Researching Real-World Web Use with Roxy, A Research Proxy

Ericka Menchen-Trevino
Northwestern University, emt@u.northwestern.edu

Chris Karr
Audacious Software LLC

Follow this and additional works at: https://scholarworks.umass.edu/jitpc2011

Part of the Communication Commons, Computational Engineering Commons, Computer Engineering Commons, Political Science Commons, Public Affairs, Public Policy and Public Administration Commons, and the Science and Technology Studies Commons

Retrieved from https://scholarworks.umass.edu/jitpc2011/2

This Article is brought to you for free and open access by the The Journal of Information Technology and Politics Annual Conference at ScholarWorks@UMass Amherst. It has been accepted for inclusion in JITP 2011: The Future of Computational Social Science by an authorized administrator of ScholarWorks@UMass Amherst. For more information, please contact scholarworks@library.umass.edu.
Researching real-world Web use with Roxy, A research proxy

Ericka Menchen-Trevino
Northwestern University

Chris Karr
Audacious Software LLC
Abstract
Outside of a lab environment, it has been difficult for researchers to collect both behavioral and self-reported Web-use data from the same participants. To address this challenge we created Roxy, open source software that collects real-world Web-use data with participants’ informed consent. Roxy gathers Web log data as well as the text and HTML code of each page visited by participants. We describe Roxy’s data gathering capabilities and search functions and then illustrate how we used the software in a multi-method study. The use case examines selective exposure to political communication during the November 2010 U.S. general election campaign.
The Web is an integral part of daily life in the developed world. According to a 2010 survey by the Pew Internet & American Life project, 79 percent of American adults use the Internet (of whom virtually all use the Web), and 66 percent have home broadband (Princeton Survey Research Associates International, 2010). Increasingly, networked technologies are so embedded in routine activities that the technology itself seems to disappear from view (Parks, 2009). In contrast to earlier visions of a separate cyberspace, the Internet is now “the utility of the masses” (Wellman, 2010 p. 20). News consumption, long a routine and even ritualized activity (Carey, 1975), has moved online. A 2010 survey of the U.S. showed that 80 percent of online Americans get news online (Princeton Survey Research Associates International, 2010), while a 2008 survey of the city of Chicago (the site of our research project) showed that 91 percent of online Chicagoans access news on the Web (Mossberger & Tolbert, 2009).

Yet it has been difficult for researchers to combine behavioral and self-report data in socio-political investigations of communication technologies. When we looked for tools to collect real-world Web behavior data for a multi-method project, the software was either not

---

1 This software project was possible due to a grant from the Department of Communication Studies at Northwestern University, and the guidance and support of James Ettema. Grants from Northwestern’s School of Communication and The Graduate School supported the research we describe in this report. The authors would also like to thank Robin Hoecker for her assistance with the abstract.

2 We use the first person plural to avoid awkward grammatical constructions, however, the research project that prompted the creation of the software and serves as the case study herein was conceived and executed by Ericka Menchen-Trevino, while the software architect and developer was Chris Karr.
available for use (trade secret software developed for consulting or ratings purposes), or not designed to adhere to the ethical principals of informed consent (software designed for workplace employee monitoring or parental control of children). This prompted us to create Roxy, a research proxy that can collect real-world Web behavior for a wide variety of research goals adhering to principals of informed consent.

Researchers have examined the Web’s role in political information gathering and political activity using different methods. Many of these studies use nationally representative sample surveys, such as the reports from the Pew Internet & American Life Project and the National Annenberg Election Survey. The disadvantages of surveys that ask participants to report their media use is that respondents have a great deal of difficulty reporting use accurately (Price & Zaller, 1993). Reporting media use is cognitively difficult, particularly in a rich environment of ubiquitous media. A recent report shows that 46 percent of American adults get news from 4 or more media platforms regularly (Purcell, Rainie, Mitchell, Rosenstiel, & Olmstead, 2010). Furthermore, self-reports of media use may be influenced by social desirability, where consuming news, for example, is generally seen as desirable, while watching reality television is not. This leads to skewed estimates in surveys (Prior, 2009).

Other studies of political information gathering online rely on ratings data from companies such as Nielsen, comScore or Hitwise (for example Hindman, 2009; Tewksbury, 2003). Using this type of data eliminates the problem of respondents’ limitations in describing their online behavior, but it creates other challenges. Ratings companies keep the identity of their participants confidential. Investigators who use this data generally do not have the opportunity to contact the people whose data they are investigating to investigate their other behaviors and motivations. Often ratings data is provided in aggregate form or limited to pre-determined Web
sites and does not provide a comprehensive picture of individual-level browsing. Furthermore, this ratings data does not include an archive of the text and HTML viewed by participants, so the exact content accessed by the individual is often uncertain. The software used to collect ratings data is typically treated as a trade secret so it is often unclear exactly how the information is recorded, and some aspects of the participant recruitment process are also opaque. This prevents the methodological transparency needed for research replication.

Other researchers have studied online political behavior in a lab setting (Iyengar & Hahn, 2009; Knobloch-Westerwick & Hastall, 2010; Menchen-Trevino & Hargittai, 2011). This approach allows the investigators to combine self-reports and observed behavior in a transparent and replicable manner. The primary drawback of this approach is that lab experiments or observations are different from real-life Web use. Participants are in an unfamiliar location using the lab computer and working under some type of time constraint. They are often assigned a task they would not ordinarily choose to perform.

Roxy allows investigators to observe and record participants’ real-world Web use. Researchers using Roxy have full control over participant recruitment, selection, and the data gathering process. The data collection effort is comparable to survey and lab-based projects. Every research design involves trade-offs. Roxy gives Web researchers another tool with which to tackle such decisions. The basic proxy server technology that Roxy uses is widespread and has existed for quite a long time. Our application of proxy servers in the research data collection process demonstrates an innovative use of this technology.

Below we discuss our development goals for the software and the technologies we used to reach them, including the refinements we made through testing. Next we describe the user’s

---

3 Roxy’s name is a blend of the words research and proxy.
experience before we detail the researcher’s work flow. Finally we describe our study where we collected eight weeks of Roxy data from 41 participants in a multi-method study that combined Web-use data with surveys and in-person interviews.

Development

Roxy is designed to balance the requirements of informed consent and privacy with the need for comprehensive Web-activity data from participants. While Roxy’s first deployment focused on online news and political information consumption, we designed and built the software to address a wide variety of Web data collection needs.

Roxy is a Web proxy server. A proxy server is a software program that acts as an intermediary in the exchange of requests and content between a respondent’s Web browser and the Internet. When a user’s browser is configured to use a proxy server, the browser submits requests for content to the proxy server, and the proxy server fetches the content on the browser’s behalf. When the proxy receives the content from the originating source, it returns it to the browser and the browser presents the content to the user. Proxy servers have been used to solve a variety of problems, such as bandwidth optimization, content filtering, or anonymization of Internet traffic.

Proxy servers are a proven infrastructure technology with wide technical support among the many platforms and systems that connect to the Internet. Using a proxy server allowed us to collect rich information while avoiding the logistical challenges of developing, installing, and supporting custom software on our participants’ computers. We provided them with simple instructions to configure their existing browsers to connect to the Roxy server. After following those few steps, they joined our data collection system without installing any software on their computer.
Since the existing proxy programs are predominately designed to transfer and filter content efficiently, we constructed our own in order to implement the rich data collection and provide effective informed consent and privacy protections for our participants. Roxy differs from other proxy packages in the following ways:

- Roxy keeps a full record of the text content that participants request. This textual content provides the corpus that allowed us to interview participants about the specific content they requested or perform any form of textual analysis. In addition to the text of visited pages, we also collect the URL, the referring page and the date and time of the content access.

- Roxy allows participants to specify pages or sites to blacklist from our system. This blacklist is used to protect participants’ privacy by allowing them to opt out of our study on a fine-grained level and to be confident that we are not collecting private information like bank account numbers and Web-based e-mail messages.

- Roxy provides participants a mechanism to review their browsing history and remove information after-the-fact. This is an additional tool to protect participant privacy.

- Roxy includes a system-wide blacklist that allows researchers to proactively block irrelevant sites using existing blacklists and site categories. We adapted others’ lists to avoid collecting data when participants visited adult sites, banking Web pages, and personal social networks.

- Roxy does not attempt to log any encrypted (HTTPS) content.

In summary, Roxy archives copies of the content that participants request for later analysis, but balances this potentially intrusive activity by giving participants a rich set of tools.
to review the gathered information, and to exercise a level of control over whether that information becomes part of our dataset. No existing proxy server program provided both the tools for rich data collection, as well as robust privacy management features.

We used a variety of technologies to build Roxy. As mentioned above, we did not require participants to install custom software on their own computers, so our efforts focused on building the central proxy server. The core of the server uses the Twisted framework for building Internet applications (Twisted Matrix Labs, 2010). Twisted provided the infrastructure for managing the proxy traffic. We used a MySQL database to store the collected information and operational data like the usernames and passwords used to identify participants (Oracle Corporation, 2010). In addition to MySQL, we also used Apache Solr to index the textual content (The Apache Software Foundation, 2010). Whereas MySQL’s text searching and analysis functions are fairly basic, Solr allowed us to efficiently search the large amount of information we collected using a rich query language.

We protected participants' privacy and maintained the highest level of data security by hosting Roxy on its own private dedicated host. Only the Roxy administrators had access to any personal identifiable information. We supplemented the site's privacy protections with a robust set of intrusion detection and access denial tools that prevented unauthorized third parties from accessing the server. When Roxy is used in this manner, the data that it captures is secure at all levels. Using Roxy does not expose the participants’ Web data to any additional vulnerability beyond what they would encounter in routine Web use. After data collection is complete, the researcher must then store the resulting dataset securely to protect participants from privacy violations.
The participant-facing portions of the server are implemented as a small custom Web application integrated into the Twisted infrastructure. This Web application includes instructions for participants to configure their systems, a small testing tool to determine if the setup was successful, and the tools described above for managing privacy. Roxy administrators (the researchers or programmers) use a private administrative site to review how many participants were active at a given moment, as well as search tools to review the data collected, described further below.

When a participant makes a request through the Roxy server, the following flow is used (see Figure 1):

1. The request is checked against a blacklist on the participant’s computer specified by Roxy’s proxy configuration (PAC) file. If the request matches, the browser does not use the proxy at all (a, b, c).

2. The request is transmitted to the proxy and is tested against the system and user blacklist (d & e). If the participant has not logged into the system, (d) is interrupted with a login dialog and resumes upon successful authentication. A matched request is tagged before being sent to the request manager (f).

3. The request manager fetches the content from the Web (g). If the content is not blacklisted, it is stored in the full-text index (h), and the database (i).

4. The request manager returns the requested content to the participant (j).

{Insert Figure 1 about here}

Refining Roxy

After we completed an initial implementation of the Roxy software, we conducted an internal beta test to evaluate the effectiveness of the system as well as tune the software to handle
the larger population of users in our first study. Since most existing proxy server packages simply relay content between the requestor and the content provider, we used this beta testing period to explore the performance ramifications of not just transmitting content, but also storing it. Our suspicion that the computational costs of storing the content would be a performance issue was correct, and we refined our initial implementation to use a multithreaded design where saving retrieved content was delegated to a collection of worker threads.

In addition to raw performance metrics, we also reviewed the content required by a typical Web page view. Web advertising introduced a situation where a single Web page view could request content from tens of third party advertising sites. For example, viewing the source of the Chicago Tribune front-page in January 2011 reveals that content loads from more than ten non-Tribune domains, the majority of them advertisers. Since banner ads, Facebook “like” buttons, and other auxiliary content was not essential to our first study, we elected not to log these requests and proxy them without saving this content to improve performance. This allowed us to accommodate more participants in our study using fixed resources. Similarly, no image or video files were archived based on their relatively lower utility to the study compared to the high resource cost. Studies with different research goals or funding can customize the software to make different tradeoffs.

Originally, advertising blacklists were implemented on the server side, but we discovered that we could conserve server resources using our proxy configuration file (PAC). A PAC file is a small JavaScript file that tells the browser when to use the proxy and when to communicate with the Internet directly. We originally deployed the PAC file to simplify browser configuration, but later began using it as a blacklist to alleviate the load on our server.
Finally, after observing Roxy in the beta test, we discovered that the large amount of text we were archiving during the study was beginning to degrade the performance of our database because extremely large databases perform more slowly. We measured the performance and activity required to begin degrading the server’s performance, and implemented a regularly scheduled off-hours maintenance process that would collect and archive each week’s data. This process reset the system’s database each week so we could continue to provide a responsive experience to our participants, who were sensitive to the slightest delays introduced to their Web-browsing experience. At the conclusion of the study, we reassembled our weekly snapshots into a complete dataset for analysis.

Roxy from a Participant’s Perspective

Roxy’s user interface is designed to cause participants as little interruption as possible, while still offering them meaningful control over the system. The username and password are saved so users do not have to enter them frequently. Once users select their session type (regular, private, or guest) they are immediately redirected to the page they were requesting before the interruption occurred. This results in a one-click user interaction in most cases, where the one click indicates the participants’ choice of session type (see Figure 2).

After a participant has been inactive for more than 30 minutes, Roxy automatically prompts them to re-authenticate. Participants may visit the Roxy login page (Figure 2) at any time to end their current session and to begin a new session with different privacy preferences. Each of the three session-type buttons display explanations that describe the session types when a participant holds their mouse cursor over an option. This allows users who may not have read or remembered Roxy’s participant instructions to understand how to use the tool based only on the login page itself. The explanatory text for the ‘Regular Session’ button is: “Roxy will record
the sites you view, except for any sites you’ve blacklisted or sites that are secure (start with https and have a ‘lock’ symbol like banking sites).” The text for the ‘Private Session’ reads: “Roxy will not record any information about the Web sites you visit in a private sessions. Please use this option sparingly to help the research project. You can blacklist specific sites using the ‘History & Blacklist’ link on the right.” The ‘Guest Session’ button explains: “If you are not a participant in this research project please use this option so that none of your Web browsing is logged.”

{Insert Figure 2 about here}

Participants can also review the data Roxy has collected from them. The ‘Browsing History’ page shows a chronological list of every page logged, with the option to delete the page from the log, or to blacklist the domain (see Figure 3). A link to the Browsing History page is available on the Roxy login page. Users can also create a personal blacklist of Web sites they do not want Roxy to log.

{Insert Figure 3 about here}

Roxy from the Researcher’s Perspective

Roxy has two features available only to researchers. These features are the User Status page and the Data Explorer. The User Status page shows each username and indicates whether or not that participant has ever logged on, if the participant is logged on now, and the date and time of last login if applicable. The Data Explorer allows researchers to search the full index of content in Roxy (see Figure 4). There are many options for targeting one’s search as depicted in Figure 4. The example illustrated in Figures 4 and 5 shows a query that returns results where the word Obama matches in the title field of the page only for the username ‘test.’ Up to 50 results will be displayed in reverse chronological order according to when ‘test’ accessed these pages, based on the settings depicted in Figure 4. The Data Explorer is a front-end for Solr search server
queries. Researchers who want an even more powerful query can go to the ‘Raw Search Interface’ and construct queries in Solr syntax directly. The flagging feature in the search results helps researchers who need to go beyond simple text searching to manually indicate interest in relevant results by clicking the ‘flag’ option in the search results (see Figure 5).

{Insert Figure 4 about here.}

{Insert Figure 5 about here}

Work Flow

Researchers must acquire a sufficiently powerful server for the project they wish to conduct. Participants are very sensitive to delays in their normal Web browsing and if Roxy significantly impacts browsing speed this will jeopardize the research project because it will cause participants to drop out of the study. Consequently, researchers must accurately predict the number of participants in a study, how active those participants are online, and how well the server fits the participant profiles. We suggest pilot testing one’s server before launching a project to detect any delays in browsing speed.

While selecting participants for a research project using Roxy, there are a few technical issues to consider. Any Web browser that connects to the Internet and allows the user to configure a proxy can use Roxy; however, not all browsers support proxy servers. The AOL Web browser and the AT&T mobile network[^4], for example, do not support custom proxy configuration. Also, participants must be able to configure their Web browser to use Roxy, so those who browse the Web primarily in tightly-controlled environments like offices will pose a challenge.

[^4]: The iPhone, which at the time of this research project was only available through AT&T, allows users to configure a proxy only for wireless hotspots, but not the AT&T mobile network.
After selecting participants and obtaining informed consent, researchers follow this process to deploy Roxy:

1. Create usernames and passwords for the participants.
2. Send the participants instructions on how to connect to Roxy and inform them of their username and password.
3. Verify that the participants have connected by viewing the user status page.
4. (Ongoing) Monitor the server and database performance for hints of a potential overload. Adjust settings as needed. Perform backup and archiving as needed, about once per week.
5. When the study is complete, ask users to remove the proxy setting that connects them to Roxy.

Researchers can search the data during and after data collection using the Data Explorer. Once the data collection is complete, researchers can export the data for analysis in specialized data analysis software packages.

Using Roxy in a Multi-method Research Project

Roxy was designed to collect data for research on the role of technology in selective exposure to political communication. Below, we begin with a brief description of the research project and then discuss some of the preliminary results. We focus on Roxy’s unique contributions to the dataset rather than substantive conclusions because the purpose of this workbench note is to demonstrate the analytical potential of Roxy, not to present research findings, which will be discussed in other publications.

Roxy was deployed for our project from September 6th 2010 until November 5th 2010. This time period includes eight weeks of the U.S. general election campaign and three days after the election on November 2nd. There were 41 total respondents from the Chicago area who
completed the study\textsuperscript{5}. During this time, the participants successfully logged 186,296 Web pages to Roxy\textsuperscript{6}. They also completed 10 online surveys and, after the election, an in-person interview lasting, on average about 1.5 hours. Participants were paid a total of $70 for completing the study.

We took great care in recruiting and selecting participants based on the theoretical concerns of the project. The research design\textsuperscript{7} required participants with different levels of political interest. Given the fairly high respondent burden and the technical requirements for Roxy use, we chose to use a quota sample drawn from a pool of recruits who expressed interest in participating by completing an online recruitment survey designed for this study. Because 629 qualified recruits filled out this survey, we were able to roughly match the U.S. population’s level of political interest\textsuperscript{8} in selected participants, and also select those who owned the computers where they did most of their Web browsing.

\textsuperscript{5} Two respondents who later dropped out of the study used Roxy during most of this time, as well as three test users who were not part of the study, bringing Roxy’s user load to a total of 46 people. To meet the needs of these users, we rented a dedicated server with the following specifications: 2.4 GHz Xeon server equipped with 3 GB RAM & 500 GB disk drive.

\textsuperscript{6} Roxy logged over 360,000 requests. Many requests were components of Web pages, such as advertising, buttons, or widgets and three non-participant test users and two users who did not complete the study also used Roxy.

\textsuperscript{7} The research design is described further in forthcoming publications by Menchen-Trevino.

\textsuperscript{8} This match was based on survey measures of political interest in the American National Election Studies post-election survey of 2008 that were also used in the recruitment survey.
The surveys, Roxy data, and the interviews worked together to inform the project. Before each interview we reviewed the participant’s Roxy data and survey responses. We assessed the Roxy data by searching for terms related to the mid-term elections, politics, and news in general. For example, we searched for local candidate names, as well as national political terms such as “Tea Party” and national political figures like Obama. We also reviewed the surveys before the interviews, which contained several open-ended questions, in part to find additional search terms to use in Roxy that were specific to the participant’s stated interests. During this process we noted the participant’s general pattern of online news access to discuss with the participant during the interview. We also used the flag feature in Roxy’s search results (Figure 5) to highlight any interesting or puzzling pages and brought her computer with her to the interview to review these flagged pages together with participants as needed during the interview.

We were not sure how often participants would use Roxy’s privacy features and how this might impact the data. Most participants did choose a private session at least once (28 out of 41), but few did so frequently. Over all 5,784 sessions, the 41 respondents chose a logged session in 91.1 percent of cases. However, the distribution is highly skewed per user (see Figure 6). While 13 participants chose logged (regular) sessions every time, one person chose only 34 percent logged sessions. The individual with the highest rate of private sessions shared her computer with other family members who mistakenly used the private session option instead of the guest session. The median among users was 98 percent logged sessions. Five participants used the blacklist feature. As requested, none of the blacklisted sites were news sites. In general participants were very cooperative with the data gathering and used the privacy features judiciously.
Roxy’s search functionality allowed us to explore political content wherever it was found. Others have found important political discussions occurring outside of sites defined as news or politics (Wojcieszak & Mutz, 2009), but studies that consider the political implications of non-political Web site use are rare. Data that addresses this issue have been difficult to collect. To illustrate the breadth of our data, we explore the range of different Web sites accessed by participants whose title contained the word “Obama.” This search had 222 results, 189 of which were unique (duplicates are due to the same participant accessing the same Web page more than once during the study). The pages came from 64 different domains that ranged from the very popular, such as yahoo.com and cnn.com, to niche sites like motherjones.com or personalfinancebulletin.com, to obscure blogs like Freedom Eden. This breadth allowed us to explore online political discussion and participation more fully in the interviews, using specific examples from participant browsing history.

Another key feature of Roxy data is that its richness allows for a detailed examination of potential quantitative measures based on Web content. Most of the search results for ‘Obama’ in the Web page title were articles about the president and his policies, such as an article from the New York Times entitled “Obama to Open Offshore Areas to Oil Drilling for First Time.” Initially it seemed reasonable to expect that participants who accessed such pages (30 of the 41 participants) would be those who were more interested in politics, as compared to those who did not access any pages with “Obama” in the title. However, this measure does not correlate significantly with the participants’ survey measures of interest in government, politics, and elections. Looking more closely at the content of the titles the reasons for this lack of correspondence become clear. For example, one participant accessed a page with the title
“Reuse, Renew, Reverse: Michelle Obama's sweater two ways”. Another respondent accessed many pages while searching for the streets that would be closed for an Obama campaign appearance in Chicago. Also, some participants accessed national political news primarily offline via TV, radio, magazines, or talking with others.

Although issues of informed consent for non-participants prevented us from logging Facebook for this study, the referrer field allowed us to observe cases where participants were referred from Facebook to another Web site. We have logged 978 Web pages from 30 respondents with facebook.com in the referrer field. This informed the interviews, which dealt with political information and Facebook in part. Also, the Roxy data alone shows which types of news articles are accessed via Facebook and the spreading of political memes such as, whatthefuckhasobamadonesofar.com, which detailed the president’s accomplishments and was accessed by three participants via Facebook.

We were not sure how participants would respond to questions based on their browsing log. It was unclear how much detail they would remember or if they realized exactly what was logged. Participants responded well to interview questions based on their browsing history. Although they did not recall the details of their actions in some cases, none of them were surprised about what was logged, demonstrating an effective informed consent process. Below is an example of an interview question about a search on a political topic.

Interviewer: Another … search that you did was looking for the Tea Party’s stance on homosexuality.

---

9 The correlation between those with at least two search results with Obama in the title (n=22) and the survey measure of political interest was also statistically insignificant.
Participant: … You know, I’m gay so like their stance on it is a big issue to me…

I actually remember, now, doing that… I think there was some discussion about … they [the tea party] are very like ‘government get out of our lives.’ And I’m like, well I wonder … if they’re true libertarians then they’re like, … ‘people can do whatever they want.’ So, I wanted to see like if they had actually made statements about … how they feel about homosexuality.

This type of information about specific behaviors added very useful context for understanding how political information searching occurs in real-world contexts. The interview question based on the Roxy data was critical in helping the participant remember and discuss the context of the search behavior.

We also asked participants directly if they did anything differently online because of the software. This question was asked near the end of the interview once rapport had been established and specific examples from the browsing data had been discussed. Some mentioned that they did change their behavior somewhat due to the observation. One participant mentioned that he may have avoided a news site he considered low-brow, but he also wanted to portray his behavior honestly:

I was focusing … more heavily on the intelligent stuff [rather] than like the Daily Beast [a website that mixes entertainment with political news]. Then I would go, “Oh, I should go check out the Daily Beast, because I have to be honest about this, I really do look at this.”

We do not know how much he would have accessed the site otherwise, but we see that this participant accessed 77 pages from thedailybeast.com over the course of the study. Most participants said they continued their normal online routines. One said of the software, “It was
not intrusive. It was in the background completely.” Another participant indicated, “I completely ignored that it was ever there.”

Conclusion

The link between researchers and participants has an important ethical dimension. This relationship should be one of mutual trust and respect. When researchers ask participants to use software that monitors their Web use they are asking for a significant amount of trust that it is secure and works in the way the researcher has described. Roxy gives participants full control over their data before and after it is collected to create a reciprocal relationship where participants are both giving and receiving trust. Participants can use their control over their data to harm the research project by deleting information that the researcher might deem useful or interesting. Participants can potentially subvert most, if not all, research methods that rely on human subjects. It is the responsibility of the researcher to build a mutually beneficial and respectful relationship with participants. Our case study demonstrates that this type of relationship is achievable in practice with participants 78 percent of whom were recruited through online advertising, primarily Craigslist.

Roxy enabled us to collect a unique dataset that was directly relevant to our research questions. We believe that researchers in many fields of study may benefit from combining real-world Web data with other methods, including library and information scientists, market researchers, and those who do lab-based studies of Web behavior and want to compare lab with real-world behavior. Roxy could also be used as the only data collection method in research that focuses on general patterns of individual-level Web use.

Most modern Web browsers can use proxy servers. This software provides a way to capture a type of data that was previously inaccessible, and can do so on a wide variety of...
computer platforms. However, if more Web browsing occurs on mobile smart phones in the future, this will present a technical challenge. Roxy is an important first step in what may be a new type of research software for observing real-life online behavior for research rather than commercial purposes.

The development of this tool from concept through launch took less than one year of part-time work by the two authors. We have accomplished a significant amount given our limited resources. We wish to develop Roxy into an even more general and scalable tool for researchers observing Web behavior. Other projects may need more or less data than the study Roxy was originally designed to support, for example, so a simple configuration page that allows researchers choose what data they would like to collect could be developed. Roxy is a data-gathering tool that is not intended to provide full analysis capabilities so we would like to provide a wider variety of export functions to transmit Roxy data to data analysis tools such as qualitative data analysis software, statistical packages, linguistic analysis, and network analysis tools.

We plan to release the Roxy software under an open-source license when we have the resources to complete the necessary documentation and modifications to make the software useful to a wide variety of researchers. Current information about Roxy will be available at http://www.roxyproxy.org/. We look forward to collaborating with other researchers and institutions to further develop this unique tool.
References


Ericka Menchen-Trevino
Northwestern University

Ericka Menchen-Trevino is a Ph.D. Candidate in the Media, Technology & Society program in the School of Communication at Northwestern University. She received her M.A. in Communication Studies at the University of Illinois of Chicago, and her B.S. in Anthropology at Loyola University Chicago. Her research focuses on the role of technology in selective exposure to political communication.

Chris Karr
Audacious Software LLC.

Chris Karr is an independent ubiquitous computing researcher & consultant. He received his M.A. in the Media Technology & Society program in the School of Communication at Northwestern University, and his B.S. in Computer Science from Princeton University. He is the primary software developer for the Roxy program and is the founder and chief developer at Audacious Software.

Correspondence concerning this article should be addressed to Ericka Menchen-Trevino at pubs@ericka.cc.
Figure Captions

*Figure 1*: Request flow diagram

*Figure 2*: Roxy login page\(^\text{10}\)

*Figure 3*: Roxy Browsing History page

*Figure 4*: Roxy data explorer

*Figure 5*: Search results from Roxy Data Explorer

*Figure 6*: Percent of private sessions per user (n=41, 13 never selected a private session)

\(^{10}\) Figure 2 is a partial screen-shot of the Roxy login page. This figure focuses on the login information.
Figure 1
Figure 2

Welcome to Roxy

Thank you for participating in this study. When you choose a session type you will automatically be directed to the web page you were on your way to before Roxy interrupted you. Roxy will ask you to choose a session type when you start using the web after you haven’t been browsing for a while.

Choose A New Session Type

Session Status: Inactive

Username: 

Password: 

- Regular Session
- Private Session
- Guest Session

Session History

Please start a session to retrieve history.

Refresh Session History

Contact Information
Figure 3
Figure 4

Roxy: Data Explorer

Search Parameters

Find: Obama

In Fields: □ Body
          □ Title

Username: test

Access Date: Start Date to End Date

URL: www.news.com

Referrer: www.referrer.com

Search Terms: search term

Max Results: 20

Sort By: Date (Latest First)

Flagged:

Search

Raw Search Interface
### Search Results

<table>
<thead>
<tr>
<th>User ID</th>
<th>Domain</th>
<th>Date</th>
<th>Flag</th>
</tr>
</thead>
<tbody>
<tr>
<td>test</td>
<td>tpmmuckraker.talkingpointsmemo.com</td>
<td>2010-10-30</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>16:01:55Z</td>
<td></td>
</tr>
<tr>
<td></td>
<td>URL: tpmmuckraker.talkingpointsmemo.com /2010/09 /obama_koran_burning_stunt_could_greatly_endanger_t.php</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Search Terms: [&quot;koran burning could endanger troops&quot;]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Request ID: f786f5f9-f2e7-4293-a3a2-9d9bb5b2708</td>
<td></td>
<td></td>
</tr>
<tr>
<td>test</td>
<td><a href="http://www.politico.com">www.politico.com</a></td>
<td>2010-10-28</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>07:24:23Z</td>
<td></td>
</tr>
<tr>
<td></td>
<td>URL: <a href="http://www.politico.com">www.politico.com</a> /blogs/bensmith /1010 /Obama_joins_it_gets_better_campaign.html</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Search Terms: [&quot;it gets better campaign&quot;]</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Request ID: b8386d57-f48a-47b0-9f81-03d526ca110b</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Figure 6

Percent of Private Sessions Per User

0% 10% 20% 30% 40% 50% 60% 70%