# APPENDI A: METHODOLOG

statistical methods. All .com domains with at least 39,000 unique visitors were selected and ranked in order of audience size. 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. 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. 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. 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

## 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. 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 and worked in pairs. Each pair was randomly assigned ten sites at a time. 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

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.

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.

# 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% ( . ., 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

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 ( ., ., 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 sites on the Web, which would be necessary to employ such a methodology, do not appear to be available.