domain" is the aggregation of all Web pages, sites, and servers using a particular domain name, defined as the word or letters immediately preceding the ".com." For example, the site "sports.yahoo.com," the page "yahoo.com/news/root.asp," and the server "www1.yahoo.com" would all be included in the domain "yahoo.com." In this study, we define "Web site" as a domain, which served as the unit of analysis for the Survey.[3]

"Unique audience" or "unique visitors" is defined as the estimate of the number of different individuals that visited a domain in a particular time period, without regard to the number of visits made to or the amount of time spent at the domain by each individual during that time period. Thus, even if a person visited a particular domain thirty times in the month of January 2000, he or she would still count as one visitor to the domain.

Nielsen//NetRatings identified all ".com" domains visited by individuals from home during the month of January 2000 and calculated the projected unique audience for each domain using

statistical methods.[4] All ".com" domains with at least 39,000 unique visitors were selected and ranked in order of audience size.[5] This list served as the sampling frame for the Random Sample. Accordingly, results from the Survey of the Random Sample can only be generalized to this population of Web sites, and not to the entire universe of ".com" domains.[6] The busiest 100 sites on the Nielsen//NetRatings list (excluding certain sites, as discussed below) constituted the Most Popular Group.

B. CREATION OF SAMPLING POOL

The following systematic sampling procedure was used to create a pool of sites from the sampling frame provided by Nielsen//NetRatings.[7] First, a target size of 350 sites was established for the Random Sample. It was estimated that up to 800 sites might need to be examined to ensure a final sample size of about 350.[8] Once this target sampling-pool size was determined, a "sampling interval" was determined by dividing 5,672 (the number of sites on the Nielsen list) by 800 (the target sampling pool size) to get an interval of 7 (rounded). The sampling interval was then used to randomly select sites from the sampling frame for inclusion in the sampling pool by the following methodology. A random number was generated, and the site appearing in the random number's slot on the sampling frame list was selected for inclusion, as was each site appearing on the list at the interval of one sampling interval. The resulting sampling pool contained 811 sites.

The 811 sites were then divided into 54 replicates of 15 sites each (with one replicate having 16 sites). Dividing 811 by 54 yielded the replicate interval of 15 (rounded), which was used to apportion sites among replicates. The first site went to the first replicate, the second to the second replicate, etc. Thus the 54th site was allocated to the 54th replicate. The process was then continued with the 55th site going to the first replicate, etc., until all sites had been allocated. This allocation ensured that the final sample would be representative of the sampling frame regardless of the number of replicates used. Note that because the replicates were created from the entire sampling frame, some sites from the Most Popular Group also appeared on

the replicates and thus were included in the Random Sample.

A similar procedure was used to create replicates for the 100 sites in the sampling pool for the Most Popular Group. Specifically, ten replicates with ten sites per replicate were created.

C. FINAL SAMPLES

Once replicates had been created, the final sample was achieved as follows. First, each of the 100 sites in the Most Popular Group was surfed. Next, for the Random Sample, one of the 54 replicates (containing 15 or 16 sites) was chosen at random, and all sites on the replicate were examined by a Committee staff member ("surfed"). This procedure was repeated until the number of sites surfed exceeded the target sample size.[9] Once a replicate had been selected, all sites on that replicate were surfed.[10]

At this stage, some sites were excluded from the Survey for one of three reasons: they were "adult" (*i.e.*, pornographic) sites, they were sites primarily directed to children 12 and under,[11] or they were inaccessible.[12] Forty-five sites were excluded for these reasons. Once the data collection described below was completed, additional sites were excluded from both samples. First, all foreign sites were excluded.[13] Second, all sites primarily directed to other businesses as opposed to consumers (*i.e.*, business-to-business sites) were excluded.[14] Finally, certain duplicate sites were excluded.[15] Altogether, 50 sites were excluded as foreign, business-to-business, or duplicates. The following chart sets forth the number of sites in the sampling pool and final sample for both the Random Sample and the Most Popular Group.

Sample	# Sites Examined	# Sites Excluded	Final Sample Size
Random	421	86	335
Most Popular	100	9	91

II. THE SURVEY

The Survey itself was divided into three separate parts: (1) a surf of all Web sites to ascertain their information collection practices and privacy disclosures; (2) a separate surf of Web sites to determine the use of third-party cookies; and (3) content analysis of the privacy disclosures found during the first surf.

A. INFORMATION COLLECTION AND PRIVACY DISCLOSURES

Sixteen Commission staff members, including attorneys, legal assistants and investigators ("surfers") surveyed the sites in the samples during a two-week period in February 2000. The surfers were not involved in designing the Survey, in the subsequent data analysis, or in drafting this report. Each surfer underwent a day's training in the technical skills of visiting and reviewing Web sites and in the use of the Survey questionnaire.[16] Surfers conducted the Survey in two rooms using computers equipped with Pentium III 500 processors and Windows 98 and connected to the Internet with a 1.5 MB SDSL link. All machines had at least the following plug-ins: RealAudio, QuickTime, and Macromedia Flash Player. Staff attorneys serving as supervisory proctors were present in the room at all times during the Survey to handle any technical difficulties and answer questions.

Surfers were randomly assigned replicates from both samples and instructed to visit each site listed on the replicate and to spend no more than twenty minutes surfing each site. Once a surfer concluded that a site qualified for inclusion in the Survey (it was not inaccessible, an "adult" site, or a site directed to children), the Survey questionnaire was completed. Surfers were instructed to determine whether the site had a privacy seal, had any information practice disclosures, and collected any personal information.[17] Surfers were instructed to print each site's home page and every page on which an information practice disclosure was located. Each site not excluded by the surfer was then examined again by a second surfer, who looked for any additional information practice disclosures.[18]

B. THIRD-PARTY COOKIES

All sites not excluded by the surfers were then examined for third-party cookie placement by six Commission interns ("cookie surfers") using two dedicated computers whose cookie cache had been cleared prior to the project. The browsers on the computer were set to notify the user if a cookie was being placed. The interns each underwent a half day's training on how to ascertain whether a third party was attempting to set a cookie on a site and how to complete the third-party cookie questionnaire.[19] Each cookie surfer was randomly assigned sites from the samples to visit. If a cookie alert indicated that a domain other than that listed on a replicate was attempting to set a cookie, the third-party cookie questionnaire was answered in the affirmative and the cookie surfer noted the URL of the domain on the questionnaire. In the event that no third-party cookie was found, a second cookie surfer would check the site to ensure the accuracy of data.

To determine whether third-party cookies observed during the online phase of data collection for the Survey were sent by network advertising companies engaged in profiling, Commission staff reviewed the completed third-party cookie survey forms and visited the Web sites associated with the domains of the observed cookies. Only companies whose Web sites explicitly stated that the company targeted banner ads on the basis of consumer characteristics were classified as "profilers."

C. CONTENT ANALYSIS

A third group of 17 Commission staff served as content analysts who reviewed the privacy disclosures of those sites that had such disclosures (either a privacy policy or an information practice statement). The content analysts underwent four half-days of training in the use of the content analysis form[20] and worked in pairs. Each pair was randomly assigned ten sites at a time.[21] Each analyst in the pair independently reviewed all of the disclosures for each assigned site and completed a content analysis form. Once both members of the pair had completed their independent review, the pair met and reconciled their answers for each site on a final content

analysis form. Where their answers were the same, they simply indicated the answer on the final content analysis form. Where their answers differed, the analysts discussed the question at issue and arrived at a consensus answer.[22] All sites with at least one privacy disclosure were reviewed and analyzed by two content analysts who ultimately agreed on the answer to the questions on the content analysis form.[23]

D. DATA ENTRY AND DATA ANALYSIS

Once all of the sites had been surfed, cookie-surfed, and, in the case of sites with privacy disclosures, content analyzed, data were entered by three pairs of data-entry personnel. The data-entry teams worked in pairs, with one member of the pair reading off answers to the second member of the pair who inputted the data. These numbers were then manually checked for accuracy by the data-entry teams. A set of queries was then run on the data to ensure that the data was internally consistent, *i.e.*, that all conditional answers were answered or left blank, as appropriate. All errors were corrected prior to substantive data analysis.

Finally, the data was analyzed using Intercooled Stata 6.0 for Windows 98/95/NT, and manually reviewed for errors by several Commission attorneys involved in the preparation of the report. Two analyses were performed with the data. One analysis focuses on the performance of Web sites and seeks to estimate the proportion of sites whose privacy policies fall into various categories. This analysis was performed on both the Random Sample and the Most Popular Group. Estimates for the Random Sample are reported together with 95-percent confidence intervals, which convey the margin for error on either side of the estimates.[24] For the Most Popular Group there is no sampling error, because the results were obtained using a 100-percent sample – a census.

The second analysis performed on the data, referred to as the weighted analysis, seeks to represent consumer experiences and gives proportionally more weight to sites with more traffic.[25] This weighting scheme shifts the focus of the analysis from sites to unique site visits.[26] For example, instead of representing the proportion of sites that post a privacy policy, the

weighted analysis represents the proportion of all unique site visits to the most heavily-trafficked sites that were made to sites that post privacy policies.[27]

It is important to note that the population from which the Random Sample was drawn excluded sites with fewer than 39,000 unique visitors in one month. Thus, the weighted results represent only the likelihood that a consumer surfing only sites with 39,000 or more visitors per month will encounter a particular practice. The weighted results represent consumer experiences only on that part of the Web from which the sample was drawn, and are not generally representative of consumers' online experiences.[28]

APPENDIX A: ENDNOTES

1. Nielsen//NetRatings provides online publishers, e-commerce companies, Internet advertising and marketing firms, and others with audience information and analysis about how people use the Internet, including what sites they visit, what ad banners they see, and the demographics of the users.

2. There were over 5,600 domains on the list; the unduplicated reach of all sites on the list was 98.3% (*i.e.*, it was estimated that 98.3% of all active Web users visited at least one of these sites at least once in the month of January 2000).

3. The sampling frame used in the 1999 Georgetown survey was a list of the top 7,500 servers. GIPPS Report, App. B at 4 (1999), available at <http://www.msb.edu/faculty/culnanm/gippshome.html>. Multiple servers for a single domain were then eliminated, and domain served as the unit of analysis for that survey as well. *Id.* at App. B at 5. The methodologies differ, however, in that domains with multiple servers had a greater chance of being included in the sampling pool and the sample in the Georgetown study, but not in the Commission's study, which used a list of domains as the sampling frame. *Id.* at App. B, n. iii.

4. Nielsen//NetRatings has recruited and maintained an Internet panel, a nationally representative sample of persons living in United States households with Internet access from the home. In January 2000, the month for which the data for the Commission's Survey was gathered, the panel included more than 40,000 individuals. The panel is constructed using a random digit dial, telephone recruitment process. Repeated attempts are made to identify and recruit all eligible households, in order to ensure that the Nielsen//NetRatings sample is as representative of the universe of Internet users as possible. Households receive a $50 U.S. savings bond every six months for the duration of their participation in the research.

   Proprietary tracking software is installed on the computers in all eligible, participating households. The software automatically tracks and collects information about Web pages viewed, ad banner viewing and clicking, and e-commerce activity, capturing all URLs and information about the time spent at each page, and transfers the data in real-time to Nielsen//NetRatings. The data are aggregated by Nielsen//NetRatings by property, domain, and unique site, or, for ad banner data, by advertiser, site, and banner.

   Nielsen//NetRatings also conducts monthly studies to determine the total size of the home Internet user population and demographic profile information. From this data, Nielsen//NetRatings is able to project from the panel data an estimated audience size for sites visited by the panelists. Further information about Nielsen//NetRatings is available at the company's Web site, <http://www.nielsen-netratings.com>.

5. Domains that Nielsen//NetRatings had classified in its "Adult Category" – which includes adult ISPs, adult content domains with age verification services, and sites that have explicit material in terms of language or content that should be viewed by adults only – were excluded from the list sent to the Commission. However, Nielsen//NetRatings applies its categories only to sites with a unique audience of 120,000 or

more. Therefore, some "adult" sites were included in the list provided by Nielsen// NetRatings and were excluded from the final sample, as described below.

6. There are over 7.8 million ".com" domains today, many of which receive few or no visitors. Network Solutions, *Network Solutions Surpasses 10 Million Domain Name Registrations*, available at <http://www.nsol.com/news/2000/pr_20000413.html> (Network Solutions alone has 10 million registered domains); Network Solutions, *Fun Facts About Domains*, available at <http://www.nsol.com/statistics/fun/fun.html> (as of January 2000, 78% of registered domains are ".com").

7. Because the Most Popular Group represents a census of the most popular sites, as opposed to a sample of those sites, the sampling techniques described below were used to create the Random Sample only.

8. As described below, "adult" sites, sites directed to children, and sites that were inaccessible for technical reasons were excluded at the beginning of the data collection process and did not qualify for the sample.

9. Because, as discussed below, additional sites were removed from the sample after data collection, more than the target number of sites were surfed.

10. Where a Most Popular Group site appeared on a replicate for the Random Sample, the site was not examined again, and the data gathered for the site during the surf of Most Popular sites was used in its place.

11. Such sites' information practices are covered by the Children's Online Privacy Protection Act, 15 U.S.C. § 6501, *et seq.*, and its implementing regulations, 16 C.F.R. Part 312, available at <http://www.ftc.gov/opa/1999/9910/childfinal>.

12. This last category included sites for which there was no DNS entry, a 404 Error message was received, or which were otherwise inaccessible.

13. Sites were deemed foreign if the registrant address, as listed in the Whois database at <http://www.networksolutions.com>, listed an address outside the U.S. A total of 31 foreign sites were excluded.

14. Two Commission staff re-examined each site in the proposed sample to identify potential business-to-business sites. A group of three Commission staff then jointly decided whether each such site was in fact a business-to-business site. A total of 13 sites were excluded as business-to-business sites.

15. "Duplicate" sites occur when two separate domain names lead to the same Web page. Duplicates were identified by a staff member visiting all the sites in the proposed samples, and identifying the homepage visited. Duplicates fell into one of three categories. Where both duplicate sites were in the Most Popular Group, the higher-ranked site was retained and the lower-ranked site was deleted, resulting in a smaller final Most Popular Group. (There were 4 such duplicates deleted.) Where one of the duplicate sites was in the Most Popular Group (*i.e.*, in the top 100 sites) and the other was not, the site in the Most Popular Group was retained and its data used in place of the duplicate site in the Random Sample. Thus, such a situation did not affect sample size.

(There were four such duplicate pairs.) Where both sites were in the Random Sample, one site was randomly deleted, resulting in a smaller final Random Sample. (There were two such duplicates deleted.) A total of six sites were deleted as duplicates, and four additional duplicate pairs resulted in the substitution of data, as described above.

For the weighted analysis, the traffic for the retained site was used. Where one of the duplicates was in the Most Popular Group and one was in the Random Sample, however, the retained site (from the Most Popular Group) and its traffic was included only in the Most Popular portion of the weighted analysis; *i.e.*, the site was not counted twice in the weighted analysis.

16. Copies of the surfers' instructions and the Survey questionnaire are included in Appendix B.

17. Staff did not ascertain whether sites in the Survey used hidden electronic means to collect personal information, but instead looked to order forms, registration pages, and other places where information was requested from consumers.

18. When additional information practice disclosures resulted in a change to an answer on the Survey questionnaire, a proctor first approved the change.

19. Copies of the cookie surfers' instructions and the third-party cookie questionnaire are included in Appendix B.

20. Copies of the content analysts' instructions and the content analysis form are included in Appendix B.

21. On occasion, fewer than ten sites were assigned for scheduling reasons.

22. The content analysts' agreement rate was 92% (number of agreements/total number of questions answered).

23. The content analysts were trained to treat inconsistencies in information practice disclosures as follows. First, questions were to be answered based on a site's treatment of *any piece* of information. Thus, for example, if a site offered choice with respect to disclosure of some but not all information collected, answers regarding such choice were answered in the affirmative. If, however, a site offered differing types of choice for different kinds of information, the least privacy-protective of those choices would control. Thus, if a site offered choice with respect to the sharing of personal information with some, but not all, third parties, the site received credit for choice as it relates to third parties. If the same site offered opt-in choice with respect to some third parties and opt-out choice with respect to other third parties, the site was classified as offering opt-out choice with respect to third parties. This methodology was necessitated by the complexity of sites' information practices and information practice disclosures, and does not reflect a policy decision on the part of the Commission.

24. These confidence intervals were constructed using the binomial probability distribution, which applies when analyzing dichotomous (yes/no) variables, such as we have here.

25. The weighted analysis is based on the data from both the Random Sample and the Most Popular Group. Data for both groups were combined in such as way as to give each group its proper weight, as dictated by the size of the population traffic it represented. (Sites appearing in both groups were counted only once.) This procedure was used (as opposed to simply assigning weights to each observation in the Random Sample) because it makes better use of the data regarding the Most Popular sites, where so much of the traffic takes place, and therefore gives a more accurate estimate.

26. The analysis treats the Nielsen//NetRatings estimates of unique site visits as precise measures of site traffic. Because this underlying traffic figure, which is based on estimates from survey panel data, actually contains some margin of error itself, the resulting weighted analysis figures are somewhat less precise than we report.

27. Some of the data is reported as a percentage of sub-samples. For example, the fair information practice figures are reported as a proportion of sites that collect personal identifying information, and not as a proportion of all sites in the samples. Where the data is reported as a percentage of a sub-sample (*e.g.*, all sites that collect personal identifying information), the weighted analysis included only those sites meeting the sub-sample's characteristics and all other sites were excluded.

28. If the sample had been drawn from the entire Web, the weighted analysis would have provided a more useful interpretation of the data. For example, in such a case the weighted analysis figure for "privacy policy" would represent the likelihood that a representative consumer would visit a site that posts a privacy policy each time he or she visits a different Web site. Audience estimates for *all* sites on the Web, which would be necessary to employ such a methodology, do not appear to be available.