Privacy Considerations for Public Storytelling

Christopher Wienberg and **Andrew S. Gordon**
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Los Angeles, CA 90094
{cwienberg,gordon}@ict.usc.edu

Abstract

The popularity of the web and social media have afforded researchers unparalleled access to content about the daily lives of people. Human research ethics guidelines, while actively expanding to meet the new challenges posed by web research, still rely on offline principles of interaction that are a poor fit to modern technology. In this context, we present a study of the identifiability of authors of socially sensitive content. With the goal of identity obfuscation, we compare this to the identifiability of the same content translated to and then back from a foreign language, focusing on how easily a person could locate the original source of the content. We discuss the risk to these authors presented by dissemination of their content, and consider the implications for research ethics guidelines.

Introduction

The popularity of the web and social media has provided a wealth of data for researchers interested in narrative, natural language, social interaction, and many other fields. The amount of content posted publicly provides information at a depth and breadth that researchers even a few years ago could only imagine. For instance, in our lab, we have gathered millions of personal stories posted publicly to weblogs for use in large-scale narrative analysis and knowledge management applications (Gordon and Swanson 2009). A project we have recently undertaken has involved using 40 stories where the author is describing socially questionable behavior he or she has performed (e.g. getting in fights, putting a child up for adoption, quitting a team) as stimuli in an fMRI study (Gimbel et al. 2013). This project has involved linguistic analysis of the 40 stories, examining the impact of qualities such as narrative structure on the neurobiological (fMRI) response of readers.

Given the content of these 40 stories, we became concerned about the attention our project may bring to these authors, and began to contemplate our ethical obligations to these authors. While these 40 stories are posted to public weblogs, many of these blogs are obscure and likely read by only a handful of people. If these 40 stories received greater attention, for instance from acquaintances of the author, it could have serious negative repercussions for the author. Given this possibility, we became concerned that our work may present more than minimal risk to the authors we drew these 40 stories from. We considered the importance of protecting the authors from unwanted negative attention, contemplating issues such as modification of stories to protect author identities, whether we should include snippets from these 40 stories when presenting our research in papers and talks, and if we could ethically distribute these 40 stories to other researchers for use in their own work.

In tension with our desire to protect these web users is our desire to be good stewards in the scientific community. An important part of good scientific research is the ability to reproduce results and build off others' advancements. Other researchers expressed a desire to access these 40 socially questionable stories for use in their own experiments. As scientists, we felt compelled to share our data.

Motivated by the above example, we examined the ethical and scientific obligations of researchers, and the potential trade-offs these obligations require. We delve into the literature on research ethics in social media research, presenting previous examples, existing systems, and the currently recommended best practices. Using the data described in the previous example, we examine the risks of publication, conducting a study to see how easily people can find the original source of this data. Finally, we look at methods to obscure the data's source while preserving its usefulness.

Ethical Social Media Research

A primary concern when conducting research involving human subjects is ethical behavior. Important practices have developed over time to guide and evaluate researchers when working with human subjects, particularly among fields where human subjects research is traditional common—such as medical and psychological research. Safeguards such as institutional review boards are well acquainted with these researchers' legal and ethical requirements.

Unfortunately, these resources have been slow to adapt to changes in how research involving people may be conducted. Improvements in technology have made large populations of human subjects available to researchers, yet the ethical guidelines for researchers often reflect outdated practices. For social media investigators, recommendations and requirements are often derived from analogies to real life that are incomplete or inadequate to capture the issues at stake for studies involving large scale social media datasets. The guidelines that do exist often focus on decisions that users have made in how public their data is. Eastham (2011) developed recommendations for the use of weblog data in the health sciences. In this work, Eastham notes a tension between what is public and what is private in weblogs, and discusses if work with such content constitutes *human subjects research*. In lieu of obtaining consent, Eastham concludes that a blogger's intended level of privacy may be inferred from attributes such as whether the blogger permits comments, uses a pseudonym, or other behaviors that might indicate that the author desires more privacy. However, it is less than clear that these criteria are sufficient, and assessments of risk and what constitutes identifiable private information are left to researchers.

To clarify this, the US Department of Health & Human Services' Office for Human Research Protections released a report with recommendations and clarifying advice on ethical obligations when conducting web research (SACHRP 2013). Unfortunately, this report raises many questions even as it answers others, and its guidance—while an improvement over the previous state of affairs—leaves many gray areas. Particularly, on the issues of *identifiable private information* and *reasonable expectations of privacy*, the report is vague, leaving such judgments to institutional review boards and researchers. This is not merely a theoretical issue. Current research indicates that internet users may be making much more information public than they intend to (Zheleva and Getoor 2009), that users frequently post material they regret (Wang et al. 2011), and that technology and public datasets have made it very easy to infer private information about users (Acquisti and Gross 2009). In our own experience, even cautious users often accidentally reveal information about themselves which can easily be used to identify them. In the simplest case, a user might accidentally forget to anonymize identifying data on domain registrations. In more extreme examples, a user may post photographs from events like foot races, where the user is wearing a race 'bib.' The race number on the 'bib' can be easily used to look up the user's name and hometown on the race organizers' website.

web research ethics are especially important when it comes to socially questionable content, which accompanies increased risk for the content's creator. In this context, questions about what constitutes public information are critical. Even if we accept the position that content creators who post publicly have no reasonable expectation of their content being private, in many cases authors may not expect that content to be amplified and spread far beyond its typical scope of a handful of readers. Take, for instance, the 'FireMe!' project, in which academic researchers filter Twitter's public API for "every twitter update that mentions, inappropriately, the author's working environment" (Kawase et al. 2013). Cast as a project to investigate awareness of the risks of such public behavior, Kawase et al. contacted users they identified as posting inappropriate content about their workplace, informing them that the tweet was publicly accessible and suggesting they delete it. They recorded the responses of users, noting what percentage of users indicated they would delete the content in question, and how many followed through. The researchers also provide a live feed of the content they uncover, amplifying negative content in a way that their own study demonstrates these users do not anticipate. This work is motivated by genuine concerns about internet privacy, and complies with a reasonable interpretation of the current standards for ethical web research. Despite this, it calls the current ethical guidelines into question, particularly when it comes to problematic content.

Identifying Users from Text

Understanding potential risks for users is important to developing informative guidelines regarding the dissemination of questionable content by researchers. To investigate this, we conducted a study to see how easily a person could find the original source of content given a selected subsection.

In our lab, we have been conducting an experiment regarding the impact of narrative framing—such as appeals to certain values or experiences of the listener—on the neurological response of people across different cultures. In this experiment, participants are instructed to read several narratives while their brains are scanned in an fMRI. We chose to use authentic narratives as stimuli rather than write narratives for the purpose of this project. To find authentic narratives, we searched a corpus of millions of English-language personal stories posted to weblogs, using existing technologies (Gordon and Swanson 2008; Gordon, Wienberg, and Sood 2012). While we initially attempted to search directly for instances of sacred framing (the type of narrative framing we were interested in, which features appeals to sacred values like the religion or patriotism of the listener), our search tools were ill-suited to find instances of sacred framing. Instead, we saw success when searching for narratives where the authors would feel compelled to justify their actions with additional context. Narratives of socially questionable actions—such as putting a child up for adoption or getting into a physical fight—were highly likely to exhibit sacred framing. We used 40 such narratives for experiments.

Given the sensitive nature of the content we drew from, we became concerned about potential negative consequences to authors whose content was unwittingly included in our experiment. Particularly, as we began sharing our results with other researchers, we had qualms about sharing the stimuli. Since we were interested in the framing of these narratives, it was important to preserve both the meaning of the content and its linguistic structure. Therefore, we could not make significant changes to these 40 stories, so as to preserve their narrative framing as much as possible. Unfortunately, simply not sharing data was not a good option, given the nature of the experiments. Since we were looking at linguistic phenomena, including examples in publications is vital to inform readers about the experiments. Additionally, other researchers requested our data to use in their own experiments, which further raised our concerns about the potential harm to the content authors.

If publishing our data will harm its authors, those authors must first be identifiable. Therefore, we undertook a study to determine how easily the content's original post could be found on the web, given only the text snippets used as stimulus data. Three people already associated with the project, but who did not know the original data sources, were tasked with using the web to find the stimuli sources. They were instructed to take up to 15 minutes for each story and to use whatever means available to the public that they wished. They were encouraged to take advantage of popular search engines and to experiment with many queries to find the original source. For each of the 40 stimulus stories, these annotators recorded their best guess as to the content's original source and the techniques they used to arrive at that guess.

Our findings suggest that we are right to be concerned about potential harm to the stimuli authors. Table 1 shows that, when presented with story data from the web, people can easily locate its source. While there is considerable variance between annotators, these results indicate that we can expect that at least a quarter of the authors could be identified from the content we use in our experiments, and that—depending on the skills of someone attempting to find the content—many more may be at risk of being identified. In addition to reporting exact matches, we also report the accuracy of the annotators when being more permissive in what guesses we are willing to accept. The dataset that we drew the 40 stimulus stories from has some content that is duplicated across the web. For instance, some stories have been re-blogged by other bloggers, and others were drawn from journalism organizations where the text has been syndicated verbatim to other sites. In other situations, the events that the narrative describe have been covered from different perspectives, with different text describing the same occurrences. We include these in table 1 because, with additional effort and scrutiny, it is likely people could find the original source.

Additionally, we looked at the strategies that the annotators reported were successful. Our annotators used a variety of commercial search engines in their efforts. Universally, they reported that using many keywords from the story—particularly searching whole sentences and taking advantage of phrase-based search features—were highly successful. This suggests that any attempt to obscure the content's source must involve substantial changes to the story vocabulary to reduce the effectiveness of these strategies.

Protecting User Identities

Given the results in the previous section, we became concerned about potential risk to the authors of the content we drew our stimuli from. While these authors posted publicly, using this sensitive content in our experiments and publicly disseminating it may bring attention that these authors did not anticipate and do not want. However, refusing to disseminate the stimuli is not an option because examples are necessary when publishing work and sharing data is important to make progress in this research area.

Basic strategies to obscure the content's source are not appropriate for this particular task. For instance, one strategy might be to re-tell the story. Under this strategy, a person is asked to read the story, and later recount its events from memory. This re-telling is recorded, and can be used with fewer worries of the original source being revealed. While this can preserve the general narrative content of the story, the structure and framing is likely to be lost, and will certainly be missing the authenticity of someone framing personal behavior and experiences.

	Annotator			
	1	2	3	Mean
Correct	31.6%	44.1%	55.6%	43.5%
Correct or ST	42.1%	58.8%	72.2%	57.4%
Correct or ST or SE	47.4%	58.8%	75.0%	62.0%

ST=Same Text, but different website.
SE=Same Event, but different perspective.

Table 1: Accuracy of annotators' guesses of the original source of the stimuli stories.

	Original	Back-translation
Correct	43.5%	5.3%
Correct or ST	57.4%	15.8%
Correct or ST or SE	62.0%	26.3%

ST=Same Text, but different website.
SE=Same Event, but different perspective.

Table 2: Accuracy of annotators' guesses of the original source of original and back-translated stimulus stories.

Instead, what is necessary is a strategy that still preserves general sentence structure but sufficiently alters vocabulary and word order to prevent against identification through search. One such strategy is back-translation, a process where text is translated from and then back into its source language. Properly conducted back-translation will preserve the purpose of each sentence, maintain the order of sentences, and maintain enough content structure to preserve the narrative framing. We confirmed this over the course of our narrative framing project. As part of this project, we conducted fMRI scans of native Farsi speakers to see their response to sacred framing. This necessitated stimuli in Farsi, which we generated by translating the original, English-language stimuli. We then performed the same narrative analysis performed on the English-language stimuli (presented by Sagae et al. (2013)) on these Farsi-language stimuli, which indicated that the narrative structure we are interested in was preserved across translation. With this finding, we are confident that back-translation preserves the narrative features we are interested in.

To test if back-translation would sufficiently alter word use to prevent identification using web search, we repeated the study described in the previous section. A new annotator was provided with the back-translated stimulus stories and asked to spend up to 15 minutes searching the web for the originally posted content. Our findings indicate that back-translation is a viable strategy to obfuscate the identity of authors. Table 2 compares the accuracy of identification for the 40 original stimulus stories to identification from back-translated stories. These results reveal that it is more difficult to identify these content authors after back-translation. This indicates that back-translation may be a good strategy when looking to protect content authors while still preserving some facets of narrative and linguistic structure.

Our annotator reported that strategies that were effective for the original stimulus stories are less effective for the back-translated stories. Particularly, searches for phrases are almost entirely ineffective, likely because of the changes to vocabulary as a result of the back-translation process. More effective than phrase-based searches was searching for keywords selected from the back-translated story.

Finally, to more objectively compare finding the original source using a back-translated or unaltered stimulus story, we looked at search rankings using our story search system. Comparing the search result rankings for the source stories using a full stimulus story as a query, it is more difficult to find the original sources using the back-translated stories. The average and median ranks of the original content using the back-translated stimuli are 494.4 and 14 respectively, compared to 7.1 and 1 for the original stimuli. This means a person attempting to identify the authors of these 40 stories will need to look much deeper in the results list to find the original source. This indicates that it would be more difficult to identify content authors based on back-translations than data derived directly from the original source.

Discussion

Important in web research is taking care to protect users whose data is used in the course of the research. To that end, we have presented an issue from our own research, argued that there is a real risk to the people whose data we have gathered across the web, and presented a way to mitigate those risks. We have shown that back-translation is a potentially valuable tool for web researchers concerned about protecting users' identities when sharing data, though additional experiments are necessary to determine if it is effective against more sophisticated methods of identification. We encourage web researchers to take care to protect web users whose information they use in their research, whether using the method we have presented—back-translation—or other techniques appropriate in their domains.

In the course of this work we have been struck by inadequacies of the current ethical guidelines for web researchers. If users post publicly, current guidance is that researchers can freely work with their data, even if the dissemination of this content presents a risk to the users. Despite the public nature of these posts, users may not expect that their content will circulate beyond a small circle of friends and family, and are highly unlikely to expect their data to receive attention as the result of its inclusion in a research effort. Researchers should be wary of publishing raw data gathered from the web, especially content describing socially questionable behavior engaged by the author, for fear of the negative attention that publication might bring to these authors.

More generally, a rethinking of our ethical frameworks for research are in order. The classic notion of interactional research with informed participants is often incompatible with how research is conducted using the web. Instead, researchers frequently gather and analyze data from web users who may not be informed of the research effort. Frequently this data contains deeply personal information that these web users may not anticipate reaching beyond a small audience. While the researchers never interact with these users, they can be studied with a similar level of detail to that of a participant in a traditional research study, all without consent or even acknowledgment. Even if this data is posted publicly, it should not absolve researchers of obligations to protect and minimize risk to these unwitting participants.

Acknowledgments

This research was sponsored by the Defense Advanced Research Projects Agency, Grant number D12AP00069. The content of this publication does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

The projects or efforts depicted were or are sponsored by the U. S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Acquisti, A., and Gross, R. 2009. Predicting social security numbers from public data. *Proceedings of the National Academy of Sciences* 106(27):10975–10980.

Eastham, L. A. 2011. Research using blogs for data: Public documents or private musings? *Research in nursing & health* 34(4):353–361.

Gimbel, S.; Kaplan, J.; Immordino-Yang, M.; Tipper, C.; Gordon, A.; Dehghani, M.; Sagae, K.; Damasio, H.; and Damasio, A. 2013. Neural response to narratives framed with sacred values (abstract). In *Annual meeting of the Society for Neuroscience*.

Gordon, A. S., and Swanson, R. 2008. Storyupgrade: Finding stories in internet weblogs. In *International Conference on Weblogs and Social Media*.

Gordon, A., and Swanson, R. 2009. Identifying personal stories in millions of weblog entries. In *Third International Conference on Weblogs and Social Media, Data Challenge Workshop*.

Gordon, A. S.; Wienberg, C.; and Sood, S. O. 2012. Different strokes of different folks: Searching for health narratives in weblogs. In *2012 International Conference on Social Computing*.

Kawase, R.; Nunes, B. P.; Herder, E.; Nejdl, W.; and Casanova, M. A. 2013. Who wants to get fired? In *Proceedings of the 4th Annual ACM Web Science Conference*.

Sagae, K.; Gordon, A. S.; Dehghani, M.; Metke, M.; Kim, J. S.; Gimbel, S. I.; Tipper, C.; Kaplan, J.; and Immordino-Yang, M. H. 2013. A Data-Driven Approach for Classification of Subjectivity in Personal Narratives. In *2013 Workshop on Computational Models of Narrative*.

Secretary's Advisory Committee on Human Research Protections (SACHRP). 2013. Considerations and recommendations concerning internet research and human subjects research regulations. US Department of Health & Human Services.

Wang, Y.; Norcie, G.; Komanduri, S.; Acquisti, A.; Leon, P. G.; and Cranor, L. F. 2011. "i regretted the minute i pressed share": a qualitative study of regrets on facebook. In *Proceedings of the Seventh Symposium on Usable Privacy and Security*.

Zheleva, E., and Getoor, L. 2009. To join or not to join: the illusion of privacy in social networks with mixed public and private user profiles. In *Proceedings of the 18th International Conference on World Wide Web*, WWW '09, 531–540.