Public Perception of Private Information on Search Engines

Tara Whalen
Carrie Gates

Technical Report CS-2005-01

Jan 25, 2005

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

Public Perception of Private Information on Search Engines

Tara Whalen
Faculty of Computer Science
Dalhousie University
Nova Scotia, Canada
whalen@cs.dal.ca

Carrie Gates
Faculty of Computer Science
Dalhousie University
Nova Scotia, Canada
gates@cs.dal.ca

Abstract

The principal strength of search engines is that they enable people to retrieve information easily. This power, however, represents a threat to privacy. Personal information, such as home addresses and membership in organizations, can be easily retrieved by individuals all over the globe. This has caused widespread concern among citizens and privacy organizations alike. Google has stated that it is not willing to censor information, but wants the internet community to decide whether or not certain types of data should be made public [1]. To help inform this debate, we are conducting a study on public perception of private information available through search engines. In particular, we focus on how people perceive certain types of information, and on its temporal nature. This paper presents the results from a pilot study designed to spark interest in the issues, and to inform the construction of future work in this area.

1 Introduction

Internet search engines, such as Google and Yahoo, have become a ubiquitous resource, used daily by a large segment of the on-line population. Google, for example, has indexed over 4 billion pages. Many of these web pages contain personal information, such as resumes, journals or blogs, photos, or favourite recipes. Such pages are not necessarily restricted to personal information about only the author of the page, but may contain information about other people, such as their friends or family. Beyond this, many employer sites will provide information about their employees, and many government sites will provide data that, while although always publicly available, now becomes extremely easy to access through cursory searches.

Search engines have further complicated the issue of information being publicly available by maintaining caches of information that has previously been deleted. This allows a user to view the content of a web page, even when the original web page is no longer available. Thus, even if a user deletes a web page containing some personal information, the user has not necessarily removed this personal information from future access by others. This is in some ways a bonus feature of search engines, allowing users to still access needed information, but the downside is the privacy implications of keeping certain information accessible.

While privacy is a key issue for both governments and citizens, studies on user attitudes towards privacy have focused on the collection and sharing of personal information by corporations and governments (e.g. [2], [7]) and on what types of information people consider to be private under these circumstances (e.g. [3]). Additionally, papers have been written on the value of private information (e.g. [6]) and technologies for ensuring privacy (e.g. [4]). Very few studies have investigated the public availability of private information via the world wide web, nor have many studies been performed to determine what information might be considered private in an on-line context. Further, no studies have been performed to date that link the availability of private information to its being indexed for searching via the web. There are also no studies that have investigated users' attitudes towards the length of time that various types of personal information should be cached on (and thus available through) a search engine.

This paper presents the results from a pilot study that was based on a questionnaire survey. The goal was to determine if there was some information that people consider private, and therefore should not be searchable via the web, and if there was a temporal aspect to this type of information. That is, do users feel that personal information should be available only for a limited time? The pilot study presented here was performed to determine if there were particular areas

of concern that should be further investigated, and to refine the questionnaire itself.

A description of the pilot study undertaken is described in Section 2, where the survey and the population are both described. Section 3 presents a summary of the answers from the questionnaire, highlighting the key issues. These results are discussed in the following section, with an emphasis on the implications they have for search engine policy. Further discussion is provided in Section 5, where our work is placed in the context of related studies. Section 6 describes how future work in this area can be guided by this study, with some concluding remarks provided in Section 7.

2 Description of Study

2.1 Population

For the pilot study, we chose to use librarians and library students as the target audience. This group was chosen because it represents a well-educated population that are both internet-savvy and familiar with the use of search engines. In addition, librarians tend to have a diverse background, including science, arts, business and education; in general, they also value access to information, in balance with privacy and policy issues. We solicited 114 individuals from a single university via an email sent from the head librarian (in the case of librarians) and from the graduate coordinator at the School of Library Sciences (in the case of the library students). There were 16 responses, representing a response rate of 14%.

All of the respondents were older than 18, with 5 respondents aged 18–25, 2 aged 26–35, 3 aged 36–45 and 6 aged 46–55. There were 11 female and 4 male respondents, along with one who did not provide gender information. Seven of the respondents were parents, 5 of whom still had children living at home.

All of the respondents had at least one university degree, with 8 library students and 6 librarians; 3 respondents did not provide their current professional status. All of the respondents have been using the internet for more than 4 years, with 12 of them having used it for 7 years or more. All of the respondents save one used the search engines daily (the one exception used search engines weekly). However, despite the frequent and long-term use of the internet, only half of the respondents have ever had a web page or a blog.

Thus the population chosen represented a wide variety in terms of age, sex and parenthood. However, the population was also consistently highly educated and well-versed in using the internet and search engines, with a great deal of experience in this area.

2.2 Survey

This study was conducted through the use of an online questionnaire survey. Participants were recruited through an email request sent by an administrator. Those who were interested in participating were asked to contact one of the principal investigators. The investigators responded with the consent form, an ID number and the URL for the survey.

The survey was conducted via the web, and was composed of five pages. The first page consisted of prompting the user for demographic information, such as their age bracket, sex, and use of search engines.

The next set of questions were designed to ascertain the level of comfort the respondent had for having certain kinds of information available through search engines. The respondent was asked to select their response from a five-point Likert scale: very uncomfortable, uncomfortable, don't care, comfortable and very comfortable. However, due to the small number of respondents, the analysis provided in Section 3 places "very uncomfortable" and "uncomfortable" into one group, and "comfortable" and "very comfortable" into one group, resulting in three categories overall. The second page of the survey consisted of 28 questions asking if specific types of personal information about the respondent should be available through a search engine. The third page required the respondent to assume that he was a caretaker for a child between the ages of 8 and 12 years old. The page consisted of 12 questions pertaining to the comfort level of the respondent knowing the availability of information about the child through search engines. The fourth page consisted of 7 questions about the comfort level of the respondent to having club or organization membership information available through a search engine.

The final set of questions were designed to determine the temporal nature of personal information available through a search engine. The fifth page of the questionnaire consisted of 15 questions of this nature, divided into personal information (e.g. home address, marital status), information on public postings (e.g. journal entries and newsgroup postings), and information on photos of the respondent. The question asked was how long this information should be available through a search engine. The possible responses to these questions were: no time (the item should not be available), less than six months, 6–12 months, 1–3 years, 4–6 years, 7 or more years, until the source is removed, or forever. In addition, the section on personal information included the option of until the source is changed. Due to the small number of respondents, some of this information was grouped. Specifically, the

responses of less than six months, 6–12 months and 1–3 years were grouped into a single category of less than 3 years. Additionally, the responses of until the source is changed and until the source is deleted were combined into "source-based" changes. Finally, because there was very low selection of responses longer than 3 years (one response for "4–6 years" and one for "7 or more years"), these two categories were ignored.

After each subsection, respondents were asked if they had been given the option of "until I ask for the information to be removed from the search engine," would they have chosen it for any of the questions, and if so, for which questions. Page five finished by providing three statements. Respondents were asked to rate their agreement with the statement by choosing one option on a five-point Likert scale (agree strongly, agree somewhat, neither agree nor disagree, disagree somewhat, or disagree strongly). For the last question, respondents were given the opportunity to elaborate.

The final page of the survey was somewhat more free-form. Respondents were asked to list their primary concerns with having their email available through the web, being provided with five possible options, along with the ability to add other comments. They were then asked to indicate their agreement with a particular statement, using the same scale provided on the previous page. Respondents were asked if they had ever found any information about themselves on-line that they had wished had not been made public, and then given the chance to elaborate if the answer was yes. They were then asked if they had ever performed an internet search for themselves and, if yes, why they had done so. Finally, respondents were asked if they had (either currently or previously) a personal web page. If yes, they were asked if they had ever limited access to one of their web pages, and if so using what technology, where the options were: password protection, `robots.txt` files, and/or creating an unlinked web page.

3 Responses

The first section of the survey consisted of questions asking how comfortable the respondent was with having various pieces of information available on the web through a simple search. These included questions about home and work addresses and phone numbers, as well as pictures, religious and political affiliations, and favourite books, foods and movies. The questions asked were:

1. your home address
2. your work address
3. your home phone number
4. your work phone number
5. map and driving directions to your home
6. your email address
7. your age
8. your occupation
9. your salary
10. your child/children's name(s)
11. a picture of your face (like a passport photo or other "headshot")
12. an unidentifiable picture of you (e.g., your name does not appear with it)
13. an unflattering picture of you (e.g., you have messy hair or a funny expression)
14. a picture of you receiving a prestigious award
15. your hobbies (e.g., sports, model railroading, gardening)
16. clubs that you belong to
17. religious affiliation(s)
18. political affiliation(s)
19. your favorite foods
20. your favorite movies/TV shows
21. your favorite books
22. your resume
23. a newsgroup posting you wrote about your child's first day of school
24. a newsgroup posting you wrote about a controversial political issue (e.g., gun control, abortion)
25. a newsgroup posting you wrote in the past that contains opinions you no longer agree with
26. a newsgroup posting you wrote that contains a recipe for chocolate cake
27. a newsgroup posting you wrote that demonstrated your knowledge, which solved a person's problem
28. a fictional story that you wrote

The responses are illustrated in Figures 1 and 2.

There were eight questions where more than half the respondents stated that they were uncomfortable having that information available. Three of these were related to home contact information (address, phone, map). Four were related to information which can be considered as very personal: salary, religious and political affiliation, and the name of their child. The last such question was related to vanity — people were generally uncomfortable with having unflattering pictures (e.g. messy hair, funny expression) on the web.

In contrast, there were nineteen questions where more than half the respondents stated that they were

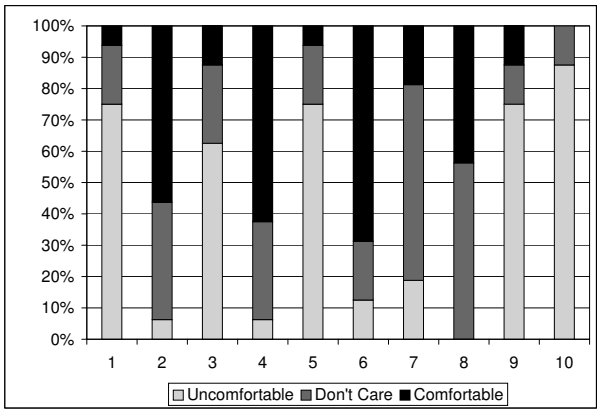

Figure 1. The results for questions 1–10

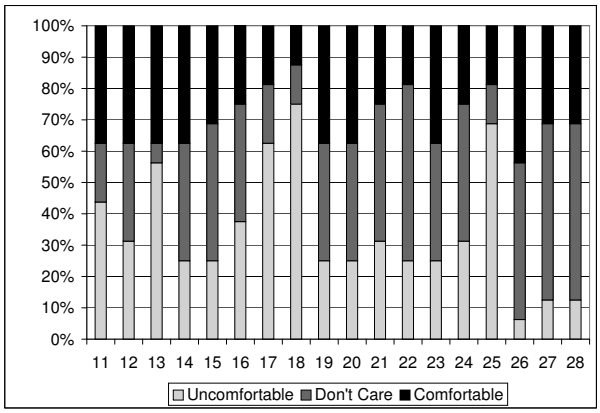

Figure 2. The results for questions 11–28, on the availability of personal information through search engines.

not uncomfortable (e.g. they were either comfortable or did not care) providing the requested information. These tended to cluster around work (e.g. work address and phone number, email, resume), non-controversial information (e.g. chocolate cake recipe, favorite foods) and postings that put the poster in a positive light (e.g. solving a problem for someone).

Two other categories—age and occupation—had very few people state that they were uncomfortable with this information being available. For age, the majority of respondents—ten of sixteen— did not care if the information was available, while three were comfortable and three were uncomfortable. For occupation, nine respondents did not care if the information was provided, while seven were comfortable with this.

The second set of questions asked the respondents to imagine that they were the caretakers of a child between the ages of 8 and 12. The questions then focused on the comfort level for the respondent of knowing that various pieces of information was available about the child. The questions were:

1. home address
2. home phone number
3. map and driving directions to child's home
4. email address
5. age
6. a picture of child's face (like a passport photo or other "headshot")
7. a picture of child receiving a prestigious award
8. an unidentifiable picture of child (e.g., their name does not appear with it)
9. child's hobbies (e.g., sports, collecting comics)
10. clubs that child belongs to
11. favorite movies/TV shows
12. a fictional story that they wrote

The responses to these questions are illustrated in Figure 3.

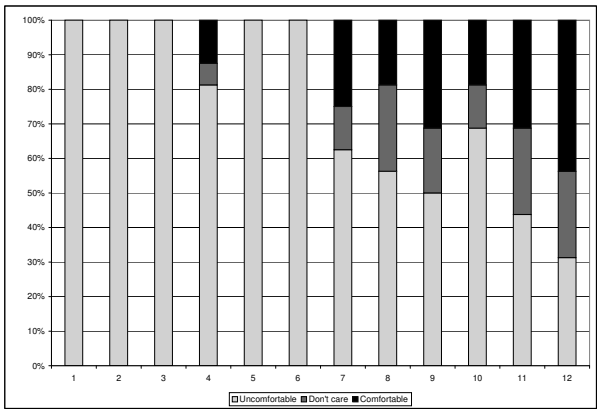

Figure 3. The results for questions about the availability of information on a child for which the respondent was a caretaker.

Perhaps not surprisingly, the respondents were consistently uncomfortable with having information about a child available through a search engine. Everyone was uncomfortable with having identifying information about a child available (home address, phone number, maps, age, passport-like photo). While the majority were uncomfortable with having an email address available, it is interesting to note that this was not the case for all respondents. Although the difference was not

shown to be significant ($p = 0.2251$[1]), there was one person did not care if this information was available and two who were comfortable with the information being available. (It is also interesting to note that the two who were comfortable with this are both parents.)

While the majority of respondents were uncomfortable with any information regarding a child being made available (with the exception of a fictional story that the child has written), there were some respondents who were either comfortable or did not care if non-identifying information were available. Such information includes the child's hobbies and favourite movies. Interestingly, four respondents were comfortable and two did not care if a photo of their child receiving a prestigious award was available through a search engine. However, these same respondents were uncomfortable with a passport-like photo of their child being available.

The third set of questions focused on any extra-curricular activities of the respondent, such as memberships in various organizations and clubs. Again, the questions focused on the comfort level of the respondent knowing that this information is available through a search engine. The organizations were:

1. political group (e.g. political party, lobby group)
2. group that some may find controversial (e.g. gun club, Greenpeace)
3. charitable group that may raise suspicions about your personal life (e.g. Mothers Against Drunk Driving, John Howard Society)
4. leisure group or club (e.g. softball, orchestra)
5. religious group (e.g. Knights of Columbus, United Synagogue Youth)
6. financial group (e.g. investment club)
7. charitable group (e.g. hospital fundraising committee)

The results are illustrated in Figure 4.

Consistent with the answers to the first set of questions, the majority of respondents were uncomfortable with having their memberships in religious or political organizations known. Respondents were also uncomfortable with information being available about any affiliation that might be considered controversial (e.g. gun club, Greenpeace), as well as any that might reflect on their financial status (e.g. investment club). However, few people were uncomfortable with their membership in leisure clubs or charities being known, and most did not care if this information was available online.

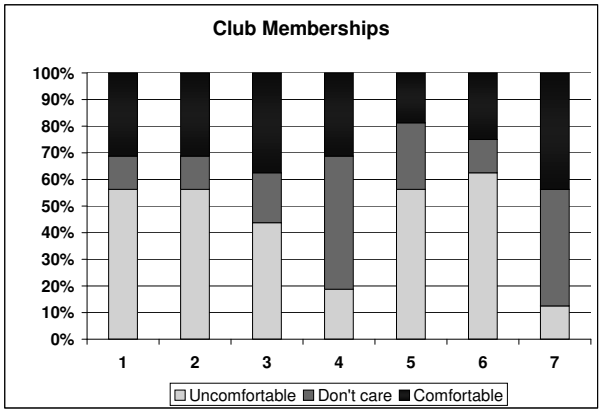

Figure 4. The results from questions on the availability through search engines of information on clubs to which the respondent belonged.

The fourth set of questions focused on how long some personal information should be available through a search engine. In particular, the questions focused on home address, resume and marital status, with the results presented in Figure 5. The answers provided for how long a person's address should be available through a search engine were inconsistent in some cases with the responses from the first set of questions. As reported above, in the set of questions on comfort level with personal information, 12 people responded that they were uncomfortable with their home address being available through a search engine. However, in this set of questions about duration, only seven of these people stated that this information should never be available. The other five people stated that the information should be available for either up to three years (one person said 6–12 months, and two said 1–3 years) or until the source was changed (one person) or deleted (one person).

Not surprisingly, the majority of respondents felt that their resume should be available through a search engine ($p = 0.01242$), however some felt that it should be available for up to three years (nine people), while others thought that its availability should be based on the source being modified or deleted (four people). Interestingly, marital status was a particularly private issue, with the majority of respondents commenting that it should never be available through a search engine.

The fifth set of questions centred on understanding how long people felt that public postings of opinions

[1] All p-values were calucated using a χ^2 test of independence.

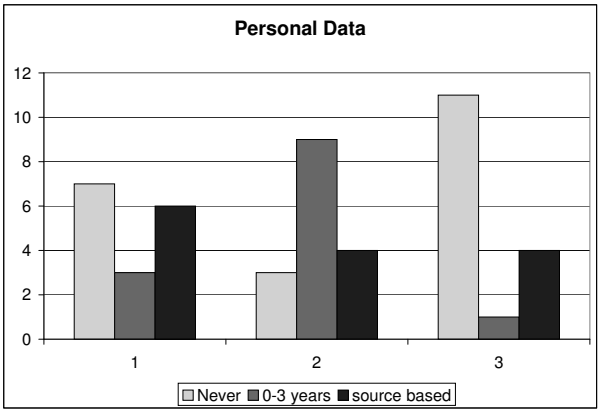

Figure 5. The results from questions discussing the length of time for which some personal information should be available through search engines. The questions were regarding (1) home address, (2) resume and (3) marital status.

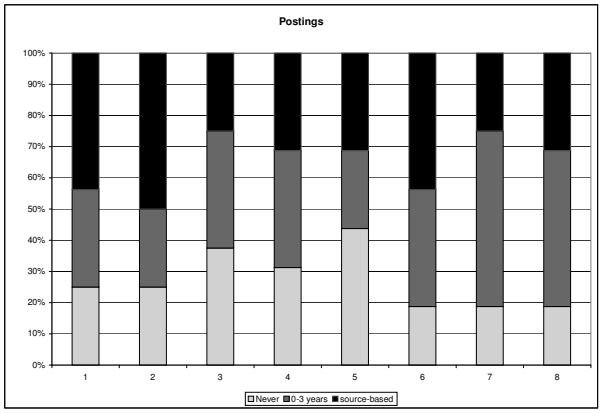

Figure 6. The results from questions discussing the length of time for which on-line postings should be available through search engines.

or activities should be available. The questions asked were:

1. a journal entry about your pet
2. a journal entry that you wrote anonymously about your garden
3. a newsgroup posting that you wrote that contains opinions about welfare that are opposed to your current opinions
4. a newsgroup posting that you wrote anonymously that contains views about immigration that you no longer agree with
5. a journal entry that you wrote anonymously about your promotion at work
6. a journal entry about your winning a prize for community service
7. a newsgroup posting that you wrote about a sensitive issue (e.g., abortion)
8. a newsgroup posting that you wrote anonymously about a personal problem (e.g. depression)

The results are available in Figure 6.

Most respondents stated that journal entries should remain available until the source was deleted (with a smaller number prefering one to three years). The one exception was a journal entry on a promotion at work. Even though written anonymously, many respondents felt that this information should never be posted. The two other types of postings that many users felt should not be available were opinions with which they no longer agreed, whether posted anonymously or otherwise. However, while many users felt this information should not be available, they were not a majority, with nearly equal numbers stating that the information should be available for up to three years, and slightly fewer people stating the information should remain available until the source was deleted or changed. In contrast, newsgroup postings, even on sensitive issues, with which the respondent still agreed (e.g. newsgroup posting on abortion, or about a sensitive personal problem such as depression) were considered to be information that should remain available, with only three people in each case stating that that information should never have been posted ($p = 0.01242$). In both cases, the majority of respondents felt that the information should be available for up to three years (nine and eight people, respectively), with fewer people (four and five people, respectively) stating that the information availability should be based on the availability of the source.

The sixth set of questions focused on the availability of photos of the respondent. Respondents were then asked how long they thought that such visual representations should be available through a search engine. The different types of photos described were:

1. an unflattering picture of you (e.g., you have messy hair or a funny expression)
2. a picture of your face (like a passport photo or other "headshot")
3. a picture of you receiving a prestigious award
4. a picture of you in which you cannot be

identified (e.g., your name is not present)

The results are provided in Figure 7.

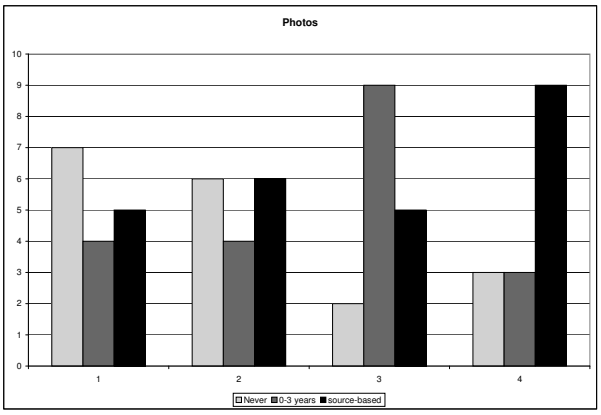

Figure 7. The results from questions discussing the length of time for which pictures of the respondent should be available through search engines.

Responses were mixed on the first two types of photos being available — both unflattering photos (e.g. with messy hair) and passport-like photos. In the first case, seven people thought this should never be available, versus four who said that it should be available for up to three years, and five who said that it should be available based on the availability of the source. For the second questions, the responses were six, four and six, respectively. In both of these instances, there were responses that were inconsistent with responses from the first set of questions. For the unflattering photo, there had been nine people who were uncomfortable having this information available, yet only five of them stated that this information should never be available, while one said that it should be available for up to six months, and three said that it should be available until the source was deleted. For the passport-like photo, there were seven people who were uncomfortable having the information available, five of which then said that this information should never be available. One person said that this information should be available for up to six months, while a second person said that this information should be available until the source is removed.

In contrast, the number of respondents who said that photos of them receiving a prestigious award, or in which they can not be identified, can remain on the web was significant. Only two respondents for the first case felt that this information should never be available ($p = 0.0027$), and only three in the second case ($p = 0.01242$). In the first case, the majority of respondents felt that the photo should be available for up to three years, while in the second case the majority felt that the information should be available until the source was removed.

Figure 8 illustrates the results where the respondents were given four statements and asked to indicate if they agreed with the statement or not. The statements were:

1. I believe it is acceptable for all online information to be stored in a permanent archive.

2. I believe that some types of information should not be available through search engines.

3. I believe it is acceptable for anonymous online information to be stored in a permanent archive.

4. I believe that certain information should not be published on the Internet.

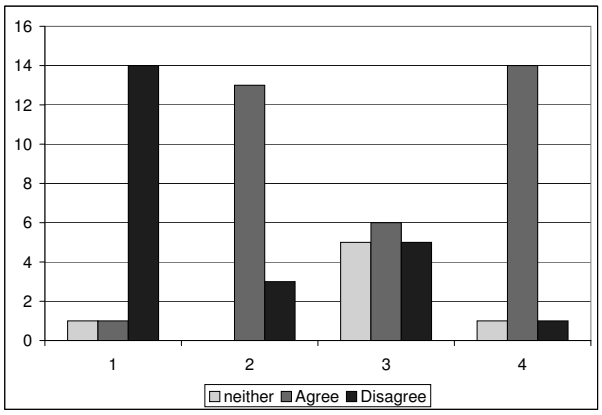

Figure 8. The results from questions that capture the overall beliefs of a respondent regarding what information should be kept available on-line.

For three of the four questions, respondents showed significant agreement amongst themselves on their agreement with the statements provided. The majority disagreed with the statement that it was acceptable for all on-line information to be stored in a permanent archive ($p = 0.0027$). In contrast, the majority agreed that some types of information should not be available through search engines ($p = 0.01242$), and that certain information should not be published on the internet ($p = 0.0027$).

However, for the third question in this series, there was considerable disagreement amongst the respondents, with five people agreeing with the statement, five people disagreeing with the statement, and six people neither agreeing nor disagreeing. This indicates that there is some uncertainty in how anonymous information should be handled. Among those who disagreed that this anonymous information should be stored permanently, three people provided examples of the types of information that they felt should not be stored. These examples included hate literature, if the context identifies the person, and patient data or personnel files. Interestingly, one respondent who agreed with the statement, then provided home addresses and social insurance numbers as examples of information that should not be available.

4 Discussion

Due to the small number of respondents for this study, there are very few questions for which the responses showed differences that were statistically significant. However, the results do suggest some overall trends, and can be used to improve the survey for future work by suggesting how to focus some of the areas.

There were two overall trends that did show significance. The first was that, out of 15 questions and 16 respondents (for 240 total responses), there were no respondents who stated that information should be available forever. This is consistent with the statements by respondents that information should *not* be permanently available, as shown in question one in Figure 8. There were 14 people who agreed with this statement, for $p = 0.0027$.

The second overall trend was that, for those responses where people felt that information should be available, only one respondent on only one question specified that information should be available for more than three years. Of the possible 240 responses, there were 74 that specified that the information should never be available, 81 where the information should be available for less than 3 years, 1 where the information should be available for 4-6 years, and 84 where the information should be available until the source changed. This suggests that, in the cases where the availability of information is not based on the availability of the source, that information should not be available through a search engine for more than three years.

In general, people considered information about their home contact information, religious and political affiliations, financial information and marital status to be private information. The majority of respondents stated that they were uncomfortable having this information available, or that this information should never be available.

In contrast, none of the respondents were uncomfortable with having their occupation known. In general, they also wanted their resume to be available, with only four people expressing that they would feel uncomfortable having this information available. Three of these four stated that this information should never be available through a search engine, while the fourth stated that it should not be available for more than six months. It is interesting to note that none of these four respondents have ever had a web page, and so might consider having their resume on-line as having been placed there by some other person (e.g. their employer), rather than being posted of their own volition.

The responses regarding the availability of personal information on children aged 8–12 was very consistent, with the majority of respondents stating that they were uncomfortable having any information that identified a child available through a search engine. As the information became less likely to identify a child (e.g. such as fictional stories written by the child, or the child's favourite books or movies), respondents were more likely to state that they were comfortable making this information available.

In general, respondents felt that journal entries either should not be available, or should only be available as long as the source was available. This might be a reflection of the fact that journal entries are generally controlled by the respondent, and therefore if the journal entry is removed, it should also no longer be available through a search engine. In contrast, the majority of respondents felt that newsgroup postings should be available for up to three years. This perhaps represents how long the respondents feel static information such as this should be available before it expires.

Interestingly, no consistent view emerged of how photos of the respondents should be treated. Some felt that unflattering photos or passport-like photos should never be available, while others felt that they should be available for up to three years. Some were comfortable with this information being available, others were not. However, respondents felt that photos of them receiving an award should be available for up to three years. This is perhaps consistent with respondents' views on how long newsgroup postings should be available. That is, perhaps they view such photos as static information provided by someone else (e.g. a newspaper), and so feel that it should be available for some amount of time that they apply consistently to all static information. Similarly, they felt that any photo in which they were not explicitly identified could remain available until the

source was removed.

Respondents expressed very consistent views on the availability of information that was not anonymous. They consistently felt that information should not be permanently available ($p = 0.0027$). In addition, they felt that there was some information that should not be published on the internet ($p = 0.0027$), and that there was some information that should not be available through search engines ($p = 0.01242$).

However, they expressed no consistency in how anonymous information should be handled. When asked if anonymous information should be permanently stored, there were equal numbers of responses agreeing with the statement, disagreeing with the statement, and stating that they neither agreed nor disagreed with the statement. These contradictory views are also shown in the section on newsgroup and journal postings, where anonymous postings received a variety of responses, apparently based primarily on the content of the message rather than the anonymous nature of it.

5 Related Work

The Graphic, Visualization, and Usability (GVU) Center at the Georgia Institute of Technology conducted 10 user surveys from January 1994 until October 1998. The surveys were intended to capture the changing attitudes of web users. While some of the questions were focused on the issue of privacy, this was set in the context of web pages collecting information on users, mass emailings, and privacy of communications (e.g. encryption), for example, and so privacy in the context of information easily available through search engines was not addressed. However, one of the questions in our survey (state if you agree or disagree with the following statement: "I believe that certain information should not be published on the Internet.") was taken directly from GVU's 10th WWW User Survey [5]. Surprisingly, our results are considerably different from those found in the GVU survey. The GVU survey found that respondents were split, with 46.6% of respondents agreeing with the statement, 44.0% of respondents disagreeing with the statement and 9.4% not expressing an opinion. Conversely, our survey found that the majority of respondents agreed with the statement (14 of the 16 respondents agreed), for a significant difference ($p = 0.003965$). This perhaps is indicative of some of the demographic characteristics of our population, all of whom deal daily with issues surrounding censorship and information as part of their profession.

The survey that matches most closely to our survey was performed in 1999 by Cranor et al. [2]. This study examined the attitudes of internet users towards privacy; however, the focus was on the information that users felt comfortable providing to a website during some transaction, and what characteristics might influence that comfort level (e.g. having a privacy policy posted, having a seal of approval from some third organization). In contrast, we are looking at information that is available publicly, rather than provided privately to a second party.

However, despite the difference in goals between the two studies, some similarities did emerge. In [2], the authors found that respondents felt uncomfortable providing their phone number, but felt comfortable providing their email address. This is consistent with our results, which found that approximately 60% of respondents were uncomfortable having their phone number available through a search engine, but that less than 15% were uncomfortable having their email address available. When our respondents were provided with a checklist of potential concerns about having their email address publicly available, 14 of the 16 respondents indicated that receiving junk email was an issue. This is consistent with the finding by Cranor et al. that respondents did not like unsolicited communications, and were unlikely to provide their email to a company who would share that information with other companies to send marketing material.

Cranor et al. [2] also asked respondents to how comfortable they would be providing various types of personal information. These questions were asked for both the respondents, and for children between the ages of 8 and 12 for whom the respondent was a caretaker. They found that respondents were consistently less comfortable providing information about children, which is consistent with our findings. Cranor et al. also found that respondents were not comfortable providing their phone numbers (11% were comfortable) or income (17%), but were comfortable providing their age (69%) and email address (76%). This is also consistent with our findings.

6 Future Work

Based on the results from this pilot study, we intend to pursue a larger study of the more general internet population. We also intend to improve the survey, based on some of the results. For example, we will investigate if privacy attitudes differ depending on whether the user posted his or her own personal information, or whether someone else posted it. In addition, we will be more specific about available material from government sites (e.g. salary information for some people, property assessments). This will likely increase the

length of the survey, as we suspect that the source of the material will affect people's perceptions of how private it is, which will require numerous questions for full exploration.

While the number of respondents to the survey were too small to allow any meaningful discussion of the differences in responses between those who had personal web pages and those who did not, this is an area on which we would like to follow up in the next study. That is, are the responses to the scenarios influenced by whether or not a person already has (or has had in the past) a web page?

Finally, in the next phase of the study, we intend to include some of the same demographics questions used by Cranor et al. in [2]. This will allow us to better compare our results to theirs on some of the questions for which there is an obvious relationship or overlap. This will hopefully continue to show support for certain statements about what information people deem to be private.

7 Concluding Remarks

In this paper we presented the results from a pilot study on people's attitudes towards privacy and on-line information available via search engines. This study has shown that there are some areas in which there is clear agreement among our sample population, and has provided insights into how to best further this study. As a result, we will be modifying some of the study questions and performing the next phase on a larger selection of internet users.

From this sample population, however, we have found that people do not want to have web-based information permanently stored. This is the case even when the information is not particularly personal (e.g. a journal article about your pet, or an anonymous posting about your garden). In the case of data that was provided anonymously, there was sharp disagreement among respondents on whether this information should be permanently stored. In general, users felt that data should not be available for more than three years.

There seemed to be relative consistency amongst respondents on what they considered to be personal information. In particular, location information (such as address or phone number), and religious and political affiliations were considered personal and so should not be available through a search engine. There was also consistency in information that respondents were comfortable in having available, which centred around items such as occupation and resumes.

Based on these initial results, the authors intend to pursue a larger study of the more general internet population to determine if the views expressed by the respondents are representative of the larger population.

Acknowledgments

The authors would like to acknowledge the IBM Canada Center for Advanced Studies and the National Sciences and Engineering Research Council of Canada for supporting this research. Thanks also to Jack Duffy (Dalhousie University) for advice on research methods, Josh McNutt (Carnegie Mellon University) for advice on statistical methods, and to Bill Maes, Joyline Makani, and Judy Dunn (Dalhousie University) for their assistance in surveying the library community.

References

[1] Associated Press. Google seeks consensus on privacy issues. *Toronto Globe and Mail*. March 23, 2004.

[2] L. F. Cranor, J. Reagle, and M. S. Ackerman. Beyond concern: Understanding net users' attitudes about online privacy. Research Technical Report TR 99.4.3, AT&T Labs, April 1999.

[3] U. Gattiker, L. Janz, M. Schollmeyer, and H. Kelley. The privacy project survey: Results. `http://www.sci.tamucc.edu/~martinas/Survey/results.html`, 1996. Last visited: 9 June 2004.

[4] I. Goldberg, D. Wagner, and E. A. Brewer. Privacy-enhancing technologies for the internet. In *IEEE COMPCON 1997*, 1997.

[5] Graphic, Visualization, and Usability Center. Gvu's 10th www user survey. `http://www.gvu.gatech.edu/user_surveys/survey-1998-10/`, 1998. Last visited: 14 June 2004.

[6] J. Kleinberg, C. H. Papadimitriou, and P. Raghavan. On the value of private information. In *Proceedings of the Eighth Conference on Theoretical Aspects of Rationality and Knowledge*, 2001.

[7] Ponemon Institute. Privacy trust survey of the United States government. `http://cioi.web.cmu.edu/research/2004PrivacyTrustSurveyoftheUnitedState%sGovernmentExecutiveSummaryV.6.pdf`, 2004. Last visited: 14 June 2004.

[8] H. R. Varian. Economic aspects of personal privacy. `http://www.sims.berkeley.edu/~hal/Papers/privacy/`, 1996. Last visited: 9 June 2004.