Privacy Management of Incidental Information During Collaboration: Data Analysis and Evaluation Challenges

Kirstie Hawkey
Faculty of Computer Science, Dalhousie University
Halifax, NS B3H 1W5
hawkey@cs.dal.ca

ABSTRACT
Privacy issues related to the viewing of incidental information can impact people's willingness to collaborate on an ad-hoc basis in a co-located setting. Management of incidental information is a complex problem due to multiple viewing contexts, individual differences, and the large volume of information involved. Solutions must balance the amount of control given to the user with the effort required to maintain the system. Our exploratory research on privacy issues for incidental information visible in web browsers has found patterns that may provide a basis for semi-automating privacy management. Data collection techniques included qualitatively annotated web browsing logs. We discuss ongoing challenges in data analysis and evaluation of a privacy management solution.

INTRODUCTION
Colleagues often gather in an ad hoc basis around a computer to collaborate on a project. However, a great deal of incidental information about past activities on the computer is then visible with casual inspection. This information may be inappropriate for the current viewing context. The normative privacy [9] usual for personal displays does not apply during co-located collaboration; the display is an object in the collaboration.

Currently, users must make tradeoffs to manage their privacy: they can either work efficiently in a familiar environment, with access to convenience features and usual layout, or work awkwardly in a more sterile environment. The growing prevalence of ad hoc co-located collaboration on laptop computers, used in a variety of contexts, makes incidental viewing of information a compelling problem. The intersection of privacy management [1, 9] and personal information management [2] results in a hard problem due to the complexity and volume of information and individual differences in behaviour.

The volume of incidental information generated makes this area of research difficult, both for data capture and for analysis. In our current research, we have used client-side logging to capture the pages visited during web browsing and required participants to annotate each log entry with an associated privacy level. This allowed us to merge actual behaviours with the users' qualitative perceptions.

In this paper, we first present our current research and discuss the difficulties that have arisen during data analysis. We then present our next steps as we begin to develop a privacy management solution and discuss the challenges of evaluation.

INITIAL FIELD STUDY
Our overall goal is to provide users with tools to manage incidental information privacy, only revealing information appropriate for the current context. We began by examining privacy issues related to the incidental information found in web browsers. Browsers have many convenience features that assist browsing but display traces of prior activity that users may prefer to remain private (e.g. AutoComplete reveals search terms and URLs).

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 gradients. It was important to explore normal web browsing activities to see if patterns exist that would make organization within privacy levels easier. We conducted a week-long field study [5], recording all web browsing conducted on the laptops of 20 participants. Participants used their laptops for the majority of their web browsing so we could get a complete picture of privacy issues during their web browsing both at and away from home.

To facilitate classification of visited websites, a common terminology was required. A four-tier privacy scheme partitioned web sites: *public*, *semi-public*, *private*, and *don't save* (see Figure 1). *Public* sites are those appropriate for anyone to view (including the Queen, hence the crown in Figure 1), *semi-public* may not be appropriate in some viewing contexts, *private* are only suitable for close confidants to view, and *don't save* includes sites that nobody, including the user, needs to see again.

Data Collection
The choice of data capture techniques for web browsing behaviour impacts the naturalness of the environment for participants, the ease of developing and supporting logging tools, and the type of data available. The ability to maintain participant privacy (not recording visited pages externally) and to gather rich information about user activity on a per-window basis led us to a client-side solution. We developed a Browser Helper Object (BHO) to record the web

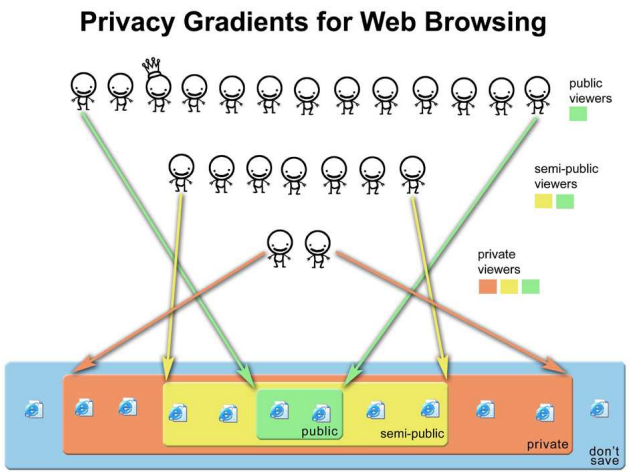

Figure 1. Privacy gradients used during field study

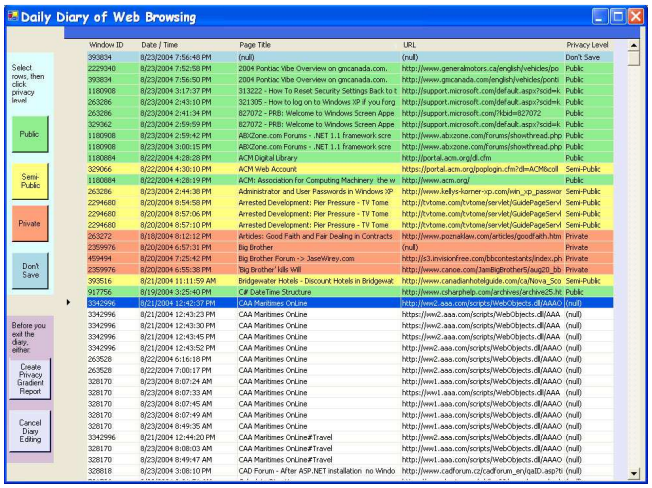

Figure 2. Electronic diary used during field study

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. An advantage of the BHO was that the users' browsing environment did not change: they were able to continue using Internet Explorer with all their normal features and settings intact.

We also developed an electronic diary (see Figure 2) to allow participants to assign privacy gradients to their web browsing daily (if possible). The diary displayed all 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 traces of the page 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 field was updated for all selected records and the row was coloured appropriately.

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 were not interested in which sites participants classified in the various privacy levels, just in the patterns of gradient application. 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. Participants also completed pre and post study questionnaires.

Preliminary Results

Results from the field study revealed that the privacy of incidental information during web browsing is indeed an issue: given advanced warning, 95% of our participants indicated they would take some action to limit visibility of this information. Trends emerged in the perceived privacy of incidental information. Participants clustered into four groups depending on their overall gradient use. Patterns emerged on a per window basis with most participants having streaks of browsing at a privacy level and few transitions between levels (as seen in Figure 3).

We also learned a great deal about the general web browsing behaviours of participants [6] that can affect the feasibility of web browsing tools. These behaviours were highly variable, both between and within participants. For example, on average, the participants in our study visited 1808 pages during the seven days (~258/day), ranging from 422 (~60/day) to 5127 pages (~732/day). Participants opened an average of 289 different browser windows during the week (from 47 to 499). In most cases, only one or two pages were viewed within each window; however, there were also several instances where large numbers of page views occurred within a browser window. The average maximum number of pages viewed in one window was 108 (from 27 to 255). Participants frequently (~37 times per day) exhibited rapid bursts of browsing with several pages loaded per minute. Overall, the average duration of a burst (defined as less than 1 minute between consecutive pages visits) was 82 seconds, but the longest burst was over 36 minutes. The average length of a burst was 7 pages, with

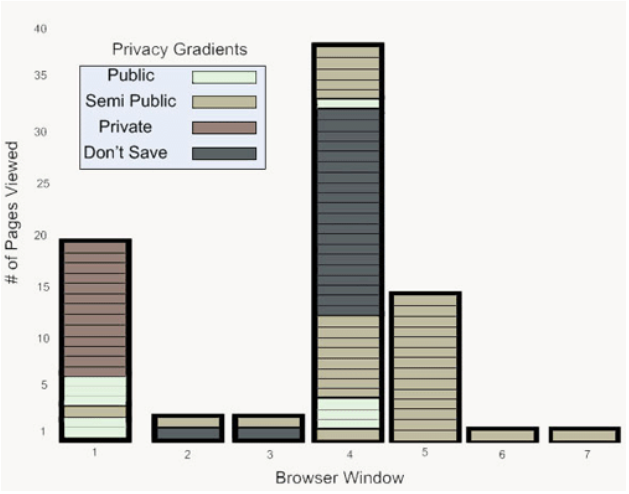

Figure 3. Example of sequential patterns of privacy gradient usage on a per browser window basis.

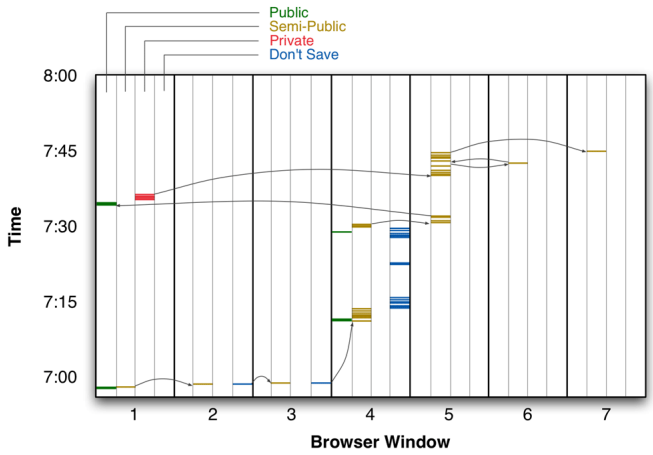

Figure 4. Example of temporal patterns of privacy gradient usage on a per window basis. Arrows indicate movement between browser windows.

bursts of up to 172 pages.

Data Analysis Challenges

While our diary reports do not require transformation (as in [7]), analysis and visualization of the data has proven to be problematic. The sheer number of sites visited has resulted in massive participant-annotated web browsing logs, making it difficult to view the patterns of interest. Figures 3 and 4 are both useful representations of web browsing activity (both show 1 hour of data from participant #1) that we created manually. Figure 3 shows the sequential patterns of privacy gradient usage on a per browser window basis, while Figure 4 shows the temporal patterns. Both representations are pertinent when looking for privacy patterns to exploit in a solution and are difficult to envision when looking at a textual data log. A visualization technique is under development to gain a richer understanding of patterns uncovered and their applicability to individual and general solutions.

Individual differences in user behaviours have also made data analysis challenging. Scripting has allowed us to tease out information such as streaks at a privacy level, transitions between levels, and bursts of activity. However, numerical averages across users do not allow us to view the individual patterns at play. It is also necessary to examine the extremes of behaviour for each individual in order to ensure the feasibility of proposed solutions. Further analysis is required to build models of individual behaviours. We will use data mining to look for more patterns within the logging data, particularly those of a temporal nature.

NEXT STEPS

Solution Requirements

Privacy management of incidental information is complex due to the multiple contexts of its creation and viewing. To protect privacy within the normal computing environment, we must balance the amount of control a user has over the environment with the time and effort necessary to provide that control. Data analyses from the field study and a survey will soon be complete. Outcomes from the studies will guide the development and evaluation of a privacy management system for the incidental information generated during web browsing. It is clear we must utilize patterns inherent during web browsing to relieve the burden of the user manually classifying all incidental information.

It is also important to understand these patterns, as web browsing is such a frequent activity. There is a continuing need to research the daily activities and gain an understanding of users' tasks and behaviours [10]. The results from our study clearly demonstrate that variability and magnitude of browsing behaviours complicate the development of any tool or technique for web browsing. The sheer number of pages that people visit while browsing means that manual tools, that operate on a per-page level, will be overly arduous and therefore impractical. Beyond the number of pages visited, the speed with which users browsed was at times staggering. The high volume of web sites visited and the rapid browsing indicate the need for seamless interactions between user and tool.

The variability of web browsing behaviours across users may make it difficult to arrive at standard solutions for web browsing tools and techniques. Furthermore, there is a high amount of variability within the browsing of a single user. Solutions must be sensitive to the changing needs and behaviours of users and allow users flexibility.

Possible Solution

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. Given the per window patterns of privacy streaks with minimal transitions revealed in our early results, one approach may be to utilize browser windows of different privacy levels.

For the purposes of displaying appropriate content when others can view your 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 (except those marked as don't save) would be visible.

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 [8]. However, integration with a more proactive approach is required in order to be manageable for users. Automatic comparison of current classifications with those made previously and intelligent defaults using identified privacy concerns about categories

of web sites may be useful. Users would also need the ability to make adjustments in classifications.

Evaluation Challenges

In order to evaluate a semi-automated approach, we must examine several different components of the system. We propose using a layered evaluation framework (as in [3]) so that we can evaluate the effectiveness of the semi-automated classifications, the user interface and the maintainability of the system. Longitudinal evaluation will be necessary to determine if the privacy management system is effective and maintainable over time.

System evaluation must occur within a natural setting. However, unless we maintain privacy of participants' computing activities, they may not engage in the normal behaviour we seek to observe. During our field study, we safeguarded the privacy of sites visited and only viewed data relating to gradient use on a per-window and temporal basis, learning more general privacy perspectives through classification tasks and questionnaires. We will take a similar approach to evaluate the effectiveness of the semi-automated privacy classifications.

By occluding information to maintain privacy during evaluation, the burden of evaluation will lie solely with participants. Indeed, even if we were to view the privacy levels and the actual web sites, researchers would be unable to determine accurately if the privacy level was appropriate due to individual differences in perceived privacy needs. We plan to augment the electronic diary to show the privacy classification scheme applied by the system. Participants can then annotate the entries with their evaluation of the appropriateness of the scheme. As before, the logs would need to be sanitized, so as not to reveal to the researchers the actual web sites visited.

Evaluation may be onerous for participants if we require them to provide fine-grained effectiveness ratings over time. Even with the electronic diary's sorting features and multiple row selection capability, participants found the exercise tedious due to the number of sites visited during the week. We therefore propose to require participants to evaluate a limited number of system classifications periodically, perhaps evaluating one day's worth on bi-weekly basis. In addition, we may alternate these evaluations with qualitative usability reports in order to not influence user satisfaction [4] by negatively associating the system with the effort of completing the diary.

We will also instrument the browser windows to log events such as the privacy type of the window opened, the frequency of switching between types, and the frequency of user modifications of system classifications. These logs, combined with the qualitative reports, will give us a clearer understanding of system usability. The tools currently under development to visualize the patterns in the data collected to date will be extensible for use during analyses of evaluation data.

CONCLUSION

We have discussed our current research including our use of participant-annotated logging data to gain insight into the privacy patterns inherent during web browsing. The general web browsing behaviours observed motivate the need for seamless interactions between users and web browsing tools. Evaluation of a privacy management solution will be difficult due to the need of maintaining user privacy and evaluating the appropriateness of thousands of automatic classifications. A combination of logging and qualitative methods with visualization tools will enable effective evaluation.

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. Boardman, R. and Sasse, M. A. (2004). "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.
3. Gupta, A. and Grover, P. S. (2004). Proposed Evaluation Framework for Adaptive Hypermedia Systems. In *Proceedings of Third Workshop on Empirical Evaluation of Adaptive Systems, AH2004*, Eindhoven University of Technology, The Netherlands.
4. Hassenzahl, M. and Sandweg, N. (2004). From Mental Effort to Perceived Usability: Transforming Experiences into Summary Assessments. In *Proceedings of CHI 2004*, Vienna Austria. 1283-1286.
5. Hawkey, K. and Inkpen, K. (2005). Privacy Gradients: Exploring Ways to Manage Incidental Information During Co-Located Collaboration. In *Proceedings of CHI2005 (to appear)*, Portland, Oregon.
6. Hawkey, K. and Inkpen, K. (2005). Web Browsing Today: The Impact of Changing Contexts on User Activity. In *Proceedings of CHI 2005 (to appear)*, Portland, Oregon.
7. Hilbert, D. and Redmiles, D. (2000). Extracting Usability Information from User Interface Events. *ACM Computing Surveys, 32*(4): 384-421.
8. Lau, T., Etzioni, O., and Weld, D. S. (1999). Privacy Interfaces for Information Management. *Communications of the ACM, 42*(10): 89-94.
9. Moor, J. H. (1997). Towards a Theory of Privacy in the Information Age. *ACM SIGCAS Computers and Society, 27*(3): 27-32.
10. Whittaker, S., Terveen, L., and Nardi, B. A. (2000). Let's Stop Pushing the Envelope and Start Addressing It: A Reference Task Agenda for HCI. *Human Computer Interaction, 15*: 75-106.