

Privacy Gradients: Exploring Ways to Manage Incidental Information During Co-located Collaboration

Kirstie Hawkey and Kori M. Inkpen

Faculty of Computer Science, Dalhousie University

Halifax, NS B3H 1W5

{hawkey, inkpen}@cs.dal.ca}

ABSTRACT

This research introduces privacy issues related to the viewing of incidental information during co-located collaboration. Web browsers were the representative application used in this research as they have several convenience features that record and display traces of previous web page visits. A one-week field study examined how individuals perceive privacy needs relating to the later incidental viewing of traces of their browsing activity. Participants used a 4-tier privacy gradient to classify the privacy of their actual web browsing. The results revealed per window patterns of privacy during browsing with streaks at given privacy levels and relatively few transitions between levels. Management of incidental information is a complex problem due to multiple viewing contexts, individual differences, and the large volume of information. These privacy patterns suggest that a semi-automated approach to privacy management may be feasible.

Author Keywords

Privacy; web browsing; client-side logging; ad hoc collaboration; contexts of use; field study

ACM Classification Keywords

H.5.3 [Information Interfaces and Presentation] Group and Organization Interfaces: Collaborative Computing; Web-based Interaction

INTRODUCTION

As computers are used, transactions are generally logged in some manner creating artifacts of the user's actions [6]. A great deal of incidental information about an individual's past activities on the computer is visible with casual inspection including file and application icons and names on the desktop, in the start menu, or in the file system itself. Many a presenter has felt uneasy when a technical problem occurs during their presentation, requiring them to interact with their computer in full view of the audience.

Unless sharing a group machine, we generally have the notion that our computing is personal. The terminology used in Windows reinforces this perception: My Computer, My Documents, My Network. However, there are many

instances where others can view your computer screen. In some instances, this viewing is invited, such as when people gather in an ad hoc basis around a computer to collaborate on a project or a professor projects her display during a lecture. As computing devices become mobile and used in a variety of settings, it becomes less clear whom all the future viewers will be and the context of viewing [6].

The prevalence of ad hoc co-located collaboration and the use of computers in a variety of contexts combine to make incidental viewing of information a compelling problem. Ordinarily, normative privacy [5] is achieved for computer displays by physically locating the display so that others cannot view it or relying on the social norms that preclude others from openly staring at information on a display within someone's 'personal zone' [7]. 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 itself becomes an object in the collaboration.

There are tradeoffs to consider when managing the privacy of incidental information. People are efficient when they work in a familiar environment and have access to their usual convenience features and layout. However, when using this environment collaboratively, incidental information relating to previous activity may be inappropriately exposed. If we want to protect the privacy of this information, it may be awkward for users to interact without their normal computing environment. We must balance the amount of control a person has over what displays in their environment with the time and effort that is necessary to provide that control.

Our goal is to provide users with tools to manage incidental information and only reveal information that is appropriate for the current context. Privacy management is a difficult problem due to the diverse privacy concerns of users [1] and the large number of potential viewers and types of information to be protected [2]. 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 [5]. The amount of control that the individual retains over the disclosure of information may also impact their level of comfort [6]. It is important to note that incidental information considered 'private' is not only 'non-work/personal' information or illicit information such as

pornography. It may just not be appropriate for the current viewing context. Issues of confidentiality can also arise with proprietary or confidential business information.

We use web browsers in a variety of contexts, often collaboratively. They have many convenience features that assist our browsing activities: the browser history allows easy access to recently visited web sites, auto complete will reveal search terms and URLs previously entered. If viewed by others, this incidental information may reveal aspects of computer use that the user may prefer to remain private. For example, a prior search for ‘foot fungus’ may be revealed when later searching for ‘four color problem’ during a lecture. To maintain privacy, users must currently choose to either turn these features off or periodically clear the stored information, with either the web browser’s tools or commercial privacy software. Commercial 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 never concurrently view sites of both types.

Before being able to develop a privacy management solution, we must examine the nature of web browsing activity with respect to privacy. This paper reports an exploratory field study we conducted examining privacy patterns inherent during web browsing. We first describe the field study, including the privacy gradient scheme used by participants. We then present and discuss the results of the study and finish with conclusions and future work.

FIELD STUDY

Obviously, privacy is a complex issue with both privacy concerns and willingness to maintain a 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 proposed 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. It was important to explore normal web browsing activities to see if patterns exist that would make organization within privacy gradients easier.

Method

The study took place in August 2004. We chose to conduct a field study over the course of a week to elicit normal web browsing behavior as much as possible. To qualify for inclusion, 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 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. Full methodological details are available in [3].

Privacy Gradients

To facilitate classification of visited websites, a common terminology was required. A four-tier privacy gradient

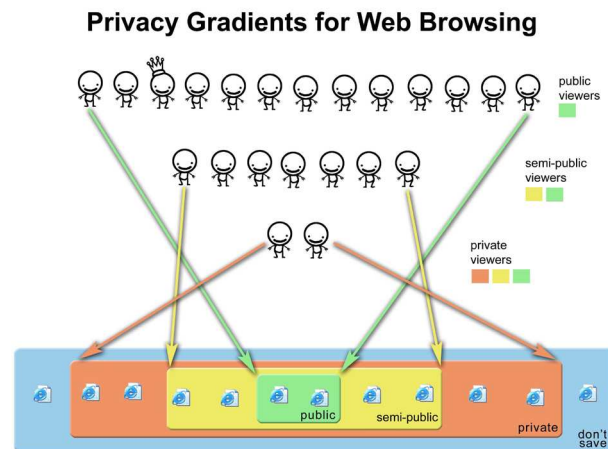


Figure 1. Diagram that participants used as a guide for classifying categories of web sites and potential viewers.

scheme was utilized to partition web sites: *public*, *semi-public*, *private*, and *don't save* (see Figure 1). If a site needs to be accessed again, traces of it should appear in the browser convenience features. These traces should be stored with some associated privacy level. *Public* sites are those someone is comfortable with anybody and everybody viewing, including the Queen of England (hence the crown in Figure 1). *Private* sites are those a person would be comfortable with only themselves and possibly a couple of close confidants viewing. *Semi-public* sites fall somewhere in between: depending on the context of the viewing, 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. the first 17 pages of a search before finding a page) or ones that are so private it is preferred that there is no record of having visited them at all.

Study Instruments

We developed client-side logging software to record the web browsing of participants over the course of the week including visited web page (URL and page title), time stamp, and ID number of the browser window in which the page loaded. We also developed an electronic diary to allow participants to assign privacy gradients to their web browsing on a daily basis. The diary displayed all the logged data and required 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. Participants could annotate individual entries with a privacy level or select multiple entries for annotation. The entries could be sorted by any field (time, url, page title), allowing participants to more easily classify groups of page visits (e.g. repeated visits to the same site). After classification, participants generated 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 hoped that the privacy afforded by the sanitized browsing record would contribute to participants' willingness to engage in

web browsing patterns that were similar to their normal actions. In addition to the diary portion of the study, participants completed pre and post study questionnaires.

Participants and Setting

We recruited participants from the general university community. Twenty participants, age 19-47, took part in the study (16 males, 4 females). Participants were highly educated where the minimum education level was some university, with 65% having completed at least an undergraduate degree in primarily technical fields (14 Computer Science, 4 Science). There were eighteen students, one professor, and an Information Technology professional. Participants were generally experienced computer users (10 years) and spent a considerable amount of time each week using their computer (29-35 hrs/wk) and web browsers (22-28 hrs/wk). On average, they reported usually browsing for personal (48%), work (16%), and educational (36%) reasons.

The homogeneity of the participants limits the validity of results for other populations with less education and computer experience and for those who do not move between contexts of use as laptop users do. However, these initial findings are an important starting point for investigation into this research question. A survey currently underway will help determine if and how the privacy concerns of various groups of users differ.

RESULTS

Current Privacy Management Strategies

Participants indicated what privacy management actions they would take given advance warning that they would work closely with another as they used their web browser. Participants selected multiple responses. One participant indicated that he would take no action. Nine participants indicated that they would chose to retain control of the keyboard/mouse and limit the browser features they would use. Sixteen participants (including 6 of those that would retain control) indicated that they would also take other actions such as clearing or modifying their favorites (11/20), history (13/20), or auto completes (13/20).

Privacy Gradients

Various patterns emerged related to classification using the privacy gradients. However, it is important to recognize that this was a field study; different participants visited and classified different sets of web pages (all pages they happened to visit during that week). As such, if two people exhibited similar behaviours, it does not necessarily mean that they have similar privacy perspectives. These patterns reflect the perceived need for privacy based on the sites that an individual visits.

Utilization of Gradients

All participants utilized all privacy categories when classifying their visited web pages (with the exception of one who never used the *don't save* category). Overall,

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.

36,170 pages were visited with 42% classified as *public*, 25% as *semi-public*, 15% as *private*, and 18% as *don't save*.

A K-means cluster analysis grouped the participants into four clusters based on the relative proportions of sites they classified into each privacy gradient (Table 1). Examination of the cluster means revealed that each of the four clusters represents a group of individuals with a relatively high proportion of web browsing in one of the privacy gradients (C1-*semi-public*; C2-*private*; C3-*public*; C4-*don't save*). The fact that C3 contains 50% of the participants suggests that this privacy breakdown may be 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 2, 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 five *semi-public* pages.

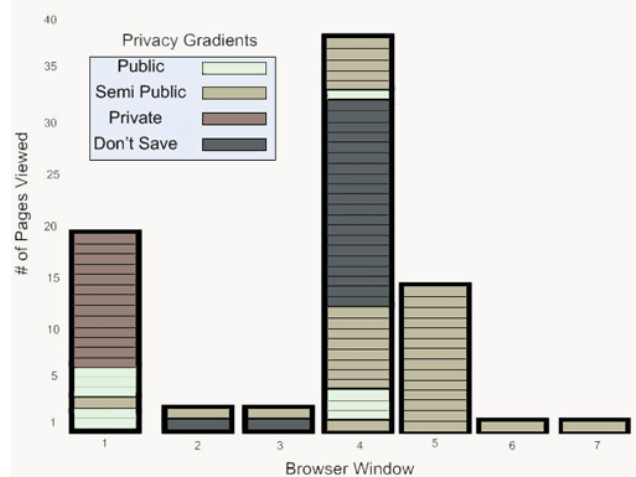


Figure 2. Example of sequential patterns of privacy gradient usage on a per browser window basis (1 hour of data from participant #1)

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, streaks were quite long (up to 166 pages).

Transitions

We define a transition to be a switch between privacy levels within a browser window. For example, in Figure 2, there are five transitions in window #4. We found 56% of browser windows contained no transitions, and on average, participants had 0.9 transitions per window. Strictly looking at the number of transitions in a browser window may be misleading. For example, 5 transitions over 11 pages would indicate that the user transitioned between privacy gradients very frequently; however, 5 transitions over 50 pages are more reasonable. Therefore, we normalized transitions ($\# \text{ transitions} \div \# \text{ pages in window}$), resulting in a numerical score between 0 and 1 where high values indicate rapid transitions. On average, participants had a normalized transition score of 0.14, ranging from 0.03 to 0.31

Goodness of Fit

After working with the privacy gradients for a week, 75% of participants reported the privacy gradients fit 'most of the time', 15% reported they fit 'all of the time', and 10% felt they fit 'some of the time'. Several participants (8/20) reported there were certain sites that did not fit well into the gradients, estimating that 15% of sites were difficult to classify. In most cases, this difficulty was with sites that had multiple purposes or variable content (i.e. news sites).

DISCUSSION

The incidental viewing of traces of previous web browsing activity is indeed a privacy issue. Participants indicated they are concerned when others can view incidental information about their previous web browsing activities: 95% of our participants would take some action to limit the viewing of this information if given advanced warning.

The magnitude of incidental information complicates any privacy management approach. In the case of web browsing, the sheer number of pages that people visit (in our study this was ~260 a day) and the speed at which browsing can occur is staggering. We observed frequent short bursts of ~5 pages per minute. Any manual solution would be overly arduous and therefore impractical.

Behaviours varied considerably in terms of the number of pages visited, number of separate windows in use, and the application of the privacy gradients. Patterns emerged from the data: most participants had streaks of browsing associated with a level of privacy and few transitions within each browser window. Given the per window patterns of privacy streaks with minimal transitions, we believe that one management approach may be to allow browser windows of different privacy levels. These windows could not only filter what incidental information is displayed, but

could also tag new sites visited in that window, similar to the extensional classification described in [4]. However, such a scheme would require integration with a more proactive approach in order to be manageable for users.

CONCLUSIONS AND FUTURE WORK

Our results from this project confirm that management of incidental information is a complex problem due to multiple viewing contexts, individual differences, and the large volume of information. The complexity of this problem will make it difficult to arrive at a standard solution for privacy management. When the incidental information is traces of web browsing activity, there are two main issues: classifying web pages and other artifacts with a privacy level and subsequently displaying the appropriate content when viewed by others. We examined actual patterns of web browsing activity with respect to privacy in an effort to find patterns that may enable a semi-automated approach to privacy management. While initial patterns look promising, we are still examining the data for temporal and individual patterns of privacy gradient use that may help guide personalized solutions.

ACKNOWLEDGMENTS

Thanks to Melanie Kellar, co-developer of the BHO logging tool developed for this study, and to the other members of the EDGE Lab for their continued support. Funding provided 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. Cadiz, J. and Gupta, A. (2001). *Privacy Interfaces for Collaboration* (No. MSR-TR-2001-82). Redmond, WA. Microsoft Research.
3. Hawkey, K. and Inkpen, K. (2004). *Privacy Gradients: Understanding Web Browsing Privacy During Ad Hoc Co-Located Collaboration* (No. CS-2004-18). Halifax, NS. Dalhousie University.
4. Lau, T., Etzioni, O., and Weld, D. S. (1999). Privacy Interfaces for Information Management. *Communications of the ACM*, 42(10): 89-94.
5. Moor, J. H. (1997). Towards a Theory of Privacy in the Information Age. *ACM SIGCAS Computers and Society*, 27(3): 27-32.
6. Palen, L. and Dourish, P. (2003). Unpacking "Privacy" for a Networked World. In *Proceedings of CHI '03*, Ft. Lauderdale, FL. 129-136.
7. Tan, D. S. and Czerwinski, M. (2003). Information Voyeurism: Social Impact of Physically Large Displays on Information Privacy. In *Extended Abstracts of CHI '03*, Ft. Lauderdale, FL. 748-749.