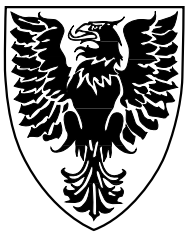

Privacy Gradiante: Understanding Web Browsing Privacy During Ad Hoc Co-Located Collaboration

Kirstie Hawkey
Kori Inkpen

Technical Report CS-2004-18

September 13, 2004

Faculty of Computer Science
6050 University Ave., Halifax, Nova Scotia, B3H 1W5, Canada

Privacy Gradients: Understanding Web Browsing Privacy During Ad Hoc Co-located Collaboration

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

ABSTRACT

This research introduces the concept of privacy issues related to the incidental viewing of traces of previous activities during ad hoc co-located collaboration. Web browsers were used as the representative application in this research, as several of the convenience features record traces of previous web page visits. We introduce a 4-tier privacy gradient to allow study participants to classify privacy levels associated with their actual web browsing over the course of a week-long diary study. Results include analysis of the privacy comfort levels of individuals, their current privacy management strategies, their browsing behaviours, and their use of the privacy gradients. This initial exploratory study provides important insight that will guide the development of a privacy management system.

Author Keywords

Privacy, web browsing, ad hoc collaboration, contexts of use, diary study, user study

INTRODUCTION

Dr. Smith is teaching a Discrete Math class. He is displaying his lecture slides with a projector connected to his laptop. During the class, Dr. Smith decides to show his students a web page that demonstrates the Four Color Problem. He loads a web browser and starts typing 'four color' into his search field in an effort to re-visit the web site. Field auto completion is enabled; and as Dr. Smith begins typing, the previous entries beginning with 'f' and then 'fo' are displayed. 'Filing for bankruptcy' and 'foot fungus' are two of the entries on the list. A collective snicker arises from the class.

Marvin is at work browsing the web at his desk as he eats his lunch. He's unhappy with his job and registers at a couple of job search sites and peruses the on-line 'help wanted' section of his local newspaper. Later, his boss sits down beside him at the computer and wants to look at the competition's web sites with him again. His boss grabs the mouse and opens up the web browser's history files as they had been looking at these sites a couple of days ago. Marvin is uncomfortable and hopes that his boss goes directly to the sites from two days ago and doesn't notice his recent job search browsing.

As computers are used, transactions are generally logged in some fashion creating artifacts of the user's actions [28]. A great deal of information about an individual's past activities on the computer is visible with casual inspection including the file and application icons and names on the desktop, in the start menu, or in the file system itself. Many applications such as web browsers offer 'convenience features' that record past interactions for future reference. For example, the browser history allows easy access to recently visited web sites and field auto complete will reveal search terms and URLs previously entered. Although often beneficial to the user for future interactions, these traces of previous activity may reveal aspects of computer use that the user may prefer to remain private. Also, URLs often include private data required by external servers as part of the query string [24] adding to privacy concerns. It is not always clear to a computer user exactly which artifacts are being created and stored and which can subsequently be viewed by others during normal computer usage [33]. As devices become mobile and used in a variety of settings, it becomes less clear who all the future viewers will be and the context under which the material will be viewed [28].

Many people use their computers for a variety of activities and contexts of use. Employees commonly use their workplace computers for personal use such as email or web browsing. When a computer is designated solely for group use, there is the expectation that others will be subsequently using the computer. Users may therefore be less likely to engage in personal activities or more likely to remove traces of such activity if they are aware of the artifacts that can be created. However, when a workplace device is used at least partially for personal use, there can be an expectation by the user of a certain amount of privacy [33]. However, users may still choose to delay much of their

Dalhousie University, Faculty of Computer Science Technical Report

personal activity until they are in the privacy of their own home. Increasingly, laptop computers accompany individuals between home and their work place. Laptop users typically engage in all their personal activities on the device, a situation that furthers the expectation of privacy. Later, when in a situation where others will view or use their personal computer, there can be a sense of exposure.

There may be many instances where others can view your computer screen. Some times it is the result of a collaborative task, as when people gather in an ad hoc basis around a computer to work on a project. It can also be more demonstrative in nature, as when a lecturer gives a seminar and their laptop connects to a large screen display.

As illustrated in the opening narratives, in many cases the information considered by an individual to be 'private' is not pornography or other illicit information. Although many people do use their computers for viewing and storing such information, much of it is of a more innocent nature. The responses to a survey we recently conducted demonstrate this. When asked to describe a case where he felt uncomfortable with traces of his previous activity being seen, one participant responded, "Spouse once discovered search term through auto complete that I wished to remain confidential (Christmas gift shopping)". Issues of confidentiality also arise with proprietary business information or personal information relating to customers, students, or study participants.

The prevalence of ad hoc co-located collaboration and the use of computers in a variety of contexts and settings combine to make incidental viewing of information a compelling problem. Ordinarily, normative privacy [27] is achieved for computer displays by physically locating the display so that others can't view it [29] or relying on the social norms that preclude others from openly staring at information on a display within someone's 'personal zone' [31]. However, normative privacy is impossible in the case of collaboration around a display, as we are inviting others to look at a particular part of the display and the display becomes an object in the collaboration [31].

Managing the privacy issues relating to all artifacts resulting from previous computer use is a broad problem. Web browsers are used during this research as a representative application for a closer inspection of the privacy dimensions that occur during ad hoc co-located collaboration. Web browsers are often used during collaboration to find information or share previously found sites and are also typically used for information gathering and entertainment of a more personal nature. This research will focus on privacy issues surrounding the recording and subsequent usage of the history, field auto completion, and bookmarks that are typical convenience features in web browsers. Users must currently choose to either turn them off or periodically clear the stored information, either through the web browser's tools or with commercial privacy software, if they want to maintain privacy. Commercial privacy management 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 users would never want to view artifacts of both types concurrently.

Before being able to develop a solution for privacy management, the nature of web browsing activity with respect to privacy concerns must be examined. This research begins this exploratory process. This paper reports our research into the nature of web browsing activity with respect to privacy concerns when others can view the traces of activity. First we will review the related literature in the areas of privacy, personal information management systems, web browser usage, and privacy management tools. The next section introduces the concept of privacy gradients, followed by the methodology of our diary study exploring the privacy patterns inherent during actual web browsing. We will then present the results of our study and, in the discussion section, examine the implications of our results on a privacy management scheme for web browsing. We conclude the paper with conclusions and future directions for this research.

RELATED LITERATURE

Privacy

Previous research has identified that individuals have fundamentally different attitudes towards privacy [1]. Additionally, the context in which the "private" information is viewed can impact the owner's comfort level. There may be different levels of privacy desired depending on the relationship the individual has to potential viewers and on the type of the information [27]. People often present themselves differently depending on their perceived audience [7, 28] and different personas may require different levels of privacy [23]. The amount of control that the individual retains over the disclosure of information may also impact their level of comfort [28]. Privacy concerns increase when displays are viewable by many people in a group and members aren't clear which information is being viewed, by whom, and how often [16].

Online privacy concerns and preferences have been examined in great detail and the Platform for Privacy Preferences Project (P3P) has developed standards [10] that allow users greater control over the use of their personal information at participating web sites. While it is important to note that on-line privacy is not the focus of this research, aspects of users' privacy concerns on-line may be relevant in a co-located setting. Research [1] has found that female users tend to be more concerned about their personal privacy online, that there are differing levels of sensitivity about personal data depending on the content, and that web users value privacy over convenience, preferring to remain in control of information rather than automating transfer of their personal information.

Privacy issues have been raised in co-located CSCW research [6, 14, 17, 29, 30, 34], although they have primarily been limited to the privacy of data within an application or on a specialized device dedicated to collaborative group work. However, this view of private information makes the assumption that all information being viewed is related to the task at hand. In the case of opportunistic collaboration around a personal computer, the information being viewed can also include other unrelated private information.

Personal Information Management Systems

Personal information management systems (PIMS) allow users to manage and integrate a large amount of personal data and transaction records both within [3, 11] and across [5, 20] applications. Users often find the management of artifacts (files, email and bookmarks) to be a burden due to their volume and employ different strategies for managing their artifacts [2, 5, 13] based on frequency of management and levels of folder organization. Artifacts can be organized by categories [3], collections [20], or user context [11]. The management strategies employed in PIMS may be applicable when privacy is the attribute that dictates membership in a category or collection.

PIMS that allow end users to provide and receive contextual information about each other [21] often have elaborate privacy management mechanisms. These are similar to those found in distributed and mixed-presence CSCW applications [4, 12, 16] that allow users to maintain privacy while displaying awareness information necessary for effective collaboration. Strategies for managing privacy include storing and presenting aggregate data where possible [4], adjusting the level of detail of information depending on the size and public nature of the display [16], and providing social privacy contracts [21].

Web Browser Usage

Current web browser convenience feature settings are hard to understand and manage [26] and often under-utilized as a result. Studies of web browsing behaviour [9, 25, 32] show that there is a high rate of web site visits that are re-visitations (60-80%); that a small number of web sites (2-3) account for the majority of re-visitations, with about 60% of pages only being visited once. However, the results of these older studies may need reevaluation against current contexts of use (e.g. increased quality and speed of internet connections, increased time people spend using their web browsers, web browser improvements). Efforts to develop and evaluate better convenience features such as the back button [25], history [18], and integrated re-visitation tools [9, 19] through visualization and different organizational models are well documented.

Privacy Management Tools

Privacy management is a difficult problem due to the diverse privacy concerns of users and the large number of potential viewers and types of information to be protected [7]. Lau et al. [22] examined classifying pages in a sharable web history application with both extensional and intensional classification schemes. Explicit classification of each piece of information with a privacy type (extensional) is easy for the user to understand as the privacy is applied as a property of each item; but classification can be unmanageable due to the large volume of items. A rule-based system (intensional) that applies a privacy level according to a set of rules, results in less work for the user, but it may be very difficult to understand. The authors [22] state that privacy interfaces should make it easy to create, inspect, modify and monitor privacy policies and that the policies should be applied proactively to objects as they are encountered. Hoccheiser's principles for privacy protection software [15] include simplicity; using privacy as the default level; and having no performance, utility, or usability penalties for privacy protection.

PRIVACY GRADIENTS

We believe there may be a variety of types of web sites that people do not necessarily want others to see traces of, for a variety of reasons. These web sites may not fit precisely into 'Public' or 'Private' categories. Whether it is a weight-loss support forum, a job search site, or an adult site, there are certain aspects of our web browsing that we may not feel comfortable sharing with all people, but would like to be able to share with some people. Our relationship to the potential viewer of this information plays a role in the level of privacy required. What may be appropriate to allow a close friend to see may be inappropriate if viewed by a supervisor or a client.

In order to enable classification of visited websites, we require a common terminology. We introduce a four-tier privacy gradient scheme that partitions web sites: *Public*, *Semi-Public*, *Private*, and *Don't Save* (see Figure 1). If a site is something that you would like to access again, you would want traces of it to appear in your browser convenience features. These traces should be stored with some associated privacy level. *Public* sites are those that you are comfortable with anybody and everybody viewing, including the Queen of England (hence the crown in Figure 1). *Private* sites are those that you would be comfortable with only yourself and possibly a couple of close confidants or a spouse viewing, people with whom you share just about everything. *Semi-public* sites fall somewhere in between. Depending on the context of the viewing, the pages would be considered to be public or private. The example given to participants of semi-public sites was the scenario of browsing for a new job and then having your boss view your web browser as you work on a task together. Traces of this browsing might be considered private if your boss is the viewer; but if a close friend is the

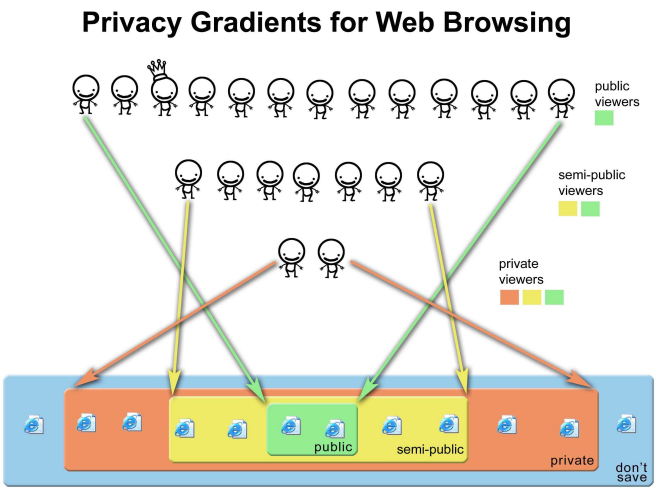

Figure 1, Privacy Gradient diagram that participants used as a guide for classifying categories of web sites and potential viewers.

viewer, it would probably be ok for them to see. Web sites classified as *Don't Save* primarily fall into one of two categories: ones that are irrelevant (i.e. the first 17 pages of a search before you found something pertinent) or ones that are very private that you do not want to have record of visiting on your computer at all.

When managing privacy of traces of web browsing activity, there are two main issues: classifying web pages and other artifacts with a privacy level and displaying the appropriate content when your display is visible by others. This research examined actual patterns of web browsing activity with respect to privacy in an effort to find patterns that may provide guidelines for a privacy management scheme to facilitate easy and effective classification of viewed web pages.

For the purposes of displaying appropriate content when there are others able to view traces of your previous browsing activity, we envision a scheme whereby you could set a browser window as being either public, semi-public, or private. The arrows in Figure 1 illustrate which artifacts would be visible in a browser window set at a specific privacy level. The only URLs, histories, auto-completions, etc. available for viewing in a public window would be those classified as public. If the window is semi-public, both the public and semi-public artifacts would be visible. If the window is private, artifacts from all previously visited sites that have been saved would be visible. We believe that the privacy level of the browser window could be used to tag new sites visited in that window, an approach similar to the extensional classification described by Lau et al.[22]; however, such a scheme would require an integration with a more proactive approach in order to be manageable.

METHOD

Obviously, privacy is a very complex issue with both privacy concerns and willingness to maintain a privacy management scheme varying on an individual basis. However, our hypothesis was that people would be willing to organize their information across a small number of privacy levels or gradients. We introduced a 4-tier privacy gradient to see if that level of granularity was appropriate to reflect the privacy needs between types of web sites and potential viewing audience. We were also interested in how the level of control at the computer and the relationship to the viewer impacts the choice of privacy level. It was important to explore normal web browsing activities to see if patterns exist that would make organization within privacy gradients easier. For example, do people use different browser windows for activities of different privacy types? Do they tend to have sequences of one type of activity or another either within a browser window or over a given time period?

Participants and Setting

We recruited participants from the general university community. Twenty participants took part in the study: 16 males and 4 females. Participants ranged in age from 19 to 47 with a mean age of 26. Participants were highly educated (the minimum education level was some university with 13/20 having completed an undergraduate degree and 5/20 having completed a graduate degree as well) in primarily technical fields (14 Computer Science, 4 Science, 1 Arts, 1 undeclared). Eighteen of the participants were students, one was a professor, and one an Information Technology professional. Participants were generally experienced computer users (10 years' computer use) and spent a considerable amount of time each week using their computer (29-35 hours per week) and using a web browser (22-28 hours per week). Participants browsed the web an average of 48% for personal purposes, 16% for work-related purposes, and 36% for educational purposes.

The study took place in August 2004. We chose to conduct a week-long diary study to elicit the normal web browsing behaviors of participants as much as possible. To qualify for inclusion in the study, participants needed to perform the majority of their web browsing on a laptop computer so that we could capture the full picture of their personal and work/school related web browsing. They also needed to have occasions where their web browser window was visible by others, so that the concept of privacy in this situation had some relevance. Participants also needed to be willing to use Microsoft's Internet Explorer (IE) as their default web browser for one week.

Study Instruments

To record the browsing activity of participants, we built a browser helper object (BHO) that worked with Internet Explorer. As each IE window opened, the BHO was automatically loaded and logged the actual web sites visited by the participants. The visited web page (URL and page

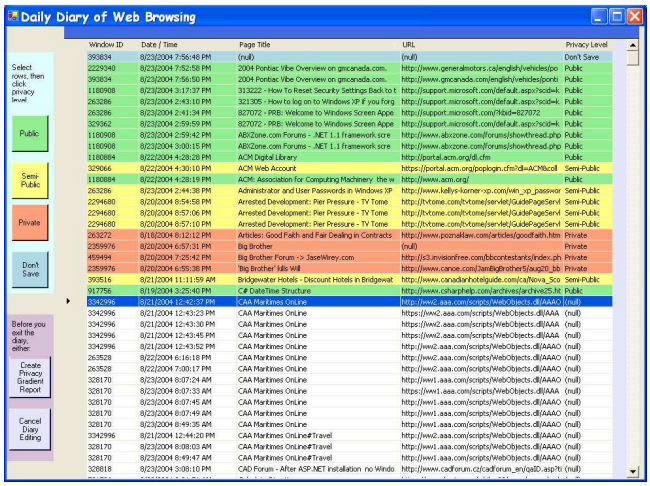

Figure 2. Screen shot of the electronic diary application used by participants to classify the web sites they had visited.

title) was recorded, along with a time stamp and the ID number of the browser window in which the page was viewed.

We also developed an electronic diary (Figure 2) to allow participants to assign privacy gradients to their web browsing. The diary displayed the logged data and allowed participants to indicate how they would classify the privacy level of each web page they visited if others were to view the history of this activity later. The data could be sorted by browser window, date and time, page title, URL, or privacy level assigned. Participants could select single rows or multiple rows using shift or ctrl keys and then classify the privacy level using one of the privacy buttons. The privacy level was updated for all selected records and the row was coloured with the same colour as the privacy button: green for *Public*, yellow for *Semi-Public*, red for *Private*, and blue for *Don't Save*.

After classification, participants could generate a report to email to the researchers. In this report, the viewing history was sanitized so that the URL and page title were eliminated. We anticipated that the privacy afforded by the sanitized browsing record would contribute to participants' willingness to engage in web browsing patterns that are similar to their normal actions. We were not interested in which sites participants classified in the various privacy levels, just in the patterns of gradient application. Participants were able to inspect the data before it was sent, but could not change it.

In addition to the diary portion of the study, participants completed pre and post study questionnaires to tease out some additional information that may be useful when trying to minimize the effort of maintaining a privacy management system. A demographic questionnaire was followed by a privacy background questionnaire to elicit information on privacy comfort levels and current privacy management practices. A 7-point Likert privacy comfort level scale (1: extremely uncomfortable, 4: neutral, 7: extremely comfortable) allowed participants to indicate how comfortable different situations made them feel in terms of privacy.

We developed two classification tasks so that participants could assign privacy levels to categories of web sites and types of viewers using the privacy gradient scheme. The web site categories and their descriptions [8] used in the web site classification task are commercially used as a means of filtering and blocking objectionable or non-productive internet content in a variety of internet management applications used by companies and government agencies. We hypothesized that it may be possible to use such categorizations as the basis for default privacy settings.

Procedure

After screening interested participants via email to ensure they were eligible for inclusion in the study, an initial installation session was scheduled at the Faculty of Computer Science. This session lasted approximately 30-60 minutes, depending on the complexity of the installation. As participants completed the questionnaires (demographic, privacy background, viewer classification task, web category classification task), the logging BHO and electronic diary were installed on the participant's laptop. Before beginning, we backed up the participant's registry and installed the Microsoft .NET framework, if necessary. After the installation, the researcher browsed a few web pages to verify that the logging software was working and to demonstrate to the participant the functionality of the electronic diary.

The participants then went about their normal activities for the week. We asked that participants complete the diary daily if possible. After one week, participants returned for the final session to complete post session questionnaires (privacy gradient evaluation, privacy background, viewer classification task, web category classification task) and to have the program uninstalled. Before uninstalling the software, it was confirmed that all data had been sent and received. Backup files stored on the participant's laptop allowed retrieval of information that had not been received. This session took 20-30 minutes. It should be noted that five participants had difficulties during the week with their software, their hardware, or their internet connection. These participants did complete 7 days worth of diary entries before returning for the uninstall session, but they were not necessarily consecutive.

RESULTS

Privacy Comfort Levels of Individuals

Questionnaire data were analyzed using Friedman Two-Way ANOVA of Ranks tests. Wilcoxon signed ranks tests were used to perform pair wise post hoc analyses using a Bonferroni adjusted p value.

The privacy background questionnaire examined participants' comfort levels related to privacy, depending

on who could view their computer display or had access to their computer. Ten types of possible viewers were explored (*parents, spouse/significant other, close friends, clients, supervisor, acquaintances, colleagues/fellow students, employee/students that you supervise, technical support staff*, and *audience at a presentation).* Analyses of these results revealed that viewer data could be clustered into three distinct categories: *spouse/significant other*, *close friend*, and *other contacts*. The 'other contacts' category included five viewer types (*acquaintances, supervisor, colleagues/fellow students, audience at a presentation, technical support staff),* given that no significant differences were found in the privacy comfort levels assigned to these types. The remaining three types (*parents*, *clients, employee/students that you supervise)* were omitted due to high variance or insufficient data points.

Using the viewer categories defined above, our results also revealed that people had different privacy comfort levels related depending on who could view their computer display. People reported that they were most comfortable with a *spouse* viewing their display, followed-by *close friends*, and then *other contacts*. All of these pair wise differences were statistically significant at the $p<.016$ level. This same trend was also found for the case where these categories of people would have access to use the participant's computer ($p<.016$). In addition, participants reported being less comfortable in terms of privacy when a person could use their computer instead of just viewing the screen, however, this difference was only significant for the *other contacts* category ($p<.05$).

Current Privacy Management Strategies

Given that browsers do offer some privacy management, we were interested in how the participants in our study currently manage their privacy related to web browser convenience features.

History

Nineteen of the twenty participants indicated that the history feature is enabled on their computer. Nine indicated that they use the default setting while the remaining ten set this feature for a particular number of days.

AutoComplete

Two participants indicated they do not use the auto complete feature; two indicated that they were unsure how they currently had this feature set in their browser; and five indicated that they use the default setting. Eleven participants indicated that they use the auto complete for web addresses; four use it for fields and forms; and four use it for usernames and passwords.

Favorites/Bookmarks

One person indicated that he does not use favorites. Thirteen indicated that they use favorites to save web addresses, using default or accurate names while the remaining six use it to save web addresses but rename some of the names to conceal the page's content.

Explicit measures taken to manage privacy

Participants indicated what privacy management actions they would take if they had advance warning that someone would be working closely with them as they used their web browser. Participants were asked to select all options that applied. One participant indicated that he would take no action. Nine participants indicated that they would chose to retain control of the keyboard and mouse and limit the functionality they would use. Of these nine, six also indicated that they would also take other actions (such as modifying their favorites, history, or auto completes).

In terms of the favorites feature, ten participants indicated that they would check their favorites and remove any inappropriate web pages, while one other indicated that he would rename any inappropriate web pages. Of these people, two additionally indicated that they may choose to erase all of their favorites.

In terms of the history feature, thirteen participants indicated that they would potentially modify their history records. Nine indicated that they would check the history and clear it if there were any inappropriate entries while four indicated that they would erase all history records.

In terms of the auto complete feature, thirteen participants indicated that they would potentially modify their auto complete data. Six indicated that they would clear it if they thought it contained any inappropriate entries, ten indicated that they would clear all passwords from their auto complete history and seven indicated that they would clear all forms from their auto complete history.

Browsing Behaviours

The browsing behaviours our participants exhibited were highly variable, both within a person's browsing and across individuals.

Number of Pages Visited

On average, the participants in our study visited 1808 pages over the course of the week (~258/day). However, the volume of web page visits was highly variable. Across participants, the number of web page visits ranged from 422 pages (~60/day) to 5127 pages (~732/day). This is a dramatic increase from previous reports of 42 web page visits per day (1999/2000) [9] and 21 web page visits per day [32].

Browser Window Usage

Given the ability to run multiple browser windows at a time, different browser behaviours can be gleaned from the number of pages viewed in each browser window. Overall, the participants in our study opened an average of 289 different browser windows over the course of the week. Again, this result was highly variable. Across participants,

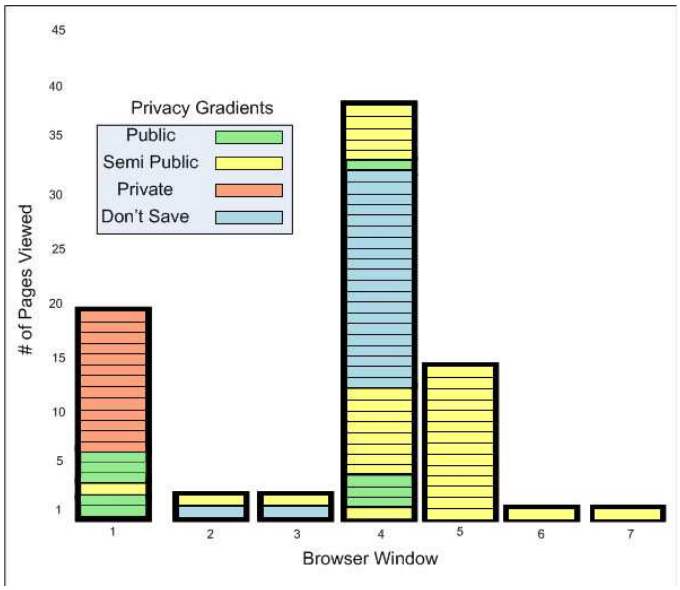

Figure 3. Example of sequential patterns of privacy gradients on a per browser window basis.

the number of different browser windows opened ranged from 47 to 799.

Figure 3 and Figure 4 show the actual privacy patterns of browsing for one participant for one hour. This participant opened 7 windows during this hour and visited a total of 81 pages during the hour. This example will be referred to throughout these results to illustrate some of the patterns observed.

Despite the fact that many different windows were opened, in most cases, only one or two pages were viewed within a browser window (for 17 participants the number of pages opened in a window had a mode of 2, the remaining three participants had a mode of 1). This is not surprising given the number of windows that get automatically spawned for a specific purpose while browsing. However, there were also several instances where large numbers of pages were viewed within one browser window. The average maximum number of pages viewed in one window was 108 across the participants (from 27 to 255).

The results from our study reveal that people frequently move back and forth between open browser windows. Analyzing the number of switches between browser windows we find that participants ranged anywhere from 22 window re-visitations to 430 re-visitations (with an average of 158). Figure 4 shows 3 browser window revisitations.

Speed of Browsing

Many participants exhibited rapid bursts of browsing with more than 10 pages being loaded every minute. We define a burst to be a rapid sequence of web visits with less than one minute's elapsed time between web pages loading. Overall, the average duration of a burst was 82 seconds. The average page length of a burst was 7 pages, with maximum bursts of up to 172 pages. The average seconds per page

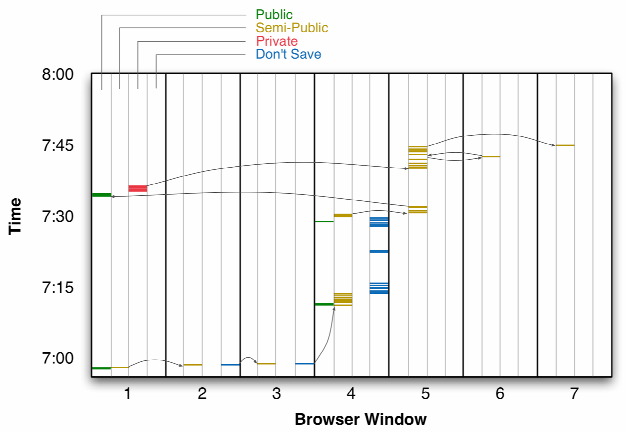

Figure 4. Example of temporal patterns of privacy gradients on a per window basis.

during a burst was 12. Figure 4 shows several examples of bursts. One burst (browser window 4) runs from 7:28:48pm-7:30:31pm with 16 pages opened in 104 seconds.

Privacy Gradients

In this study we introduced the notion of privacy gradients with four different levels (public, semi-public, private, and don't save). As we discuss the various patterns observed related to classification using the privacy gradients, it is important to recognize that this was a field study run over a one-week period. Therefore, different participants visited and classified different sets of web pages (whatever pages they happened to have visited during that week). As such, just because two people exhibited similar behaviours, does not necessarily mean that they have similar privacy perspectives. Despite this, it is still valuable to examine the way people applied the privacy gradients to their normal web browsing behaviour. These patterns reflect the perceived need for privacy based on the sites that an individual visits. It also provides real data concerning individuals' use of the privacy gradients.

Utilization of Gradients

All participants utilized all of the categories when classifying their visited web pages (with the exception of one participant who never used the *don't save* category). Overall, 15140 pages were classified as being public, 9083 pages were classified as being semi-public, 5543 pages were classified as being private, and 6404 pages were classified as ones that the participants didn't want to save.

For each participant, we computed the percentage of visited sites that were classified into each privacy gradient. A K-means cluster analysis grouped the participants into four clusters based on the relative proportions of sites classified into each of the privacy gradients. The results of this clustering are shown in Table 1. Examining the cluster means shows that each of the four clusters represents groups of individuals who had a relatively high proportion

Clusters		C1	C2	C3	C4
Privacy Gradient	**Overall**	**Final Cluster Centers**			
Public	42%	22%	36%	62%	18%
Semi-Public	25%	58%	21%	16%	28%
Private	15%	9%	36%	11%	9%
Don't Save	18%	11%	7%	11%	46%
Number of Participants		3	5	10	2

Table 1. Results of cluster analysis of Privacy Gradient use.

of web browsing in one of the privacy gradients (C1-*semi-public*; C2-*private*; C3-*public*; C4-*don't save*).

The fact that ten participants were clustered in C3 suggests that this privacy breakdown may be fairly representative of general browsing behaviour (~60% public, with the remaining categories being roughly equal). Even for those participants with a relatively high proportion of private sites (C2), there were still only 36% of sites considered private.

Streaks

We define a streak to be two or more consecutive web pages of a given privacy gradient, within a browser window. For example, in Figure 3, four streaks occurred in browser window #4: there was a single *semi-public* page, followed by a streak of three *public* pages, a streak of eight *semi-public* pages, a streak of twenty *don't save* pages, a single *public* page, and, finally, a streak of six *semi-public* pages.

Detailed analyses of the number and duration of streaks revealed that 85% of all pages visited occurred within a streak and the average length of a streak was 6.5 pages. In some instances, the length of a streak was quite long, ranging up to 166 pages. A repeated measures ANOVA revealed a significant difference between the average length of a streak depending on the four privacy levels, $F_{3,57}=4.11$, $p=.025$, $\eta^2=.178$. Given that the sphericity assumption was violated, a Huynh-Feldt correction was used, and the corrected degrees of freedom and significance levels are reported. The *Don't Save* gradient was a loosely defined category that could have been interpreted in many ways. As such we chose to analyze the data excluding the *Don't Save* gradient. This analysis revealed no significant difference between the average streak length of the remaining three gradients (public, semi-public, private) ($F_{2,38}=1.14$, $p=.316$, $\eta^2=.057$). Again the Huynh-Feldt correction was applied.

Transitions

We define a transition to be a switch between privacy gradients within a browser window. For example, in Figure 3, five transitions occurred in browser window #4. Detailed analyses of the number of transitions revealed that participants exhibited an average of 214 transitions over the course of the week. In addition, we found that 56% of browser windows contained no transitions, and on average, participants had 0.9 transitions per browser window.

Strictly looking at the number of transitions in a browser window may be misleading. If there were 5 transitions and only 11 pages visited, this meant that the user transitioned between privacy gradients very frequently (possibly after every second page). However, if there were 5 transitions and 250 pages visited, this number of transitions may seem more reasonable. To account for this we computed a normalized transition count, dividing the total number of transitions by the number of pages in a window. This gave us a numerical score between 0 and 1 where low values indicated low transition rates while high valued indicated rapid transitions. The results of this analysis revealed that participants on average had a normalized transition score of 0.14.

Viewer Classification

If people were to use privacy gradients to filter their content, it is also important that they associate potential viewers with these gradients (i.e. types of people that fit into the different privacy gradients). Results from the viewer classification task revealed that most people classified their spouse/significant other as a private viewer (13/19) and their close friend as a semi-public viewer (13/20). The results for parents were highly variable ten participants classifying them as semi-public, seven classifying them as public, while three classified them as a private viewer. Acquaintances, technical support staff, and colleagues/fellow students were classified either as being public viewers (12-16/20) or semi-public viewers (4-8/20). The remaining types of people (*supervisor, employees/students that you supervise, audience at a presentation and clients*) were classified as public viewers by most participants (17-20/20).

Web site Category Classification

The data gathered from our participants' electronic diary entries classified web pages that our participants chose to visit. However, this does not give us a sense of the similarity of perceptions of the privacy level of certain categories of web sites. In addition, because the diary entries were sanitized (URL and page title were removed), we have no record of which pages were categorized into which privacy gradient. Instead, we used a web site classification task to gain a sense of how different participants would apply the privacy gradients. We found some consistency across participants in terms of their classification of web site categories. Of the 52 categories, 5 categories were classified as *don't save* (e.g. violence/hate/racism; web advertisements); 5 were classified as being *private* (e.g. adult/mature content; financial); 11 were classified somewhere between *public* and *private* (e.g. newsgroups; shopping); 17 were classified as being *public* (e.g. arts/entertainment, search engines). None of the remaining 14 categories had any consensus across the participants.

Goodness of Fit

After working with the privacy gradients for a week, most of our participants (15/20) indicated that they felt the privacy gradients fit 'most of the time'. Of the remaining five participants, three felt that the gradients fit 'all of the time' and two felt that they gradients fit 'some of the time'. When asked if there were any web sites that didn't fit into the gradients, several participants (8/20) reported that there were sites that they felt didn't fit and estimated that approximately 15% of sites were difficult to classify. In most cases, this difficulty was because of sites that had multiple purposes or variable content (i.e. newspaper sites – it would depend on the article).

DISCUSSION

Complexity of This Problem

The results from our study clearly demonstrate that any privacy management approach is complicated by browsing behaviours. First, the sheer number of pages that people visit while browsing, means that any manual solution will be overly arduous and therefore impractical. Beyond just the number of pages visited, the speed with which users browsed was staggering.

The results from our study indicate that people's behaviours vary considerably in terms of the number of pages they visit, number of separate windows they use, when they choose to perform private browsing, and how they classify individual pages in terms of privacy. This high variability will make it difficult to arrive at a standard solution for privacy management.

While there was some agreement in the classification of categories of web pages, it was not consistent. As such, privacy filtering using pre-existing web categories would not be effective. However, the web page classifications may be a tool that could assist with personal profiling. Personalized privacy management schemes may help alleviate some of the burden from the user by applying default categories according to their privacy attitudes.

Guidelines for Solutions

The results from our study show that there are indeed patterns of privacy associated with web browsing that may help simplify a privacy management solution. However, there is a high amount of variability both across users and within the browsing of a single user. This variability indicates that potential solutions must be sensitive to the changing needs of the users and allow flexibility in the way they handle their changing privacy needs. The high volume of web sites visited and the rapid browsing indicate the need for seamless setting of levels.

CONCLUSION AND FUTURE WORK

We are just beginning to touch the problem, but this was an important first step. This study advanced the understanding of privacy issues relating to ad-hoc co-located collaboration around a computer that contains web browsing artifacts that are unrelated to the task at hand and may be of a private nature. In particular, we advanced the understanding of patterns of web browsing activity with respect to privacy and the mapping users have between their current browsing activity and their subsequent privacy needs. This is a preliminary study; its results will guide the development of a technological solution that will then require evaluation for its usability and effectiveness at managing users' privacy requirements.

ACKNOWLEDGMENTS

Thanks to Melanie Kellar for co-creating the logging software; Robert Hawkey and Chris Power for scripting assistance during analysis; Colin Bate, Vicki Ha, and Malcolm Rodgers for diagrams; and the members of the EDGE Lab for their continuous support. This research was funded in part by NSERC.

REFERENCES

1. Ackerman, M., Cranor, L., and Reagle, J. (1999). Privacy in E-Commerce: Examining User Scenarios and Privacy Preferences. In *Proceedings of ACM Conference on Electronic Commerce*, Denver, CO. 1-8.
2. Balter, O. (2000). Keystroke Level Analysis of Email Message Organization. In *Proceedings of CHI 2000*, The Hague, The Netherlands. 105-112.
3. Balter, O. and Sidner, C. L. (2002). Biforst Inbox Organizer: Giving Users Control over the Inbox. In *Proceedings of NordCHI 2002*, Arhus, Denmark. 111-118.
4. Begole, J., Tang, J. C., and Hill, R. (2003). Rhythm Modeling, Visualizations and Applications. In *Proceedings of UIST 2003*, Vancouver, Canada. 11-20.
5. Boardman, R. and Sasse, M. A. (204). "Stuff Goes into the Computer and Doesn't Come out" a Cross-Tool Study of Personal Information Management. In *Proceedings of CHI 2004*, Vienna, Austria. 583-590.
6. Brown, L. D., Hua, H., and Chunyu, G. (2003). A Widget Framework for Augmented Interaction in Scape. In *Proceedings of UIST 2003*, Vancouver, Canada. 1-10.
7. Cadiz, J. and Gupta, A. (2001). *Privacy Interfaces for Collaboration* (No. MSR-TR-2001-82). Redmond, WA. Microsoft Research.
8. Cerberian Web Filter Categories. http://www.webrootdisp.net/audit/rating-descriptions.htm.
9. Cockburn, A., Greenberg, S., Jones, S., McKenzie, B., and Moyle, M. (2003). Improving Web Page Revisitation: Analysis, Design and Evaluation. *IT & Society, 1*(3): 159-183.
10. Cranor, L., Langheinrich, M., Marchiori, M., Presler-Marshall, M., and Reagle, J. (2002). *The Platform for Privacy Preferences 1.0 (P3p1.0) Specification*. Retrieved September 11, 2004, from http://www.w3.org/TR/P3P/

11. Ducheneaut, N. and Bellotti, V. (2001). E-Mail as Habitat, an Exploration of Embedded Personal Information Management. *Interactions, 8*(5): 30-38.
12. Greenberg, S. (2001). The Notification Collage: Posting Information to Public and Personal Displays. In *Proceedings of CHI 2001*, Seattle, WA. 514-521.
13. Gwizdka, J. (2004). Email Task Management Styles: The Cleaners and the Keepers. In *Proceedings of CHI 2004*, Vienna, Austria. 1235-1238.
14. Hinckley, K. (2003). Distributed and Local Sensing Techniques for Face-to-Face Collaboration. In *Proceedings of the 2003 Int. Conf. on Multimodal Interfaces*, Vancouver, Canada. 81-84.
15. Hochheiser, H. (2000). Principles for Privacy Protection Software. In *Proceedings of Computers, Freedom and Privacy*, Toronto, Canada. 69-72.
16. Huang, E. M. and Mynatt, E. D. (2003). Semi-Public Displays for Small, Co-Located Groups. In *Proceedings of CHI 2003*, Ft. Lauderdale, FL. 49-56.
17. Izadi, S., Brignuli, H., Rodden, T., Rogers, Y., and Underwood, M. (2003). Dynamo: A Public Interactive Surface Supporting the Cooperative Sharing and Exchange of Media. In *Proceedings of UIST 2003*, Vancouver, Canada. 159-168.
18. JasonSmith, M. and Cockburn, A. (2002). Get a Way Back: Evaluating Retrieval from History Lists. In *Proceedings of Fourth Australasian User Interface Conference*, Adelaide, Australia. 33-38.
19. Kaasten, S. and Greenberg, S. (2001). Integrating Back, History and Bookmarks in Web Browsers. In *Proceedings of CHI 2001*, Seattle, WA. 379-380.
20. Karger, D. R. and Quan, D. (2004). Collections: Flexible, Essential Tools for Information Management. In *Proceedings of CHI 2004*, Vienna, Austria. 1159-1162.
21. Kaufman, J., Ruvolo, J., and Ford, D. (2001). Tempus Fugit and the Need for an E-Social Contract. In *Proceedings of Agent Supported Cooperative Work*, Montreal, Canada. 77-84.
22. Lau, T., Etzioni, O., and Weld, D. S. (1999). Privacy Interfaces for Information Management. *Communications of the ACM, 42*(10): 89-94.
23. Lin, D. and Loui, M. C. (1998). Taking the Byte out of Cookies: Privacy, Consent, and the Web. *SIGCAS Computers and Society, 28*(2): 39-51.
24. Martin, D. M., Jr. , Smith, R. M., Brittain, M., Fetch, I., and Wu, H. (2001). The Privacy Practices of Web Browser Extensions. *Communications of the ACM, 44*(2): 45-50.
25. Milic-Frayling, N., Jones, R., Rodden, K., Smyth, G., Blackwell, A., and Sommerer, R. (2004). Smartback: Supporting Users in Back Navigation. In *Proceedings of WWW 2004*, New York, NY. 63-71.
26. Millett, L. I., Friedman, B., and Felten, E. (2001). Cookies and Web Browser Design: Toward Realizing Informed Consent Online. In *Proceedings of CHI 2001*, Seattle, WA. 46-52.
27. Moor, J. H. (1997). Towards a Theory of Privacy in the Information Age. *ACM SIGCAS Computers and Society, 27*(3): 27-32.
28. Palen, L. and Dourish, P. (2003). Unpacking "Privacy" for a Networked World. In *Proceedings of CHI 2003*, Ft. Lauderdale, FL. 129-136.
29. Rodden, T., Rogers, Y., Halloran, J., and Taylor, I. (2003). Designing Novel Interactional Workspaces to Support Face to Face Consultations. In *Proceedings of CHI 2003*, Ft. Lauderdale, FL. 57-64.
30. Shoemaker, G. B. D. and Inkpen, K. M. (2001). Single Display Privacyware: Augmenting Public Displays with Private Information. In *Proceedings of CHI 2001*, Seattle, WA. 522-529.
31. Tan, D. S. and Czerwinski, M. (2003). Information Voyeurism: Social Impact of Physically Large Displays on Information Privacy. In *Proceedings of CHI 2003*, Ft. Lauderdale, FL. 748-749.
32. Tauscher, L. and Greenberg, S. (1997). Revisitation Patterns in World Wide Web Navigation. In *Proceedings of CHI '97*, Atlanta, GA.
33. Weisband, S. P. and Reinig, B. A. (1995). Managing User Perceptions of Email Privacy. *Communications of the ACM, 38*(12): 40-47.
34. Wu, M. and Balakrishnan, R. (2003). Multi-Finger and Whole Hand Gestural Interaction Techniques for Multi-User Tabletop Displays. In *Proceedings of UIST 2003*, Vancouver, Canada. 193-202.