Self-monitoring of Web-based Information Disclosure

Kulsoom Abdullah
Georgia Institute of Technology
Atlanta, Georgia
kulsoom@gatech.edu

Gregory Conti
United States Military Academy
West Point, New York
gregory.conti@usma.edu

Edward Sobiesk
United States Military Academy
West Point, New York
edward.sobiesk@usma.edu

ABSTRACT
Free online tools such as search, email and mapping come with a cost. Web users obtain such services by making micropayments of personal and organizational information to the web service providers. Web companies use this information to create customized advertising and tailored user experiences. Individually, each transaction appears innocuous, but when aggregated, the result is often highly sensitive. The impact of AOL's inadvertent disclosure of 20 million nominally anonymized search queries underscores the pressing need for increasing web privacy and raising user awareness of the problem. Rather than advocate extreme legal and policy measures to address the dilemma, this paper proposes an equitable self-monitoring solution. Self-monitoring allows individual users and large enterprises to regulate their web-based interactions intelligently and still allow online companies to innovate and flourish. The primary contributions of our work includes exploration of visualization techniques that support self-monitoring, a human-centric evaluation and the results of a user requirements survey.

Categories and Subject Descriptors
H.3.3 **[Information Search and Retrieval]**: Information Search and Retrieval, K.4.2 **[Computers and Society]**: Social Issues, K.4.3 **[Computers and Society]**: Organizational Impacts, K.4.4 **[Computers and Society]**: Electronic Commerce.

General Terms
Security, Human Factors

Keywords
query visualization, googling, information disclosure, search, privacy, anonymity, Google, AOL, Yahoo, MSN

1. INTRODUCTION
Since the creation of the World Wide Web in the early 1990's, web users have accessed billions of websites generating a continuous stream of interaction data. As electronic commerce began to rise to dominance on the web, many companies found successful business models by providing free online services paired with advertising. Initially starting with naively targeted advertising, companies soon learned to exploit the information users provided in order to more accurately profile individuals, and hence generate more effective advertising. This model has flourished, and we now see a wide variety of powerful, free online tools available to web users.

The sensitivity of web-based information disclosure associated with such tools received its greatest media attention in August 2006 when AOL inadvertently released a search query dataset containing more than 20 million searches by 657,426 AOL users. Many of these queries included sensitive information such as medical conditions, addresses, business dealings and other exceptionally personal information. Despite being nominally anonymized (AOL user names were replaced with numbers), New York Times reporters Michael Barbaro and Tom Zeller quickly demonstrated the ability to move from these "anonymized" search queries back to real world individuals [1]. Shortly thereafter, a number of collaborative analysis websites appeared, seeking to analyze, and in some cases identify, users [2,3,4].

Unfortunately, web users are largely unaware and insensitive to the significant privacy risks that their aggregated information disclosures pose. Despite the initial media attention and public outcry against AOL, several months after the event we found that 84% of the college undergraduates we surveyed were completely unaware of the AOL data spill, and 81% of these undergraduates admitted to having conducted searches for information they would not want disclosed to their current or future employer [5].

Rather than advocate extreme legal and policy measures to address this problem, we propose the development and proliferation of self-monitoring tools as an equitable first step to mitigate this ever increasing privacy challenge. Self-monitoring would allow individual users and large enterprises to self-regulate their web-based interactions intelligently and still allow online companies to innovate and flourish.

The primary contributions of our work include exploration of visualization techniques that support self-monitoring, a human-centric evaluation of these techniques, and the results of a user requirements survey. Ultimately, the uniqueness of our approach springs from our focus on efficient and effective visualizations for self-monitoring of web activity.

2. RELATED WORK
Self-monitoring of web-based information disclosure is largely an unexplored area. In 2003, Battelle first popularized the importance of search queries in his "Database of Intentions" writings [6,7]. In 2006, Conti more formally codified the threat in "Googling Considered Harmful" [8].

Currently, only one researcher we have discovered has created a search query specific visualization [9]. Entitled SearchClock, the visualization plots the most frequent search terms using a circular format. Each ring represents one week's worth of activity. Search terms are plotted around the circle to mirror an analog 24-hour clock. SearchClock is an interesting initial prototype that focuses on the entire 657,000 user AOL dataset, not on individual or business scale requirements.

WPES'07, October 29, 2007, Alexandria, Virginia, USA.

The most readily available tool currently capable of self-monitoring is the history function included in most modern browsers. However, the history functionality included in today's browsers is designed primarily to help users locate previously visited websites. It is incapable of providing a coherent picture of information disclosure. Although, as in parents monitoring their children, it can crudely help monitor web surfing activity over short periods of time. Parental monitoring tools such as Net Nanny [10] do offer the capability to log and monitor web activity, but provide primarily text-based reports, again with an emphasis on websites visited and not on information disclosure.

There are a number of browser plug-ins that provide users with additional information regarding their online activities. Page Addict is a Firefox plug-in that shows the user how much time he or she has spent on different websites; reports are available in text list and simple chart formats [11]. Packet Garden plots Internet activity on a globe [12]. It uses a garden metaphor to "grow" plants based on online activity. While aesthetically pleasing, Packet Garden is not designed for efficient self-monitoring because it focuses primarily on the raw amount of data sent and received, not on information disclosure.

A number of commercial tools to monitor web activity, such as Websense [13], are now available. These tools, though, focus on preventing access to websites with undesired content, not on an organized method for user self-monitoring.

There are a number of text-based visualization techniques that are related to our work. While they do not directly address query visualization, they do provide useful insight into text visualization. PaperLens shows the interplay between research topics, researchers and research sources [14]. Lin's "Visualization for the Document Space" provides useful insight into how to visualize and create category groupings [15]. Themeriver is valuable to consider because of its approach to visualizing themes over time [16].

3. SELF-MONITORING VISUALIZATION

The most significant contribution of our work involved empirically exploring different methods of visualizing self-monitored information from web searches. This paper does not present our work in browser, host and network based collection of disclosure data. To conduct our visualization experiments, we carefully selected data from three different users in the AOL data set representing (1) sporadic use, (2) light use and (3) heavy and frequent use of web search. Our strategy was to first create mock-up visualizations for these three representative users and to then to perform an evaluation of those visualizations. By analyzing these mock-up visualizations, we feel we have identified user feedback that will contribute to the design of operational visualization tools. For our visualizations, we used queries and timestamps from the dataset as well as manually added categorical information.

A maximum resolution of 1024 by 768 was used for our visualizations. The amount of queries that can be seen in one screen is limited by this size, and is most significant with the heavy use user.

The primary task we are addressing is that of individual user self-monitoring of web-based information disclosures. To help assess user specific requirements for our visualizations, we conducted a focus group session with 18 undergraduate college students. We deliberately solicited participants from non-technical majors because we believe they are more representative of our projected user base. To help put our work in context, we began the session with a short discussion of the AOL dataset disclosure and our desire to provide the means for users to monitor their web-based information disclosure. After this initial discussion we asked session participants to suggest tasks that our system might facilitate. Suggested tasks included:

- *providing a time-sequence listing of disclosures, preferably including date and time*
- *categorizing and grouping information disclosed by content and destination site*
- *monitoring most frequently used search terms*
- listing most frequently visited sites
- helping monitor cookies, including the number sent per site and the expiration date
- listing the time spent at different websites
- listing their activities at each site
- identifying whether login was required for each site
- providing a way to highlight disclosure of sensitive information
- determining if they had shopped on a given site

Based on this session, and our own assessment, we chose to address the first three (italicized) tasks. In order to further scope the problem we focused only on web search activities, but suggest future work should address all forms of web-based information disclosure, such as online mapping, email, instant messaging and financial transactions.

To facilitate informed self-monitoring we explored showing web search activity over time. In particular, we wished to provide users with the ability to rapidly scan their activities over varying times scales in ways that allow them to self-assess the sensitivity of their aggregated disclosures. While aggregating many user flows into a single, enterprise-level visualization is very relevant future work, we focus here on only a single user.

Ultimately, we explored four different visualization techniques. These were a histogram, a bubble chart, a Windows file Explorer-like hierarchical view, and a Seesoft-based view [17]. Please see Figures 1-4 on the last page of this paper for examples and descriptions of these four techniques.

4. EVALUATION

After we designed our visualization mock-ups, we conducted a 52 user evaluation with college undergraduates. Our primary goal for evaluation was to determine the strengths and weaknesses of our approaches in order to provide a foundation for ourselves and other researchers seeking to provide efficient and effective techniques for self-monitoring.

For our user study we again chose only students that did not major in a technology related field in order to gain insight into more typical, non-technical end users. Both qualitative and quantitative questions were used. The questions addressed the following three categories.

Questions on the usefulness of the visualizations:

- What visualization is best for allowing self-monitoring of your online search activities?
- Was the visualization easy to understand?
- How effectively could you self-monitor your activity?

Questions related to how much search queries reveal:

- What percentage of the queries would you consider sensitive? Scan and find the most sensitive activities.
- What can you tell about the person based on this query visualization?

Other evaluation questions:

- What is the maximum number of queries this technique can handle before it becomes too crowded or otherwise unusable?
- How does it fare with various realistic time scales?

- How reasonable were our text size and time scale decisions?
- Did truncation matter to users (what length is best)?
- How would the user like to interact with the visualization?

5. ANALYSIS

The surveys were given in an online format, and visualization specific questions were displayed alongside images of each respective visualization technique. There was no testing of interaction functions.

The majority of users thought the visualizations were easy to understand (70%) and the text and time scale decisions were reasonable (75%). Many said the Seesoft view made good use of space versus the histogram, and it maintained context better by including a space for non-active days (73%). Some users wanted to interact with the Seesoft view to selectively remove dates from the display as well as browse and search the entire history of queries. A large percentage said they would like to be able to click on the queries to get more detailed information (47%), such as the specific time the query was made and what the resulting action was after the query submission. The majority liked the explorer view (74%) better than the bubble category view. Those that preferred the bubble chart thought it was more aesthetically pleasing and provided a faster way of seeing category activity. Those that preferred the explorer view thought it was user-friendly, well organized and helped break down the information in a useful way. Being able to see overall categories with the option of viewing the detailed query list was seen as useful. Many thought or alluded that the bubble charts wasted space. About half thought that 50 bubbles were too many to view (47%), and half thought the categories made sense (55%). The overall favorite was the hierarchical explorer view (55%), with the bubble chart coming in second (31%), and Seesoft view was last (14%). Since the top visualization design only received 55% of the votes, this could be taken to indicate that future visualization tools should include several display formats. Most users thought that not showing duplicative queries causes some loss of important information (63%). One idea to counter this is to show the query and a number next to it to represent how many times that query was used.

The majority thought it would be useful to monitor their own query activity (86%) to see what information was being revealed, and additionally, for personal information management purposes, liked the ability to go back to queries they already made and access those results again. Looking at these visualizations, the users were able to give accurate assessments and opinions on the AOL users despite their initial unfamiliarity. Some additional suggestions were made including: allowing visualization of queries by different parameters, such as frequency, time and date in the same view as well as letting the user know how much time was spent at a particular site.

Our results show that we were successful in meeting the requirements of providing a time-based view of disclosed queries, categorizing queries, and monitoring more frequently used search terms as well as raising user awareness overall. The next step will be creating a functional prototype after a design is fine-tuned. One likely approach would be a plug-in for the individual user and the other a stand-alone application for the enterprise level. At each stage, the best visualizations would be selected and refined.

6. CONCLUSION

In this paper we analyzed a number of visualization techniques that allow users to self-monitor their information disclosure. We feel we have laid the groundwork for future visualization and interface designers. We believe that self-monitoring is a powerful tool that raises awareness to the threat and empowers both individuals and enterprises to regulate the amount of information they disclose, while, at the same time, only minimally impacting web services.

In the future, we see the potential for widespread deployment of self-monitoring technologies for both individual browsers and stand-alone enterprise level appliances. For future work, we plan to seek additional self-monitoring visualization techniques as well as interaction improvements. We have focused on self-monitoring at the user level, but a logical next step is to extend our research to include self-monitoring of enterprise scale datasets. Finally, our visualizations provided satisfactory results, well beyond the current browser history function, with the Windows Explorer-like hierarchical visualization being most favored.

7. REFERENCES

[1] Michael Barbaro and Tom Zeller. "A Face is Exposed for AOL Searcher No. 4417749." New York Times, 9 August 2006.

[2] AOL Stalker. http://www.aolstalker.com/, last accessed 14 March 2007.

[3] AOL Search Logs. http://data.aolsearchlogs.com/, last accessed 14 March 2007.

[4] AOLpsycho. http://www.aolpsycho.com/, last accessed 6 February 2007.

[5] Gregory Conti and Edward Sobiesk. "An Honest Man Has Nothing to Fear: User Perceptions in Data Retention." 2007 Symposium on Usable Privacy and Security (SOUPS).

[6] John Battelle. "The Database of Intentions." 13 November 2003. Available online at http://battellemedia.com/.

[7] John Battelle. "The Search." Portfolio Hardcover, 2005.

[8] Gregory Conti. "Googling Considered Harmful." New Security Paradigms Workshop, 2006.

[9] Chris Harrison. "SearchClock: Visualizing Searches over Time." http://charrison.net/projects/searchclock/, last accessed 25 March 2007.

[10] Net Nanny Parental Controls from Content Watch. http://www.netnanny.com/, last accessed 25 March 2007.

[11] Page Addict Project Homepage. http://www.pageaddict.com/, last accessed 25 March 2007.

[12] Packet Garden Project Homepage. http://www.packetgarden.com/, last accessed 25 March 2007.

[13] Websense Corporate Homepage. http://www.websense.com/, last accessed 28 March 2007.

[14] Bongshin Lee, Mary Czerwinski, George Robertson and Benjamin Bederson. "Understanding Eight Years of InfoVis Conferences using PaperLens." InfoVis, pages 53–54, 2004.

[15] Xia Lin. "Visualization for the Document Space." IEEE Visualization, pp. 274-281, 1992.

[16] Susan Havre, Elizabeth Hetzler, Paul Whitney and Lucy Nowell. "ThemeRiver: Visualizing Thematic Changes in Large Document Collections." *IEEE Transactions on Visualization and Computer Graphics, 8*(1), 2002, pp. 9-20.

[17] Stephen Eick, Joseph Steffen and Eric Sumner. "Seesoft-A Tool for Visualizing Line Oriented Software Statistics." IEEE Transactions on Software Engineering, November 1992, pp. 957-968.

of Defense or the United States Government.

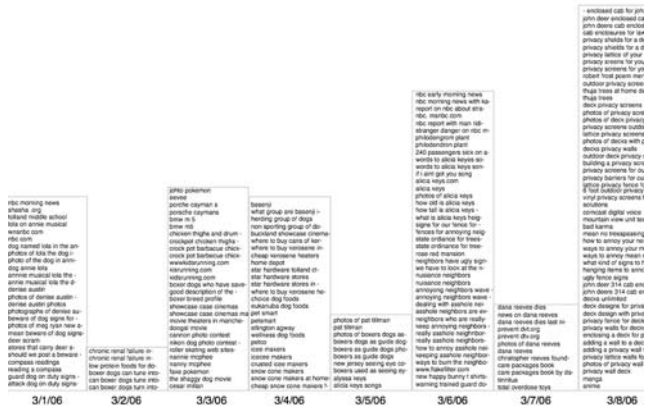

Figure 1. Search term histogram over time using 24 hour increments. This prototype displays search queries on the vertical axis and time on the horizontal axis. This view allows the users to see over a week's worth of search activity at a glance as well as see the magnitude of activity via the height of each column.

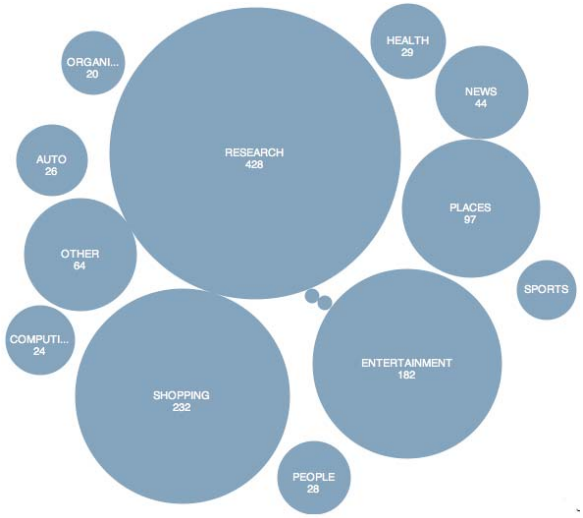

Figure 2. Using a bubble chart to display search activities. This figure depicts search categories as well as the magnitude of related searches, but could also be used to display query terms or query phrases. Because the size of each bubble is proportional to the number of aggregated queries, it uses a great deal of screen space for a relatively small amount of information.

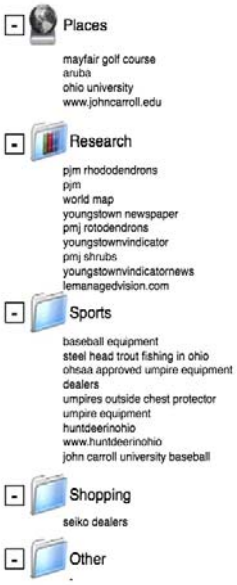

Figure 3: Hierarchical display of user search terms using a visualization similar to the Microsoft file Explorer tool. Search queries were manually grouped by category, but could also be grouped by time or destination website. Because this technique was similar to the familiar Explorer tool, we found that users were readily able to understand its use.

Figure 4. Using a modified Seesoft visualization to view search queries. Search queries were grouped by date and displayed in multiple columns to maximize use of screen space. We chose to include dates with no activity to provide context, but these empty entries could be removed in a future system to help conserve space and provide higher data density.