Privacy Research: A Mixed Methodology Approach

Kirstie Hawkey
Faculty of Computer Science, Dalhousie University
6050 University Avenue
Halifax, NS, Canada B3H 1W5
hawkey@cs.dal.ca

ABSTRACT

This paper describes our experiences researching privacy of incidental information using a mixed methodology approach. An on-line survey examined privacy concerns related to the incidental viewing of traces of previous web browsing activity. Topics examined include the scope of the privacy issues in this domain; how browsing behaviours affect the content that may be visible; and the role of content sensitivity, level of control, and viewer on privacy comfort levels. Two field studies allowed us to examine how participants felt in terms of privacy about specific instances of visible content (the web pages they had visited that day) and to examine patterns in the application of privacy levels to that content. This detailed information was necessary to explore the feasibility of possible privacy management approaches.

KEYWORDS

Privacy, methodology, survey, field study

INTRODUCTION

All research methodologies have inherent flaws and benefits in terms of the ability to generalize results, measure behaviours and attitudes precisely, control confounding factors, and conduct the research within a realistic context [7]. Privacy is a complex issue with both privacy concerns and willingness to maintain a management scheme varying on an individual basis. There is often a dichotomy between responses on attitudinal surveys and the actual privacy preserving behaviours observed (or lack thereof) (e.g. [2]). Attitudinal surveys may measure an ideal privacy standard; however, in practice privacy issues are not so clear cut. Users must make the tradeoff between the cost of configuring and maintaining a privacy management system and the potential benefits of guarding that privacy.

Survey research is popular as surveys are fairly easy to develop, administer, and analyze. While a carefully sampled survey may increase ability to generalize results, it is limited to measurement of self-reported attitudes and behaviours. This can be particularly troublesome with the sensitive nature of privacy research as the attitudes and behaviours reported by participants may be skewed due to participants' tendency to give socially desirable responses [7]. Attitudes may also be impacted by situational and cultural relativities [3]; for example, recent events (e.g. a privacy violation) can temporarily heighten sensitivity.

Laboratory studies allow researchers to observe privacy practices in action in a controlled fashion; however, it is difficult to provide a sufficiently realistic experimental setup that will compel participants to engage in normal behaviours. This is particularly challenging in privacy and security research due to the highly personal nature of the data at stake. It can be difficult to motivate participants to make the effort and take the same actions with study data as they would normally take if the data was their own. For instance, three participants in a study of privacy preferences for an awareness application indicated that they set preferences at the team level instead of the group level because it would allow them to more quickly finish the study [8]. Similarly, in a study of the cues that participants view to evaluate the security of a web site, real participant data (e.g. credit card numbers) could not be used and participants had difficulty treating the dummy credit card number with the same care as their own [9].

Field research theoretically allows the study of actual behaviours in a realistic environment. However, the act of observing or recording a participant's personal interactions may cause them to alter those behaviours. For example, behaviours deemed to be socially inappropriate may be avoided during the period of the study. As well, participants may also be unwilling to have logging software installed that may record personal interactions (e.g. a keystroke logger may pick up passwords), particularly if that software logs data across applications.

The remainder of this paper presents the mixed methodology approach of a survey and field studies that we have used to study the domain of incidental information privacy (full details available in [4-6]). We first give a brief description of incidental information privacy and then discuss our research methodology and the challenges met. We conclude with future research plans.

PRIVACY OF INCIDENTAL INFORMATION

Our research involves the privacy issues that arise during co-located collaboration around somebody's personal display. In addition to the documents relevant to the task at hand, often other incidental information is visible on the display, both within the operating system (e.g. file names) and within applications (e.g. IE history files). Our research has focused on the incidental information visible in web

CHI 2006 Workshop, Privacy and HCI: Methodologies for Studying Privacy Issues

browsers. Web browsers are often used during co-located collaboration and have a variety of convenience features such as auto-completion, history, and bookmarks that assist users in navigating to previously visited pages. These convenience features display traces of previous web browsing activity that may or may not be appropriate for the current viewing context.

Through an examination of related work on privacy theory in other privacy domains and our research results to date, several dimensions of incidental information privacy that impact a user's comfort level have been identified and explored (see Figure 1). Four dimensions that directly impact an individual's *privacy comfort level* in a given situation include their *inherent privacy concerns*, their *level of control* over their computer, their *relationship to the viewer* of the display, and the potentially *visible content*. Furthermore, the visible content may depend upon recent *browsing activity*, *browser settings*, and any *preventative actions* taken. Browsing activity itself may vary depending on the *location* of the activity and the type of *computer in use*. While Figure 1 shows the major influences on privacy comfort levels, these dimensions are often inter-related. A multi-method approach is required to examine these dimensions fully.

MIXED METHODOLOGY APPROACH

A survey examined privacy concerns related to the incidental viewing of web browsing traces. The survey objectives were threefold: 1) to understand the scope of the privacy issues in this domain, 2) to examine how browsing behaviours affect the types of visible content, and 3) to investigate the role of content sensitivity, level of control, and viewer on privacy comfort levels. However, the survey alone only represents users' self-reported perceptions of their concerns; it was important to build a more complete picture grounded in actual behaviours

The survey allowed us to examine participants' stated privacy comfort levels for varying levels of control and relationships to viewers; however, the content was limited to scenarios of varying sensitivity. In contrast, the two field studies allowed us to examine how participants felt in terms of privacy about specific instances of visible content (the web pages they had visited that day) and to examine patterns in the application of privacy levels to that content. This detailed information was necessary to explore the feasibility of possible privacy management solutions. We next present the survey and field studies in more detail and discuss the challenges encountered.

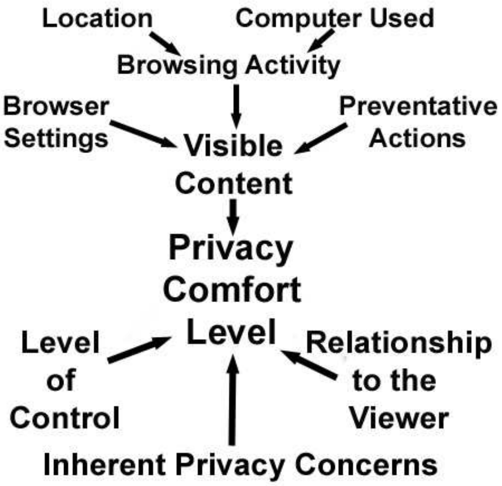


Figure 1. Dimensions that affect the comfort level of users during incidental viewing of traces of prior web activity.

Incidental Information Privacy Survey

The survey (see [6] for details) allowed us to ask many questions of many people (155 participants). The survey was on-line, so participants could complete it at a convenient time and location. One limitation of survey research is that participants must reflect upon their attitudes and experiences while not in the context of those experiences. However, in the incidental information domain, current privacy management is largely a matter of speculation: What traces of my past activities will be visible on my monitor? Who will be able to view it? Should I clear my history files? Additionally, people have to speculate about how others would regard these traces of activity that they have conducted in the past. In this regard, a survey was a good choice to explore attitudes and get self-reported data about typical web browsing behaviour and current privacy management practices.

Depending on the domain under study, there can be a huge volume of information to be considered and many contexts in which the information may be viewed. We elected to use general cases in our survey (e.g. viewer categories such as 'close friend'), but there is also a need to look at specific instances in order to increase the realism of the scenario. Some researchers have had participants instantiate an attribute (e.g. give the name of a close friend and use that in the questions). However, even an instantiated attribute may not reflect the spectrum of possible situations. A participant may consider several people to be close friends but may not share information with them all equally. Even for a specific person, privacy concerns may fluctuate (e.g. after a disagreement).

Privacy Gradient Field Studies

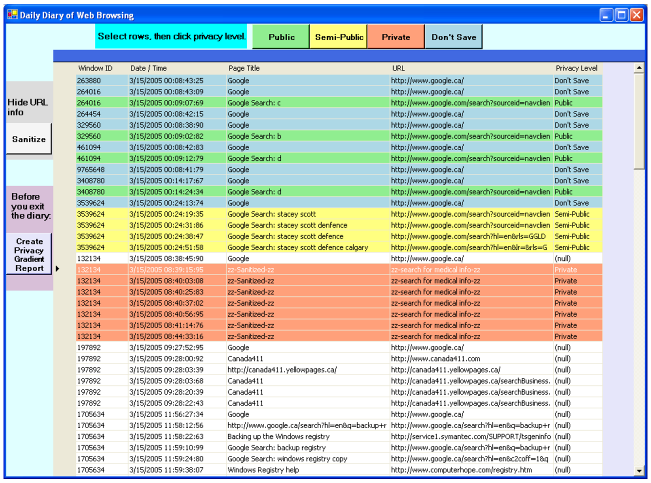
The two week-long field studies, Privacy Gradients 1 (PG1) (see [4] for details) and Privacy Gradients 2 (PG2) (see [5] for details) afforded an in-depth look at the privacy concerns for a more limited number of people (20 in PG1, 15 in PG2). It was important to explore normal web browsing activities to see if patterns exist that would make organization within privacy levels easier.

Our hypothesis was that people would be willing to organize their information across a small number of privacy levels. Participants were asked to partition visited websites using a four-level privacy scheme: *public*, *semi-public*, *private*, and *don't save. Public* sites are those someone is comfortable with anybody and everybody viewing. *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. The privacy gradients gave consistent terminology but allowed flexibility for participants to indicate their own privacy concerns for visited pages regardless of which viewer types they would consider to be at the confidant level, at the public level, or somewhere in between.

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 work with Microsoft's Internet Explorer (IE) and 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. 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.

Beyond recording participants' web browsing behaviours, we needed to have them qualitatively annotate the web pages visited with a privacy level. We also needed to ensure that doing so did not change their behaviours. We therefore elected to not interrupt the flow of their web browsing by having then annotate web pages as they visited them, but to have them do so on a daily basis. Participants were provided with an electronic diary (see Figure 2) that displayed details of all the visited pages (browser window ID, date/time stamp, page title, URL). Participants were required to indicate how they would classify the privacy level of each web page if others were to view traces of this activity and 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.

We were very concerned that participants' normal browsing activities were not altered because their browsing was being recorded. We therefore elected to remove the page title and URL from the data before it was sent to the researchers ensuring participants' privacy. After classifying their web pages with a privacy level, a report was generated and displayed so that participants could see what data (browser window ID, date/time stamp, and privacy level of each visited page) would be sent.


Daily Diary of Web Browsing
Select rows, then click privacy level.
Public
Semi-Public
Private
Don't Save
Hide URL info
Sanitize
Before you exit the diary:
Create Privacy Gradient Report

Window ID	Date / Time	Page Title	URL	Privacy Level
263880	3/15/2005 00:08:43:25	Google	http://www.google.ca/	Don't Save
264016	3/15/2005 00:08:43:09	Google	http://www.google.ca/	Don't Save
264016	3/15/2005 00:09:07:69	Google Search: c	http://www.google.com/search?sourceid=navclien	Public
264454	3/15/2005 00:08:42:15	Google	http://www.google.ca/	Don't Save
329560	3/15/2005 00:08:38:90	Google	http://www.google.ca/	Don't Save
329560	3/15/2005 00:09:02:82	Google Search: b	http://www.google.com/search?sourceid=navclien	Public
461094	3/15/2005 00:08:42:83	Google	http://www.google.ca/	Don't Save
461094	3/15/2005 00:09:12:79	Google Search: d	http://www.google.com/search?sourceid=navclien	Public
9765648	3/15/2005 00:08:41:79	Google	http://www.google.ca/	Don't Save
3408780	3/15/2005 00:14:17:67	Google	http://www.google.ca/	Don't Save
3408780	3/15/2005 00:14:24:34	Google Search: d	http://www.google.com/search?sourceid=navclien	Public
3539624	3/15/2005 00:24:13:74	Google	http://www.google.ca/	Don't Save
3539624	3/15/2005 00:24:19:35	Google Search: stacey scott	http://www.google.com/search?sourceid=navclien	Semi-Public
3539624	3/15/2005 00:24:31:86	Google Search: stacey scott denfence	http://www.google.com/search?sourceid=navclien	Semi-Public
3539624	3/15/2005 00:24:38:47	Google Search: stacey scott defence	http://www.google.com/search?hl=en&rls=GGLD	Semi-Public
3539624	3/15/2005 00:24:51:58	Google Search: stacey scott defence calgary	http://www.google.com/search?hl=en&lr=&rls=G	Semi-Public
132134	3/15/2005 08:38:45:90	Google	http://www.google.ca/	(null)
132134	3/15/2005 08:39:15:95	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:03:08	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:25:83	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:37:02	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:40:56:95	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:41:14:76	zz-Sanitized-zz	zz-search for medical info-zz	Private
132134	3/15/2005 08:44:33:16	zz-Sanitized-zz	zz-search for medical info-zz	Private
197892	3/15/2005 09:27:52:95	Google	http://www.google.ca/	(null)
197892	3/15/2005 09:28:00:92	Canada411	http://www.canada411.com	(null)
197892	3/15/2005 09:28:03:39	http://canada411.yellowpages.ca/	http://canada411.yellowpages.ca/	(null)
197892	3/15/2005 09:28:03:68	Canada411	http://canada411.yellowpages.ca/searchBusiness.	(null)
197892	3/15/2005 09:28:20:39	Canada411	http://canada411.yellowpages.ca/searchBusiness.	(null)
197892	3/15/2005 09:28:22:43	Canada411	http://canada411.yellowpages.ca/searchBusiness.	(null)
1705634	3/15/2005 11:56:27:34	Google	http://www.google.ca/	(null)
1705634	3/15/2005 11:58:12:56	http://www.google.ca/search?hl=en&q=backup+r	http://www.google.ca/search?hl=en&q=backup+r	(null)
1705634	3/15/2005 11:58:22:63	Backing up the Windows registry	http://service1.symantec.com/SUPPORT/tsgeninfo	(null)
1705634	3/15/2005 11:59:10:99	Google Search: backup registry	http://www.google.ca/search?hl=en&q=backup+r	(null)
1705634	3/15/2005 11:59:24:80	Google Search: windows registry copy	http://www.google.ca/search?hl=en&c2coff=1&q	(null)
1705634	3/15/2005 11:59:38:07	Windows Registry help	http://www.computerhope.com/registry.htm	(null)

Figure 2. Screenshot of electronic diary (PG2 version) used by participants to annotate visited pages with a privacy level

PG1 examined participants' perception of the privacy of their web browsing and found patterns to their applications of privacy levels. However, as the content was blinded, 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.

We designed PG2 to gather information about browsing activity both in an effort to learn more about what content is potentially visible, the relationship of the content to the privacy levels that participants apply to visited pages, and to evaluate the feasibility of users classifying privacy on a per-category basis. Additionally, participants in the first field study were laptop users with a primarily technical background. In order to study users with other characteristics, three different classes of participants were recruited: 5 technical desktop users, 5 non-technical desktop users, and 5 non-technical laptop users.

In PG2, the client-side logging software was modified to record additional contextual information about participants' web browsing such as focus events and location. The electronic diary was modified so that 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". 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.

Content categories (see [1]) were used by participants to theoretically specify their privacy comfort for each category and by researchers to partition participants' actual browsing using the page title and URL to determine the content categories of visited pages. We examined how privacy levels change according to the category of visited page, how similar participants were in their privacy level applications, how consistent participants were at classifying their browsing, and how accurate participants were at choosing a theoretical privacy level for the categories.

Results revealed that the categories of web pages clustered into five groups based on participants' overall application of privacy levels to their web browsing. However, 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 a content category, results showed that participants were personally consistent within most categories. This suggests that a personalized scheme may be feasible. However, 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.

CONCLUSION & FUTURE WORK

Our mixed methodology approach has allowed us to examine the privacy of incidental information both in terms of general attitudes and also based on actual behaviours. We are currently working on integrating results from the three studies to build a privacy model for this domain. We are also investigating methods of determining participants' inherent privacy concerns.

The field studies motivate the need for an intelligent system support approach to privacy management; the sheer volume of pages visited and the rapid bursts of browsing observed would make a per-page management solution unwieldy. The second field study reveals that a personalized scheme based on content categorization of visited web pages may be a suitable approach. There are also privacy patterns (e.g. streaks at a given privacy level) and temporal patterns (e.g. rapid bursts of browsing) to web browsing activities (see [4] for details) that may be capitalized upon in an intelligent systems approach. Further analysis of the contextual data from the second 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. We are also examining methods of supporting users in their understanding and control of the privacy of incidental information.

ACKNOWLEDGMENTS

Thanks to Dr. Kori Inkpen and the members of the EDGE lab for their support and guidance. This research is funded in part by NSERC and NECTAR.

REFERENCES

1. *Cerberian Web Filter Categories*, in http://www.webrootdisp.net/audit/rating-descriptions.htm.
2. Acquisti, A. and Grossklags, J., *Privacy and Rationality in Individual Decision Making*. IEEE Security & Privacy Magazine, 2005. **3**(1): 26-33.
3. Clarke, R., *Statement for Panel on Information Privacy in a Globally Networked Society: Implications for I.S. Research*. 2002, http://www.anu.edu.au/people/Roger.Clarke/DV/ICIS2002.html: ICIS 2002.
4. Hawkey, K. and Inkpen, K., *Privacy Gradients: Exploring ways to manage incidental information during co-located collaboration*, in *Ext. Abstracts CHI '05*. ACM Press. (2005) 1431-1434.
5. Hawkey, K. and Inkpen, K.M. Examining the Content and Privacy of Web Browsing Incidental Information *WWW 2006 (to appear)*, (2006),
6. Hawkey, K. and Inkpen, K.M. Keeping Up Appearances: Understanding the Dimensions of Incidental Information Privacy. *CHI '06 (to appear)*, ACM Press (2006),
7. McGrath, J.E., *Methodology matters: doing research in the behavioral and social sciences*, in *Human-computer interaction: toward the year 2000*, J.G. R. Baeker, W. Buxton, and S. Greenberg, Editor. (1995). 152-169.
8. Patil, S. and Lai, J. Who Gets to Know What When: Configuring Privacy Permissions in an Awareness Application. *CHI '05*, ACM Press (2005), 101-110.
9. Whalen, T. and Inkpen, K.M. Gathering evidence: use of visual security cues in web browsers. *Graphic Interface*, (2005), 137-145.