Examining the Content and Privacy of Web Browsing Incidental Information

Kirstie Hawkey and Kori M. Inkpen
Faculty of Computer Science, Dalhousie University
6050 University Ave., Halifax, NS Canada B3H 1W5
{hawkey, inkpen}@cs.dal.ca

ABSTRACT
This research examines the privacy comfort levels of participants if others can view traces of their web browsing activity. During a week-long field study, participants used an electronic diary daily to annotate each web page visited with a privacy level. Content categories were used by participants to theoretically specify their privacy comfort for each category and by researchers to partition participants' actual browsing. The content categories were clustered into groups based on the dominant privacy levels applied to the pages. Inconsistencies between participants in their privacy ratings of categories suggest that a general privacy management scheme is inappropriate. Participants' consistency within categories suggests that a personalized scheme may be feasible; however a more fine-grained approach to classification is required to improve results for sites that tend to be general, of multiple task purposes, or dynamic in content.

Categories and Subject Descriptors
H.5.3 [**Information Interfaces and Presentation**]: Group and Organization Interfaces – *Collaborative Computing, Web-based Interaction.* H.5.4 [**Information Interfaces and Presentation**]: Hypertext/Hypermedia – *User Issues, Theory*

General Terms
Human Factors, Experimentation, Theory

Keywords
Privacy, web browsing behaviour, personalization, web page content, ad hoc collaboration, field study, client-side logging.

1. INTRODUCTION
Web browsers are used in a variety of contexts, including during ad hoc collaboration such as when a small group gathers around a personal computer to work on a task. Web browsers have several convenience features (e.g. history, auto-complete, bookmarks) that assist with revisitation of web pages by storing traces of web browsing activity. However convenience features are often problematic in a group setting as the traces may reveal incidental information (i.e. information unrelated to the task at hand) that is inappropriate for the current viewing context. For example, information suitable for a friend to see may be inappropriate if viewed by an acquaintance or an authority figure with whom one would prefer to present a more formal persona [14].

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-323-9/06/0005.

It is not always clear to users exactly which traces of activities are being created and stored and which can subsequently be viewed by others during normal computer usage [28]. Nor is it clear whom all the future viewers will be and the context under which material will be viewed, particularly when devices are mobile and used in both personal and business settings [25]. Recent visualization advances, such as thumbnails of web pages in history files, may help users recognize a desired page [19], but also exacerbate privacy concerns due to the increased visibility of incidental information for others.

The content that is potentially visible to others depends on the browser settings, the preventative actions that people take (e.g. clearing history files), and the web pages that the users visit. A previous field study [15] examined participants' perception of the privacy of their web browsing and found patterns to their applications of privacy levels. However, the page title and URL were removed prior to the data being sent to researcher. It was unclear if differences in overall patterns of privacy application were due to differences in the inherent privacy concerns of participants or in the content being classified.

To maintain privacy in situations when web browser windows are visible to others, users must currently choose to either turn the convenience features off or periodically clear the stored information. Commercial privacy tools tend to assume that the vast majority of items are public in nature, with a small subset needing to be password protected, and that sites of both types are never viewed concurrently. A more nuanced approach is required for privacy management in this highly personal domain [21; 15]. Furthermore, given the magnitude of web pages visited [16], it is clear that some form of automation is required to classify the generated traces with an appropriate privacy level.

One approach may be to automatically classify pages as being one of several content categories and then to apply an appropriate privacy level to each category of content. However before we can design such a system, we must first understand the relationship between the privacy of web browsing traces and their content. If people hold common views on the sensitivity of content within a category, a general approach to privacy management may be feasible. If not, a personalized approach may be appropriate, allowing each user to set a default privacy level for a category. However, a personalized approach will only work if people are consistent within each category, applying a single privacy level to all visited pages.

We designed a field study to gather information about browsing activity both in an effort to learn more about what content is potentially visible and to evaluate the feasibility of users classifying privacy on a per-category basis. Page title and URL were used to determine the content categories of visited pages. In this paper, we examine how privacy levels change according to the category of visited page, how similar the participants were in

their privacy level applications, how consistent the participants were at classifying their browsing, and how accurate the participants were at choosing a theoretical privacy level for the categories. We first present related literature in the areas of privacy, privacy management tools, and web browsing behaviour. The field study is then described and our results are presented. We then discuss the feasibility of general and personalized solutions to privacy management and conclude with future work.

2. RELATED LITERATURE

2.1 Privacy

Online privacy concerns have been examined in great detail; for example, the Platform for Privacy Preferences Project [2] has developed standards that allow users greater control over the use of their personal information at participating websites. However, online privacy research has a different focus from the web browser privacy issues we present here. Online privacy research generally examines issues concerning the transfer of personal data to business or governmental entities; the relationships are between consumer and corporation. This is quite different from the privacy concerns associated with others viewing traces of previous web browsing activity in a co-located setting: there is no data being transferred (just viewed) and relationships are primarily interpersonal. Results from research into on-line privacy and other domains may not be directly applicable to the incidental information privacy domain, but can provide insights.

Ackerman et al. [6] examined privacy preferences for Internet users through a survey in 1998. The authors found that participants had differing levels of sensitivity about personal data, ranging from little concern about providing such information as their favourite television show to great concern over credit card and medical information. The authors suggest that an individualized approach is necessary given the large variance in reactions between participants.

Research about the privacy of information shared electronically has investigated privacy comfort for various types of information and recipients of that information. Cadiz and Gupta [9] found that privacy was highly nuanced; however, in general people were open to sharing information except with strangers. Olson et al. [24] examined the privacy of several types of information. They found that personal activities (e.g. viewing non-work related websites) and transgressions (e.g. viewing erotic material) are considered more sensitive than content such as availability and contact information. Activities convey the essence of a persona and knowledge of those activities can be even more sensitive when a user's identity is known since their hidden personae may be revealed [22]. With web browsing traces, a person's actions in one area (e.g. personal browsing) may later be viewed in another area (e.g. workplace). Additionally, there are likely several levels of sensitivity within the traces, the amount of highly sensitive content may fluctuate over time, and users may be less aware of what content is potentially visible.

2.2 Web Browsing Tools for Viewing Privacy

While there are commercial products that allow users to erase traces of browsing activities, those traces are often valuable for future transactions and may decrease productivity if removed entirely. There is no ability for users to record and later view a subset of their activities within a browsing session. As an example, WebRoot Software's Window Washer [3] allows a user to quickly delete traces such as auto completions, histories, and recent documents. However, with the exception of the ability to save selected cookies, the decision to erase a class of traces erases all instances indiscriminately.

COLLABCLIO [21] is a research system developed to support automated sharing of web browsing histories between colleagues. A binary classification scheme (public/private) allowed users to indicate which web sites in their history files could be shared. Users of the system expressed a desire for a more fine-grained classification scheme to reflect differing privacy needs for sub-groups of people.

2.3 Web Browsing Behaviour

Web browsing behaviour has been studied from a variety of perspectives. Beginning with the early work by Catledge and Pitkow [10] investigating how people were using their web browsers, there has been a great deal of research about how users navigate through the web both in the general case (e.g.. the study of web page revisitation patterns, as in [27]) and for specific areas such as information seeking behaviour (e.g. searching, as in [12]). Task-related research is particularly relevant to our examination of the types of web pages that people visit. During a diary study of knowledge workers in 2002, Sellen et al. [26] interviewed participants in front of their history lists and had them describe the web activities they had recently completed. Activities consisted of: transactions (5%), communications (4%), housekeeping (5%) and information seeking (86%) such as fact finding, information gathering, and browsing. A 2005 field study conducted by Kellar et al. [20] found that transactions accounted for 47% of the visited pages, with email being the most common transaction. Information seeking (fact finding, information gathering) accounted for 32% of visited pages and browsing for 20%. It is hard to compare results from these two studies directly as Sellen et al. presented their findings based on the percentage of activities participants recalled conducting and Keller et al. presented their results based on the percentage of visited pages logged. Keller et al. also found that task impacts which web browser convenience features are used.

Most of the research categorizing WWW use (e.g. Byrne et al.'s taxonomy [8]) focuses on the actions that people take and not on the type of content that is being viewed. Typically, content is examined through self-reports of the types of activities (e.g. shopping) participants engage in on the web (as in [23]). One exception is research by Curry [13] who sampled the URLs viewed by public library users and classified them by format and by subject. The author found that 39% of visits were email related; but, as not all pages received a subject categorization, content analysis in terms of relative amount of activity is limited.

There are many content classification schemes in commercial use, such as the Yahoo! Directory [4] which categorizes web pages using fourteen main headings and hundreds of subcategories. There are also commercial tools (e.g. [5]), both for corporate and parental use, for filtering out content that is deemed inappropriate. These tools may classify web pages into categories or use some combination of keywords and URL lists to filter inappropriate content and sites. However, web content filters suffer from both over blocking sites that shouldn't be blocked and under blocking sites that should be blocked [17]. A recent examination by Consumer Reports [11] shows that although research continues to improve content filtering, commercial systems are still often ineffective.

3. FIELD STUDY

The main goal of the study was to gather information about regular web browsing activity. This was required to enable examination of the relationship of the content of the browsing activity to the perceived privacy comfort levels that participants apply to visited pages.

3.1 Methodology

A week-long field study took place in March 2005. Earlier research in this area [15] examined laptop users who were primarily technical in nature. In order to study users with other characteristics, three different classes of participants were recruited: technical desktop (TD) users, non-technical desktop (ND) users, and non-technical laptop (NL) users. A screening process assessed participants' technical background and identified computers on which they conducted their web browsing. Participants were required to have logging software installed on their computer(s) so that we could capture the full picture of their personal and work/school related web browsing. Participants also needed to have had occasions in the past where their web browser window was visible by others, so that the concept of privacy in this situation had some relevance

3.1.1 Privacy Levels

Participants were asked to partition visited websites using a four-level privacy scheme: *public*, *semi-public*, *private*, and *don't save* (see Figure 1). *Public* sites are those someone is comfortable with anybody and everybody viewing, including the Queen (hence the crown in Figure 1). *Private* sites are those someone would be comfortable with only themselves and possibly a close confidant viewing. *Semi-public* sites fall somewhere in between: depending on the viewing context, pages may or may not be appropriate. Web sites classified as *don't save* primarily fall into one of two categories: ones that are irrelevant (i.e. would not want to revisit) or ones that are so private it is preferred that there is no record of having visited them at all.

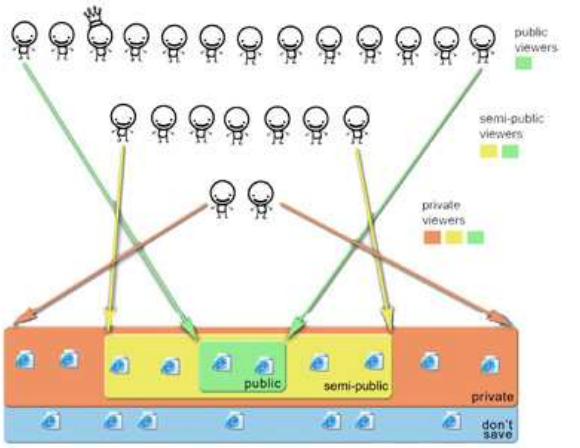

Figure 1. Privacy levels that participants used when classifying categories of web sites.

3.1.2 Study Instruments and Data Collection

Client-side logging software was developed to record contextual information about participant's web browsing during the week-long field study. We built a browser helper object (BHO) to work with Microsoft's Internet Explorer (IE). As each IE window opens, the BHO loads and logs all web sites visited until the window closes. All pages viewed in the browsing process were logged, even if navigation continued before the document fully loaded. We did not record individual frames or images loaded within a web document. Data collected included the web page visited (URL and page title), time stamp, and ID number of the browser window in which the page loaded.

Participants were also provided with an electronic diary (Figure 2) to use on a daily basis. The diary displayed details of all the visited pages (browser window ID, date/time stamp, page title, URL) and required participants to indicate how they would classify the privacy level of each web page if others were to view traces of this activity. Participants could annotate single or multiple entries with a privacy level. The entries could be sorted by any field, allowing participants to easily classify groups of pages. Participants could choose to sanitize entries in the diary by removing the page title and URL after applying a privacy level. Participants were asked to give a general reason for the sanitized browsing (e.g. "looking for medical information"); however, the default label was "no reason given". After classification, participants generated a report to email to the researchers. We hoped that the privacy afforded by participants' ability to selectively sanitize their browsing record would contribute to their willingness to engage in normal web activities while still providing us with context for most visited pages.

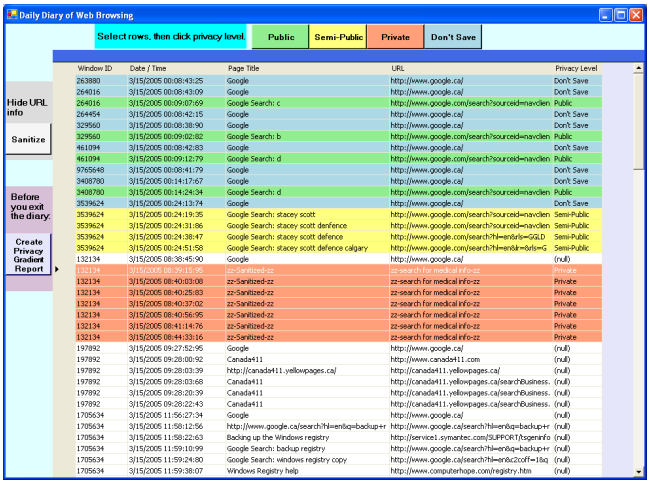

Figure 2. Screenshot of electronic diary used by participants to annotate their web browsing with a privacy level.

In addition to the diary portion of the study, participants completed pre and post study questionnaires. One of the questionnaires was a theoretical classification task where participants were asked to assign one of the four privacy levels to each of 55 web site categories, indicating their comfort if a site of this type appeared in their web browser (see [1] for a category listing). The categories (e.g. online games, news/media) and their descriptions presented to participants were based upon those used in commercial products to filter and block objectionable or non-productive internet content.

3.1.3 Content Categorization of Logged Data

Researchers used the same set of categories to classify all of the browser activity conducted by participants over the course of the week. The parental control feature of Zone Labs Security Suite [5] was enabled and all 34 categories offered (a subset of the classification task categories) were blocked. All browsing was

sorted by URL; each distinct URL was pasted into the address bar of a browser window. If the web site was blocked, its category was given as a reason. If the site was not blocked (approx. 50% of the time), we manually classified it according to the category descriptions and examples used in the classification task.

We classified some pages as *web content* management when it was clear that participants were using a content management tool within their browser rather than actually visiting a web page. Additionally, some entries were classified as *empty window* (a log entry with no URL). These entries occurred when an image (such as a web advertisement) was loaded into an empty pop-up window, when no home page was set in the browser, or as a result of scripting on a page.

3.2 Participants

Fifteen participants were recruited from the general community at Dalhousie University. Table 1 shows the breakdown of participants in each of the three groups recruited in terms of their age, sex, occupation, stated reasons for web browsing, and computer experience. Although we intentionally recruited participants with different technical backgrounds and types of computers used, participants within each group were not balanced by age, sex, or computer experience. We therefore will not attempt to make any comparisons between groups as privacy is a domain known for individual variability.

Table 1. Demographic breakdown of recruited groups of participants.

	Overall	Non-technical desktop	Non-technical laptop	Technical desktop
Age	27.8 (18-44)	27.8 (18-40)	22.8 (18-30)	31.2 (25-44)
Sex	5 M, 10 F	1 M, 4 F	1 M, 4 F	3 M, 2 F
Occupation	11 students 4 office staff	3 students 2 office staff	5 students	3 students 2 office staff
Computer Experience	9.7 yrs. avg. (6-20)	8.0 yrs. avg. (6-10)	11.2 yrs. avg. (6-15)	10.0 yrs. avg. (6-20)
Usual reasons for browsing	37% personal 18% work 45% school	31% personal 30% work 39% school	39% personal 3% work 58% school	42% personal 22% work 36% school

The ability to generalize our results may be limited as our participants are more highly educated than the general public and many were students. Given the educational domain from which participants were recruited, browsing activities may include more educational and reference sites than if participants were recruited from another domain.

4. Results

We first present results related to general browsing activity and privacy levels applied by participants followed by descriptive statistics of the categories of web sites visited by participants. We then examine the consistency of privacy labels applied within website categories. Results from the theoretical classification task are then presented and used to determine how accurate participants were at predicting their actual privacy classifications for categories. Finally, these results are used to examine category characteristics that impact consistency and accuracy.

4.1 General Browsing Activity and Privacy

The fifteen participants visited a total of 31,160 pages during the week. All participants used all privacy levels when classifying their web browsing, with the exception of two participants who didn't classify any visited pages as *don't save* (see Table 2 for a per participant breakdown of visited pages and privacy levels). These results demonstrate the highly individual nature of both web browsing behaviour and the application of privacy levels.

Table 2. Number of pages visited, privacy levels applied to pages, and number of distinct web site categories (total and with 10+ pages) per participant.

		Privacy level application (%)				# categories	
ID	# pages	public	semi-public	private	don't save	overall	With 10+ pages
ND1	699	36.6	50.4	0.6	12.4	20	9
ND2	3123	17.8	11.8	26.0	44.3	18	11
ND3	1084	36.5	59.7	3.5	0.3	19	9
ND4	936	14.9	59.4	9.0	16.8	15	10
ND5	2174	27.9	15.3	31.1	25.7	24	15
NL1	1261	64.7	2.1	9.5	23.7	15	10
NL2	1161	37.6	59.8	2.7	0.0	19	13
NL3	3284	52.5	29.3	8.2	10.1	26	16
NL4	1002	19.9	12.2	47.1	20.9	16	9
NL5	2019	18.2	3.5	24.7	53.6	28	17
TD1	1338	39.3	11.4	27.1	22.1	21	13
TD2	4070	23.6	15.9	55.9	4.6	29	18
TD3	4966	79.8	0.6	19.5	0.1	29	18
TD4	3125	24.0	33.6	39.1	3.3	25	16
TD5	918	83.8	9.3	7.0	0.0	20	11
Total	31160	--	--	--	--	41	37
Avg.	2077	40.0	19.6	25.3	15.1	21	13

Overall, 40.0% of visited pages were classified as *public*, 19.6% as *semi-public*, 25.3% as *private*, and 15.1% as *don't save*. Participants varied in their overall application of privacy levels; examination of the privacy levels applied to the categories of browsing is required to determine whether differences were due to inherent privacy concerns of participants or differences in the sensitivity of web sites visited.

4.2 Category of Visited URL

Participants visited sites from 41 of the 55 possible web categories used in the theoretical classification task (see [1] for full list of categories). Each participant visited a subset of those categories (15-29, avg. 21). Table 2 gives the number of categories per participant. Only 21 categories included page visits by at least half the participants. Table 3 lists the categories and gives the total number of pages visited, the number of participants with browsing in each category, and the number of participants that visited 10 or more pages in each category (ordered by total participants in the category and then by overall pages visited). It is important to note that participants had very different usage patterns within a category. For example, news/media appears to be a very popular category with 14 participants visiting a total of 1320 pages; however, a single participant accounted for 1032 of those pages and only 7 participants visited 10 or more pages categorized as news/media.

Categories with less than 40 total cases each were grouped into *other*, including chat/instant messaging, cult/occult, gambling, gay/lesbian, hacking/proxy avoidance, military, sex education, and vehicles. Only 6 participants sanitized some of their web page visits, accounting for 433 pages. Of these, 107 did not have sufficiently detailed explanations to assign the page to a web

Table 3. Per category descriptive statistics: overall pages, number of participants with page visits (total, 10+ pages), within category consistency, accuracy, predominant privacy levels applied, and cluster membership. Highlights show expected dominant privacy level(s) based on cluster membership.

Category	Overall page total	# part.		Consistency (%)	Accuracy (%)	Predominant privacy level				Cluster
		Total	10+ pages			Public	Semi-Public	Private	Don't Save	
Search Engines/Portals	6310	15	15	61	46	6	3	4	1	C4
Education	3315	15	14	65	57	10	3	1		C4
Email	5082	14	14	81	77	1	5	8		C5
Reference	2055	14	13	76	51	8	3		2	C4
News/Media	1320	14	7	96	95	7				C2
Shopping	770	14	10	80	38	6	3		1	C1
Arts/ Entertainment	665	14	12	81	59	5	3		4	C1
Society/ Lifestyle	1136	13	8	93	10	5		1	2	C1
Web Advertisement	158	12	3	71	55	2	1			C1
Computers/Internet	146	12	5	66	55	4	1			C4
Financial Services	510	11	10	90	75		1	8	1	C5
Government/ Legal	385	11	5	88	78	2	3			C2
Web Communication	660	10	6	76	32	3	1	2		C4
Sports/Rec./ Hobbies	431	10	5	91	39	3	1		1	C4
Travel	366	10	7	80	45	3	2	1	1	C4
Software Downloads	236	10	6	83	61	5			1	C2
Health	165	10	6	92	16	3	2	1		C4
News Group	1303	9	3	78	70	1	2			C3
Job Search/ Career	449	9	4	80	86		2	2		C3
Business/Economy	178	8	4	84	60	1	1	1	1	C4
Religion	127	8	2	78	44	1			1	C4
Online Games	520	7	5	90	74	2	1	1	1	C2
Streaming Media/MP3	148	7	4	76	69	2	1		1	C1
Web Content Mgmt.	598	6	4	80	--	1	2	1		--
Political /Activism/Adv.	57	6	2	95	71	2				C2
Dating/ Personals	600	5	4	88	18		1	3		C5
Internet Auction	101	5	3	92	95	1	2			C3
Humor/Jokes	77	5	1	79	73		1			C3
Restaurants/ Dining/Food	279	4	3	99	88	1	1		1	C2
Pornography	258	4	2	88	86			2		C5
Web Hosting	60	4	2	80	29		2			C3
Real Estate	147	3	1	100	99	1				C2
Brokerage/Trading	110	3	1	95	0	1				C2
Int. Apparel/ Swimsuit	94	2	1	97	95			1		C5
Other	229	13								
Empty Window	21115	15								
Total	31160	15								

browsing category. A further 14 pages could not be classified as the page was no longer accessible at the time of coding and did not have sufficiently descriptive URLs or page titles.

4.3 Privacy Levels Applied

Not surprisingly, participants classified different categories of browsing with varying privacy sensitivities. A K-means cluster analysis of the 33 most common categories, grouped them into five clusters based on the relative proportions of pages that were classified at each privacy level (see Table 4 for cluster centers, Table 3 for cluster membership, and Section 4.7 for an in-depth discussion of the characteristics of categories within each cluster). Examination of the cluster centers reveals the predominant privacy levels that characterize each cluster:

C1: *public/don't save*
C2: *public*
C3: *semi-public*
C4: *mixture*
C5: private

Table 4. Results of cluster analysis of web page categories by applied privacy levels. Highlights indicate the privacy levels that characterize each cluster.

	Clusters	C1	C2	C3	C4	C5
Privacy Level	**Overall**	**Final Cluster Centers**				
Public	40.0%	48%	84%	23%	51%	3%
Semi-Public	19.6%	10%	8%	72%	22%	10%
Private	25.3%	3%	2%	3%	16%	81%
Don't Save	15.1%	39%	6%	2%	11%	6%
Number of Categories	5	8	5	10	5	
% of Total Page Visits	9.2%	9.8%	6.4%	44.1%	21.0%	

The cluster analysis alone gives no knowledge of whether the different privacy levels applied within a category are a result of participants not being in agreement with each other (between participant consistency) or not being consistent in how they assigned privacy levels to pages within that category (within category consistency). Furthermore, as some participants contributed much more data than others (i.e. visited more pages within a category), their privacy patterns may dominate.

4.4 Consistency

For consistency, we report on normalized data for each participant. For each participant with 10 or more pages of browsing in a category, we determined the predominant privacy level that they applied to their browsing in that category and calculated the percentage of pages that were classified at that privacy level. We omitted instances where a participant had fewer than 10 page visits in a category; these categories were deemed to be less relevant to participants and their consistency less reliable.

4.4.1 Between Participants Consistency

Between participants consistency examines how much agreement there is between participants in their privacy classification of page visits in a category. We compared the predominant privacy level applied by participants within each category (see Table 3 for a breakdown of the number of participants that classified the majority of their page visits in the category with each privacy level). In only 4 of the 30 categories with two or more participants was there complete agreement between participants with respect to which privacy level was applied.

Furthermore, over half of those categories (16/30) have a subset of participants whose predominant privacy level in that category that was not consistent with the category's cluster membership. The highlights in the privacy level cells in Table 3 represent the expected predominant privacy levels according to the cluster membership of the category. For example, for On-line Games, the

overall application of privacy levels resulted in this category falling in Cluster 2 (public). If participants were consistent with each other, we would expect all participants to have public as their primary privacy level (hence the highlight in the Online Gaming/Public cell in Table 3). However, of the 5 participants with 10 or more online gaming page visits, an examination of their predominant privacy levels reveals that only 2 of the participants labeled most visited pages as *public*; the other 3 participants each labeled most of their visited pages with one of the other privacy levels (*semi-public*, *private*, and *don't save)*.

4.4.2 Within Category Consistency

Within category consistency examines how consistent participants were in assigning privacy levels to pages in that category, regardless of which privacy level was applied predominantly. For each category, for each participant with 10 or more page visits, we computed the consistency in each instance as the number of pages classified at the primary privacy level divided by the total number of page visits, thus normalizing the consistency on a per-participant basis. The overall consistency for each category was obtained by averaging the per-participant results. Across all categories, the average consistency was 81% (61-100%, see Table 3 for per-category results). For many categories, participants may be able to set a default privacy level that classifies most pages accurately, but some categories (e.g. Search Engines/Portals, Education) are problematic.

4.5 Website Classification Task

During the theoretical website classification task, participants assigned a single privacy classification to each of the web categories. The results are shown in Figure 3, which illustrates how differently participants felt about the sensitivity of the categories; in only two categories (News/Media, Computers/ Internet) did all participants use the same classification. It should be noted that the classification task was completed in terms of privacy of content, not relevance. Therefore, use of *don't save* may be more likely an indication that a category was considered 'extremely private' rather than 'irrelevant'. The dual nature of this privacy level may have contributed to classification inaccuracies.

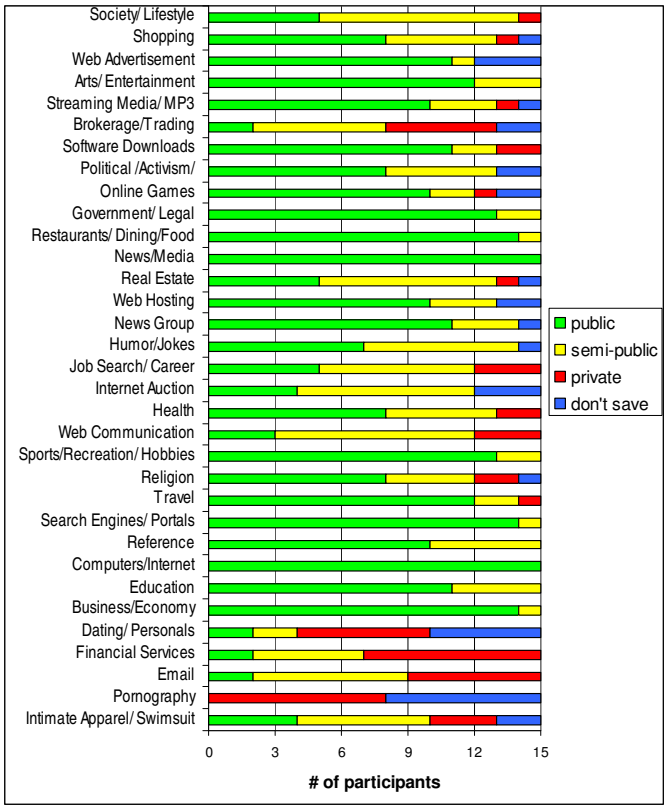

Figure 3. Results of theoretical website category privacy classification task, ordered by cluster.

4.6 Classification Accuracy

We examined how accurate the classification task was as a predictor of a participant's actual labeling of their browsing. For each participant, we computed accuracy as the number of web page visits that were labeled at the same privacy level that the category was labeled during the theoretical classification task. Overall, 57.8% of the page visits in the 32 most common categories were classified accurately (see Table 3 for per category results, no accuracy results are available for web content management as it was not a category in the classification task). Accuracy varied greatly by category, ranging from 0% (Brokerage/ Trading) to 98.6% correct (Real Estate). Accuracy also varied greatly by participant (36%-82%, avg. 58%).

4.7 Category Characteristics

The clusters identified in Section 4.3 were formed based on the overall applications of privacy levels by participants. We now use the clusters to frame a discussion of the characteristics of categories that impact consistency and accuracy results.

4.7.1 Cluster C1: Public/Don't Save

This cluster accounts for 9.2% of all pages visited and included the categories Arts/Entertainment, Shopping, Society/Lifestyle, Web Advertisements, and Streaming Media/MP3 (see Figure 4). These categories are fairly general and may contain pages with content of varying sensitivities. Participants labeled most (80-95%, avg. 87%) of the pages in each category as being either *public* or *don't save.* Still 5-15% of pages were classified as private or semi-public (i.e. potentially private) depending on the viewing context. Given the high amount of public browsing, for these categories, the *don't save* label most likely means a page is irrelevant, rather than being extremely private, with the possible exception of the Streaming Media/MP3 category which exhibits a lower percentage of public pages.

Participants were not very consistent (67-84%, avg. 77%) in their application of privacy levels for categories in this cluster. They

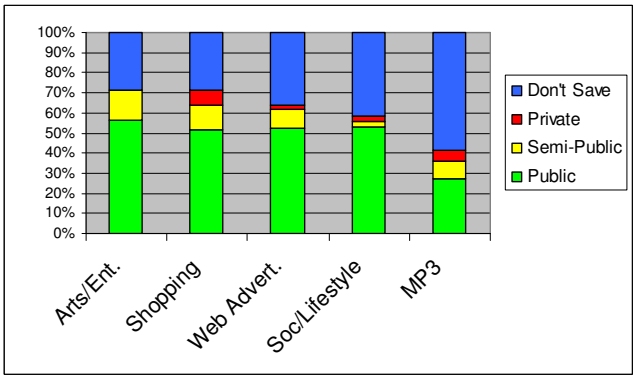

Figure 4. Relative privacy levels of categories in C1 (public/ don't save).

also exhibited poor accuracy (10-69%, avg. 46.1%) in their ability to predict their labeling of web pages in these categories. This was likely due to the large percentage of *don't save* labels applied to the web browsing. Accuracy was reduced as few participants classified the categories during the theoretical task with a privacy sensitivity of *don't save* (i.e. extremely private).

4.7.2 Cluster C2: Public

This cluster accounts for 9.8% of all pages visited and included the categories Real Estate, News/Media, Brokerage/Trading, Government/Legal, Political/Activist/Advocacy, Restaurants/ Dining/Food, Online Games, and Software Downloads (see Figure 5). Participants labeled the majority (75-100%, avg. 84%) of the pages in each category as being *public*. However, there were still some potentially sensitive pages occurring within these categories (i.e. 11-20% of the visited pages labeled as either private or semi-public for 5/8 categories).

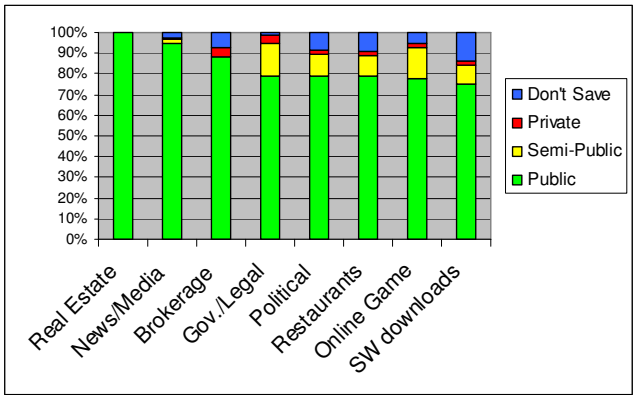

Figure 5. Relative privacy levels of categories in C2 (public).

Participants were very consistent (83-100%, avg. 92%) in their application of privacy levels within this category with the exception of Software Downloads (67%). The variable content of the downloads (e.g. free software updates, purchased products, warez) may have reduced the consistency for this category. Categories were quite accurate (61-99%, avg. 81%) with the exception of Brokerage/Trading which had 0% accuracy. Examination of the data revealed that the 3 participants with browsing in this category were conducting diverse activities, from visiting informational sites (e.g. finance.yahoo.com) to logging in to conduct secure trading transactions. The large number of public pages reflects informational pages, while the secure transactions were primarily classified as private. This category has characteristics very similar to categories in cluster C4 (mixture), it is possible that with a larger sample (this category only had 3 participants and a total of 110 page visits), a different overall privacy profile would have emerged.

4.7.3 Cluster C3: Semi-Public

This cluster accounts for 6.4% of all pages visited and included the categories News Group, Job Search/Careers, Humor, Web Hosting, and Internet Auction (see Figure 6). Participants classified the majority (64-78%, avg. 74%) of pages in each category as *semi-public*, indicating that the pages may be public or private depending on the viewing context. Interestingly, with the exception Job Search/Careers, these categories had very few (in 3 cases, none) pages indicated as being private.

Participants were not very consistent in their application of privacy in these categories (78-80%, avg.79%) with the exception

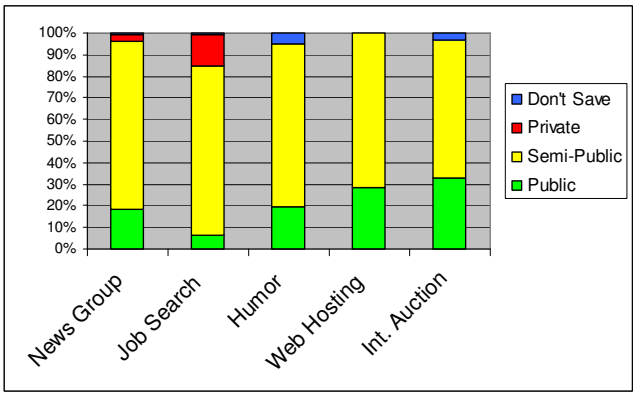

Figure 6. Relative privacy levels of C3 categories (semi-public).

of Internet Auction (92%) This is likely due to the general nature of News Groups sites (each with different topics of discussion), the varying subject matter of Humor sites and the personal content found in sub-pages of Web Hosting sites. The Web Hosting category exhibited atypical accuracy for this cluster (29%). The four participants with page visits categorized as Web Hosting indicated they would classify sites of this type as public; but, 3 of the 4 classified the majority of the sites as semi-public instead. Upon further investigation, some of these pages received a secondary classification of web content management (e.g. PageBuilder functionality on GeoCities) or contained personal content (e.g. photos on photobucket.com). The web hosting site itself may be considered public by most, but actual content on sub-pages may be more sensitive. This personal content was not apparent in the category descriptions provided to participants.

4.7.4 Cluster C4: Mixture

This cluster accounts for 44.1% of all pages visited and included the categories Education, Web Communication, Sports/ Recreation/Hobbies, Business/Economy, Computers/Internet, Reference, Search Engines/Portals, Religion, Travel, and Health (see Figure 7). These categories were frequently visited, both in terms of number of pages (165-6310 pages per category) and in number of participants (8-15 participants per category). Categories in this cluster were characterized as having a more even spread across privacy levels than in other clusters (*public*: 30-64%, avg. 51%; *semi-public*: 14-36%, avg. 22%; *private*: 1-37%, avg. 16%; *don't save*: 0-24%, avg. 11%).

Participants exhibited relatively low consistency (60-92%, avg. 78%.) in their application of privacy levels to their browsing in

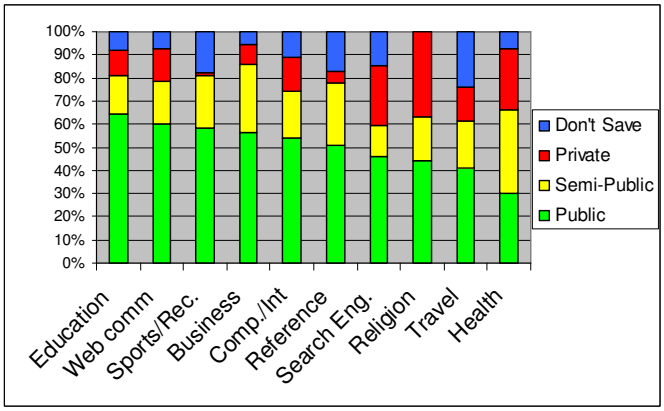

Figure 7. Relative privacy levels of categories in C4 (mixture).

these categories. Their accuracy rates at predicting which privacy level they would apply to these categories were also low (16-60%, avg. 44%). Participants were most consistent (91-92%) at classifying their Health and Sports/Recreation/Hobbies page visits. Further analysis of the categories with lower results revealed that many were multi-purpose (e.g. a general university site may have sub pages related to specific assignments and grades), had varying tasks associated (e.g. a travel page can be informational or a transaction such as a secure flight booking), or had sub-pages at varying content sensitivities (e.g. search results reveal more sensitive content than the search engine home page).

4.7.5 Cluster C5: Private

This cluster accounts for 21.0% of all pages visited and included the categories Intimate Apparel/Swimsuit, Dating/Personals, Pornography, Financial Services, and Email (see Figure 8). Categories in this cluster are characterized as being *private* (58-94%, avg. 81%) or potentially private depending on the viewing context (total private/semi-public: 85-97%, avg. 91%). For these categories, it is likely that those pages classified as don't save include some that are extremely private rather than just irrelevant.

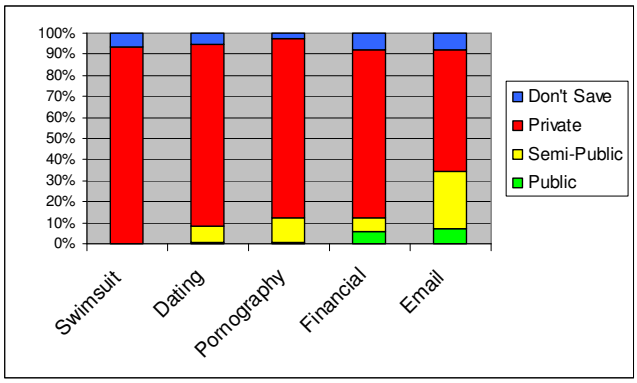

Figure 8. Relative privacy levels of categories in C5 (private).

Participants were fairly consistent (81-97%, avg. 89%) in their application of privacy levels in this cluster. With the exception of Dating/Personals (18%), participants were quite accurate (75-95%, avg. 83%) at predicting how they would label page visits in this cluster. For the 4/5 participants with more than 10 page visits in the Dating/Personals category, 2 accurately predicted the majority as *private*; 1 predicted *private* and labeled most as *semi-public*; and 1 (accounting for 76% of the total page visits) predicted *don't save*, but labeled most as *private*.

For sites such as Financial Services and Email (i.e. personal, but not sexually explicit) one marker of content sensitivity appeared to be whether or not a secure transaction was taking place. Across all browsing, there were 6963 secure pages (https); categories that had a high proportion of secure pages included Email (71%), Financial Services (74%), Web Communication (46%), Search Engines/Portals (42%), Brokerage/Trading (17%), and Travel (16%). Overall, 57% of secure pages were classified as private and 13% as public. The converse was true for pages that were not secure (14% private, 52% public); the proportion of don't save and semi-public pages remained consistent.

5. DISCUSSION

5.1 General Privacy Management System

For a general privacy management system (i.e. one size fits all) to be suitable, there would need to be universal agreement between users on an appropriate privacy classification for each category of web page. The results of the theoretical classification task showed that participants differed greatly in their privacy classifications of categories; indeed only two of the categories had complete agreement between participants. Examining the actual privacy labels applied by participants and the clusters that formed, we find that some categories did exhibit basic agreement among participants. However even for those categories that were predominately labeled with one privacy level (e.g. categories in clusters C2 (public), C3 (semi-public), and C5 (private)), there were some pages that were labeled differently. Inconsistencies were found to be both between participants (with respect to the predominant privacy level) and also within participants' classifications. This was particularly true for the categories in C1 (public/don't save) and C4 (mixture) where a variety of privacy levels were applied. As these two clusters account for over 50% of the pages visited, we conclude that a general privacy management scheme would not be effective.

5.2 Personalized Privacy Management System

For a personalized privacy management system to be feasible, participants would need to be fairly consistent at their desired privacy level within each category of web browsing activity. Many categories were very consistent; 12/34 categories examined had greater than 90% consistency. However, many categories exhibited higher inconsistencies; 13 of the categories have more than 20% inconsistency between the actual labels applied and the predominant privacy level. This was most pronounced for those categories in clusters C1 and C4 (public/don't save and mixture) which tended to have lower consistency results.

Participants would also need to be able to specify the default privacy level for each category of web browsing. Prediction accuracy varied greatly and some participants were unable to predict correctly the majority of their labeling. Some of the inaccuracy is due to categories with low consistencies; if the pages in a category are fairly evenly divided, any predicted privacy level will fail to accurately classify the majority of pages.

Clearly, we must be able to improve consistency results for those categories with low consistency ratings and also improve participant accuracy in assigning default privacy levels for personalized privacy management system to be effective. We next discuss some of the characteristics of the web site categories that lead to inconsistent and inaccurate privacy ratings and then give recommendations for increasing accuracy.

5.3 Reasons for Inconsistency and Inaccuracy

Recent research (such as [6]) has been cautioning that actual behaviour with respect to privacy practices often does not follow stated privacy concerns. For example, attitudinal information about on-line privacy practices gathered in a survey often did not match actual behaviour during a purchasing scenario [18]. People may idealize their privacy preferences, but at the time of action other circumstances may influence their actions. Acquisti [7] has proposed enriched privacy models to increase predictive accuracy by including psychological models of personal behavior such as immediate gratification and self-control. A disconnect between privacy preferences and labeling behaviour was likely not a major source of inconsistency during our study due to its short term and theoretical nature. Any effects due to social desirability (i.e. participants specifying a privacy level that they feel is the socially acceptable answer) should have been mirrored in both the

theoretical classification task and the classification of their actual web browsing. One cause of inaccuracy may have been that the example websites and category descriptions given in the theoretical classification task may not have adequately conveyed to participants the sensitivity range of content that may be visible.

Some of the inconsistency and inaccuracies within website categories may be due to the "it depends" nature of the semi-public privacy level. The uncertainty of whether visited web pages within a category should be public or private is often due to what is appropriate for the various categories of potential viewers. However, it may also be due to the variety of potential content in a given category. The potential viewing context is therefore partially resolved when a specific page is viewed. For example, the Web Communication category was predominately predicted to be semi-public and in actuality, the dominant privacy level was split between public (3/6 participants), semi-public (1/6) and private (2/6).

Similarly, the dual nature of don't save (irrelevant or extremely private) causes inconsistencies related to privacy. In some cases it is applied as a fourth privacy level (extremely private) and in other cases it was applied as a mechanism for not cluttering the convenience features with irrelevant pages (i.e. those that a participant would never bother to visit again). This dual nature was intentional during the study, allowing participants to classify the end result (not having a page saved) without having to admit to extremely sensitive browsing, Much of the inconsistency (particularly for cluster C1 (public/don't save)) may be resolved if the dual nature is separated.

As presented in the results, there were several characteristics of web page categories that led to inconsistencies and inaccuracies. Some of the categories used were very *general* so sites with very different content would be applied with the same privacy level. For example, the category News Group may be applied to forums that discuss very different topics in terms of sensitivity. The content must be examined to determine the appropriate privacy sensitivity with respect to future viewing. Participants may be unable to give a single default privacy level for these categories.

Categories may also include sites with varying task purposes (i.e. informational and transactional). For example, a page categorized as Brokerage/Trading may give general information or contain details about an individual's personal transactions. Often transactional web sites have an entry page that is less sensitive than the sub pages. Login pages may serve as markers for the transition between more public viewing and the subsequent secure pages that may be more private in nature.

Websites may also be very complex and are often dynamic in nature. Such sites may have varying content sensitivities depending on the content visible on a given page or at a given time. For example a News/Media site may have specific news stories that may be more sensitive than others. A Search Engines/Portal page may be considered public; the search results may be more sensitive in nature.

5.4 Recommendations to Increase Accuracy

To increase accuracy, two main issues must be resolved. The first is finding methods of further categorizing websites to resolve inconsistencies due to the generality, multiple task purposes and dynamic nature of sites. The second is improving participants' ability to predict the privacy levels they would apply.

Some heuristics exist that may help resolve some of the inconsistencies within categories. For those sites that are very general, being able to categorize the content at the sub-page level may improve accuracy. One method would be to use a customizable list of keywords. A similar scheme could be used for categories with pages that are often dynamic. In order to distinguish between informational web sites and transactional sites, it may be necessary to identify log-in pages or secure pages (https) and modify the privacy level accordingly.

Whatever the categorization scheme, it must be effectively communicated to users. While the classification scheme we used provided both descriptions and example web sites, in some cases it did not appear to be apparent to participants just how diverse categories were with respect to the types of pages and content that may be included. When determining an appropriate privacy level, the cost of others viewing traces of a previous web visits can only be determined if it is clear to participants what sorts of information may be visible.

6. CONCLUSIONS AND FUTURE WORK

We examined the privacy comfort levels that participants had if others were to view traces of their web browsing history. Content categories were used by participants to theoretically specify their privacy comfort for each category and by researchers to partition participants' actual browsing. Results revealed that the categories of web pages clustered into five groups based participants' overall application of privacy levels to their web browsing.

Inconsistencies between participants, both for their theoretical and actual privacy classifications, suggest that a general privacy management scheme is inappropriate. While participants often applied different privacy levels from each other for categories, results showed that participants were personally consistent within most categories. This suggests that a personalized scheme may be feasible; but a more fine-grained approach to classification is required to improve results for web sites that tend to be very general, have multiple task purposes, or have dynamic content. Additionally, participants' overall poor accuracy at specifying theoretically how they will actually label the web sites in a category indicates that better descriptions of the types of sites that may fall within a category is required as well as the types of sensitive information that may be encountered.

Previous research has shown that there are privacy patterns (e.g. streaks at a given privacy level) and temporal patterns (e.g. rapid bursts of browsing) to web browsing activities [15]. Further analysis of the contextual data from this field study will be used to explore how the content categories of visited web pages impact these patterns. Additionally, our data suggests that browsing is often partitioned with more sensitive browsing occurring in a single window while other windows have less sensitive content. We will use the content categorizations to gain a clearer understanding of how users partition their web browsing activities between windows.

7. ACKNOWLEDGMENTS

Funding provided in part by NSERC and NECTAR. Thanks to Melanie Kellar for collaboration with the data collection software, the members of the EDGE Lab for their support and feedback, and to the participants who allowed us to record their web browsing.

8. REFERENCES

[1] Cerberian Web Filter Categories. www.webrootdisp.net/audit/rating-descriptions.htm.

[2] Platform for Privacy Preferences (P3P) Project. http://www.w3.org/P3P/.

[3] WebRoot Software | Window Washer. http://www.webroot.com/consumer/products/windowwasher/.

[4] Yahoo! Directory. http://dir.yahoo.com.

[5] ZoneAlarm Internet Security Suite Datasheet. http://download.zonelabs.com/bin/media/pdf/zaiss60_datasheet.pdf.

[6] Ackerman, M., Cranor, L. and Reagle, J. (1999). Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences. In *Proc. of EC '99*, Denver, CO, 1-8.

[7] Acquisti, A. (2004). Privacy in Electronic Commerce and the Economics of Immediate Gratification. In *Proc. of EC '04*, New York, New York, 21-29.

[8] Byrne, M., John, B., Wehrle, N. and Crow, D. (1999). The Tangled Web We Wove: A Taskonomy of WWW Use. In *Proc. of CHI '99*, Pittsburgh, PA, 544-551.

[9] Cadiz, J. and Gupta, A. (2001). Privacy Interfaces for Collaboration. Microsoft Research, Redmond, WA. Technical Report No. MSR-TR-2001-82.

[10] Catledge, L. and Pitkow, J. (1995). Characterizing Browsing Strategies in the World-Wide Web. In *Proc. of WWW 1995*, Darmstadt, Germany, 1065 - 1073.

[11] Consumer Reports (2005) Filtering software: Better, but still fallible. http://www.consumerreports.org/cro/electronics-computers/internet-filtering-software-605/overview.htm.

[12] Cothey, V. (2002). A Longitudinal Study of World Wide Web Users' Information-Searching Behavior.

[13] Curry, A. (2002). What are Public Library Users Viewing on the Internet?: An Analysis of the Transaction Logs of Burnaby, Brantford, Calgary, Winnipeg, and Halifax Public Libraries, National Library and Archives Canada Virtual Collection of Monographs and Periodicals, http://tinyurl.com/8v7qc.

[14] Goffman, E. (1959). The Presentation of Self in Everyday Life. Garden City, New York, Doubleday Anchor Books.

[15] Hawkey, K. and Inkpen, K. (2005). Privacy Gradients: Exploring ways to manage incidental information during co-located collaboration. Ext. Abstracts CHI 2005, ACM Press: 1431-1434.

[16] Hawkey, K. and Inkpen, K. (2005). Web Browsing Today: The impact of changing contexts on user activity. Ext. Abstracts CHI 2005. Portland, Oregon, ACM Press: 1443-1446.

[17] Hunter, C. D. (2000). Social Impacts: Internet Filter Effectiveness Testing: Over- and Underinclusive Blocking Decisions of Four Popular Web Filters. *Social Science Computer Review* 18(2): 214-222.

[18] Jensen, C., Potts, C. and Jensen, C. (2005). Privacy practices of Internet users: Self-reports versus observed behavior. *International Journal of Human-Computer Studies* 63: 203-227.

[19] Kaasten, S., Greenberg, S. and Edwards, C. (2002). How People Recognize Previously Seen WWW Pages from Titles, URLs and Thumbnails. In *Proc. of Human Computer Interaction 2002*, 247-265.

[20] Kellar, M., Watters, C. and Shepherd, M. (2006). The Impact of Task on the Usage of Web Browser Navigation Tools. In *Proc. of Graphics Interface*, Quebec City, Canada, (to appear).

[21] Lau, T., Etzioni, O. and Weld, D. S. (1999). Privacy Interfaces for Information Management. *Communications of the ACM* 42(10): 89-94.

[22] Lederer, S., Mankoff, J. and Dey, A. K. (2003). Towards a Deconstruction of the Privacy Space. Workshop on Ubicomp Communities: Privacy as Boundary Negotiation, UBICOMP 2003, http://guir.berkeley.edu/pubs/ubicomp2003/privacyworkshop/papers/lederer-privacyspace.pdf

[23] Nie, N. H. and Erbring, L. (2000). Internet and Society: A Preliminary Report, Stanford Institute for the Quantitative Study of Society, http://www.stanford.edu/group/siqss/Press _Release/Preliminary_Report.pdf.

[24] Olson, J. S., Grudin, J. and Horvitz, E. (2005). A Study of Preferences for Sharing and Privacy. Ext. Abstracts of CHI '05. Portland, Oregon, ACM Press: 1985-1988.

[25] Palen, L. and Dourish, P. (2003). Unpacking "Privacy" for a Networked World. In *Proc. of CHI '03*, Ft. Lauderdale, FL, 129-136.

[26] Sellen, A. J., Murphy, R. and Shaw, K. L. (2002). How Knowledge Workers Use the Web. In *Proc. of CHI '02*, Minneapolis, MN, 227-234.

[27] Tauscher, L. and Greenberg, S. (1997). Revisitation patterns in World Wide Web navigation. In *Proc. of CHI '97*, Atlanta, GA.

[28] Weisband, S. P. and Reinig, B. A. (1995). Managing User Perceptions of Email Privacy. *Communications of the ACM* 38(12): 40-47.