Barriers to Security and Privacy Research in the Web Era *

Chris Grier[1], Kurt Thomas[2], and David M. Nicol[2]

[1] University of California, Berkeley grier@cs.berkeley.edu
[2] University of Illinois at Urbana-Champaign {kathoma2,dmnicol}@illinois.edu

Abstract. This paper argues that in order to enable security and privacy research on the web, modifications are required to existing legal and ethical guidelines that unduly restrict research. First, we propose that ethics review boards should update their definitions of public and private data in the context of web studies. Further, we argue that the terms of service provided by many of the most popular web applications hinder research and should be amended to facilitate access to researchers. We demonstrate how each of these issues impede web related research by examining the legal and ethical requirements of common web experiments.

1 Introduction

As web applications replace traditional desktop counterparts and personal data is migrated online, users are being exposed to novel security and privacy risks. This same migration is marked by increasingly restrictive policies set down by site operators that prevent research access to data necessary to analyze and expose risks. This is exemplified by terms of service obligations that place undue burdens on web researchers, forbidding automated analysis of web applications or scraping web data. Further complications arise as ethical guidelines enforced in the United States by Institutional Review Boards (IRBs) have become outdated, the web redefining our understanding of public and private data and human subjects research. In all of this, researchers are left to balance their ethical obligations against their ability to perform research.

Our position is twofold: first, existing guidelines for the protection of human subjects should be better defined for dealing with human created content on the web; second, requirements to adhere to terms of service must be reconsidered as the terms should not be a blanket statement limiting researchers. Web security research plays an important role in preventing attacks that target infrastructure and shut down web services, as well as other cybercrime, including fraud and identity theft [1]. Security researchers have been working to determine ethical guidelines [2] for network security research while overly restrictive policies already in place are slowing research progress [3]. The current system of guidelines and rules prevents researchers from improving the security and privacy of web applications that millions of people interact with every day.

2 Challenges

Research targeting web services and applications faces a number of legal and ethical hurdles including rights to privacy, ownership, and risks surrounding data

* This work supported in part by the National Science Foundation grant NSF-0433702. The views and conclusions contained here are those of the authors and do not necessarily reflect the views of the National Science Foundation.

retention and release. We explore each of these challenges and the limitations of existing guidelines, showing how the web era challenges basic assumptions that were foundational to outlining ethical procedures for researchers. In Section 3 we discuss specific examples that demonstrate how each of these issues impact researchers.

2.1 Privacy

With the widespread adoption of web services and applications, extensive personal data and usage information is becoming available to researchers. This includes user content exposed through application programming interfaces (APIs) provided by companies such as Google and Facebook, content posted on blogs and forums, and usage statistics released by projects such as Mozilla Test Pilot [4]. Analyzing web generated content for security and privacy flaws places researchers at the forefront of determining how to respect and maintain user privacy.

Personally Identifiable Information. Currently, the collection and study of personal information in the United States is restricted by the Common Rule [5]. The federal mandate requires any federally funded research involving human subjects to obtain prior approval from an organization's Institutional Review Board (IRB) in order to minimize risks posed to research subjects. A number of univeristies extend this coverage to any human subjects research being conducted, regardless of federal funding [6].

The Common Rule defines a human subject as any individual whom a researcher obtains "identifiable private information" or interacts with to gather data [5]. While the guidelines for IRBs were initially codified to combat unethical medical research, the broad definition of a human subject requires numerous fields to apply for IRB approval [7, 6] including security and privacy research.

Exemption from IRB review exists for studies strictly accessing *publicly* available data. However, the distinction between public and private data as understood by the Common Rule has become outdated for use in the web era. For instance, while an individual's name and address would be classified as identifiable private information, users regularly make this data publicly accessible on social networking sites, blogs, and personal web pages. For a researcher gathering this information, which classification takes precedent: public or private? Alternatively, if access to personal information is restricted to a subnetwork such as a geographic region or organization, is the data considered public or private?

The ambiguity surrounding the definition of public data poses a serious challenge to researchers as IRBs control whether a study qualifies for exempt status. Variations in interpretations across review boards can mean the difference between a lengthy administrative process and no oversight.

Data Retention & Release. A second challenge of privacy is confidentiality in regards to data retention and release. The Common Rule stipulates that all data gathered by IRB approved research must be retained for "at least 3 years" [5]. Further restrictions require that "the confidentiality of the personally identifiable information will be maintained throughout the research and thereafter" [5].

When considering the anonymization of a data set where personally identifiable information is removed before release, the web has redefined the meaning of what data is personally identifiable. Classical research surrounding anonymous data release [8–10] is being challenged by the vast overlap of data available from the web, turning what might be considered innocuous data into a potential identifier [11, 12].

The AOL search term release and subsequent privacy outcry [13] or the Netflix challenge and deanonymization risks [14] are prime examples of the challenges surrounding anonymous data release. Without a clear understanding of confidentiality requirements, researchers are faced with potential ethical and legal risks by releasing data they gather.

2.2 Terms of service

A website's terms of service constitute a set of rules that users must abide in order to access and utilize a web service. These restrictions fall under the domain of contract law, establishing a contract between the user and site operator [15]. Terms carry arbitrary obligations that can vary widely depending on the discretion of the provider. Deferring on whether a terms of service is enforceable or not, in this section we examine common restrictions encountered during web security and privacy research that place research methods at odds with content owners.

The terms of service appearing throughout the web are exemplified by those provided by Google, Facebook, and MySpace [16–18]. With respect to researchers, three common themes appear that regulate *access*, *storage*, and *transmission* of data. These sweeping restrictions include forbidding the creation of temporary accounts generated with random data, automated access to a service, crawling, scraping, or using a service in an unintended manner. Further restrictions forbid copying, duplicating, storing, or disseminating the content provided by the service in addition to protections offered by federal copyright law.

These terms come in regular conflict with web security and privacy research, often times resulting in violations that researchers are completely unaware of. For instance, a common method for examining security defects of a web application entails running it through a debugger to assist analysis. Though seemingly benign, this action can violate a number of rules set in the terms of service including, but not limited to: automation, access through an unauthorized means, and copying or duplication. Other studies that crawl the web in search of drive-by downloads or malicious content violate restrictions on automated access, copying, and storage. This issue is exacerbated when a large scale crawl of the internet is necessary where researchers must attempt to comply with hundreds of individual terms with unique restrictions, a problem similar to the tragedy of the anticommons [19] for patents and copyrights. Automation is necessary to achieve a representative data set and statistically validate studies that require thousands or millions of samples to conduct. Researchers need to clearly understand each terms of service contract they sign into and whether their research is legal.

3 Examples

In this section we explore a number of hypothetical scenarios where security and privacy research comes in conflict with ethics requirements of human subjects research, adherence to terms of service, and data retention and release. Each of these scenarios are taken from actual situations and are meant to be representative of many of the studies currently being conducted by researchers, yet highlight the risk of how ethical issues can be overlooked or overly burden researchers.

3.1 Privacy risks in social networks

As millions of users flock to online social networks such as MySpace and Facebook [20, 21], serious questions are being raised about the risks surrounding

revealing personal information to network owners, third parties [22, 23], and network participants [24]. In the absence of public data sets, a researcher must gather their own data set for study, placing them at odds with human subjects restrictions and terms of service obligations.

Consider a scenario where a researcher is interested in measuring adoption rates of access control restrictions within the MySpace community. For profiles that are accessible to the public, the user's content is stored for future analysis to determine how often personally identifiable information such as gender, a zip code, or birthday is present in the network. Despite the relative simplicity of crawling and storing data, the researcher is breaking a number of ethical requirements set down by the Common Rule and MySpace's terms of service.

With respect to the Common Rule, accessing profile information and storing personal data falls under the onus of human subjects research. Whether it qualifies for an exemption is largely based on an IRB's interpretation of public versus private data. The researcher is clearly collecting personally identifiable information that constitutes a human subject, but the data is publicly accessible to any party. Potential interpretations include:

- The data is publicly accessible and falls outside the definition of human subjects research; users do not have an expectation of privacy.
- The data contains private, identifiable information that poses a risk to subjects. Minimal risk of the study would need to be proved and a waiver of consent required else all subjects would need to be contacted for consent.

Alternatively, had the study been conducted on Facebook, where pages are restricted behind a login screen, further complications arise in differentiating between public and private data. Here, any user can sign up for an account, though this is not true when considering a crawl of a subnetwork of Facebook where a specific email such as `@illinois.edu` is necessary to acquire an account. When content is restricted to a specific university or behind a login, there are no clear IRB guidelines for determining if data should be considered public.

Beyond the ambiguity surrounding human subjects definitions, crawling MySpace and storing profile data is fraught with confidentiality and terms of service requirements. MySpace strictly forbids any use of automation, scraping, or downloading profile data [18]. These requirements amount to a double standard, where millions of users every day have access to the personally identifiable information present on MySpace, but researchers are restricted access. In the event gathering profiles does not violate the terms of service, simply removing the names of each subject before publishing results is not enough to protect a user's identity. Malicious parties can try to match profile attributes against those of public MySpace users, de-anonymizing the data set. Here, researchers must weigh their responsibility of upholding confidentiality against their desire to publish results that go beyond aggregate statistics.

3.2 Web application security

Web application security focuses on examining application code for vulnerabilities in order to protect against exploits or exposing sensitive user data. In recent years a number of evaluation techniques have appeared trying to protect against attacks such as cross-site scripting and SQL injection [25, 26].

An example of analyzing application security includes examining and modifying a popular web application to ensure it is free from any bugs that might

be exploitable. For a concrete example, consider a JavaScript based application, such as a document editor or web email, that relies heavily on the web browser for execution. Researchers have developed taint-tracking, static, and dynamic analysis techniques inside the browser that use programming languages techniques to analyze JavaScript based attacks. This involves developing custom tools that analyze the application provided, and for most web applications this is a combination of JavaScript and HTML and requires no reverse engineering. Then using the results of this analysis, tools automate rewriting the web application code to demonstrate safe transformation techniques. To evaluate the analysis and rewriting, the tools are demonstrated on a few popular web based applications.

This scenario can pose a number of legal and ethical concerns depending on the terms of service and agreements between users and the application provider. If researchers were examining the document editor provided by Google, the relevant terms of service [16] specifically prohibit:

- "You agree not to access (or attempt to access) any of the Services by any means other than through the interface that is provided by Google..."
- "You specifically agree not to access (or attempt to access) any of the Services through any automated means (including use of scripts or web crawlers)..."
- "You agree that you will not reproduce, duplicate, copy, sell, trade or resell the Services for any purpose."

Going down the list, analysis tools written by researchers access the web application in a manner that is different than intended (i.e. not through a web browser), and accessing an application automatically is also a violation of the terms of service. Further, any downloading, copying or modifications of the web application being studied (even for non-commercial, proof of concept purposes) are also a violation of the terms, limiting researchers ability to maintain temporary copies for examination. In addition to the terms of service limits on copying and duplication, other legal consequences could arise due to federal copyright law.

3.3 Analyzing malicious content online

One of many techniques for an attacker to gain control of a computer system is through crafted payloads that exploit a victim's web browser while visiting a web site [27–29]. Drive-by attacks such as these have grown in popularity and can leverage sophisticated server-side applications to detect the browser version and deliver the most effective exploit payload.

Finding and analyzing attacks delivered to the browser is a ongoing effort by the web security community and being conducted by companies and researchers alike [28, 30]. Consider a study that crawls the web and analyzes web content for malicious payloads. The first step is developing a crawler capable of scaling for the entire web and performing useful security analysis, a technically difficult feat. However, another difficulty lies in adhering to the terms of service presented by each website in a combined and coherent way. Simply locating and reading the terms of service for each site in a large scale study can be a impossible feat for researchers. Understanding and adhering to the terms is yet another issue.

Terms of service policies that strictly forbid automation or copying content pose a significant barrier to carrying out large scale web analysis. These studies are important for evaluating security techniques such as efforts to identify and protect users from drive-by download attacks, but are impossible to carry

out without automation. Archiving the data gathered throughout the crawling process may also potentially violate duplication restrictions set down by a site's terms of service, even if the data would greatly aid future research efforts in understanding the prevalence of web exploits and techniques to prevent them.

4 Discussion

As users adopt web applications and migrate personal information to online hosting, existing ethical guidelines and restrictions need to be re-evaluated. The scenarios we examined show how privacy and terms of service obligations can impact our ability to conduct security and privacy on the web. In this section, we present our recommendations for how to clearly define human subjects research in the web era and propose an alternative to existing terms of service restrictions. We hope to initiate a discussion on the unique challenges faced by web researchers and potential solutions that would open up new avenues for research.

4.1 Clearly defined IRB exemptions

As the web has redefined our understanding of personally identifiable information, we put forward a new classification of public and private data for determining exemptions from human subjects research.

Public data includes any information posted to the internet accessible to a crawler. Blogs, social network profiles, and user studies conducted by companies are all forms of public data. Data restricted behind a login is still public if (1) anyone can create an account and (2) there is no expectation of privacy, such as for newsgroups and forums. All public data should be exempt from IRB oversight and free of anonymization restrictions for public release.

Private data includes any information gathered by active interaction with users and extends to data gathered from sites where users have a reasonable expectation of privacy. Examples of private websites include restricted newsgroups, private wikis, and private forums where access is typically restricted by a login requiring moderator approval. The study and release of private data falls into the scope of human subjects research and must acquire IRB approval.

4.2 Terms of service for security and privacy researchers

Diverse terms of service and limiting language prohibits researchers from engaging in studies on many popular web applications. We propose the development of terms of service specifically designed for researchers to conduct proactive security analysis of existing services. These terms can be modeled after Flickr, whose current terms impose a lax set of restrictions that allow general use of their web applications for both users and researchers [31]. Alternatively, stricter terms can be adopted such as Google's Search and Translate which specifically grants university researchers open access to Google content, but require an application process and non-competitive obligation [32]. Using these two examples as guides, other companies should adopt similar terms of service to open up research access.

In developing new terms of service, companies should avoid clauses that restrict automation and crawling, as well as provide limited copy and modification rights for non-commercial use. Similarly, new terms should conform to a uniform standard of clauses and exceptions. Common terms of service restrictions would remove the uncertainty involved in adhering to thousands of arbitrary obligations when conducting large scale studies. While the existence of research-friendly terms of service is encouraging, widespread adoption by companies is necessary to enable future studies.

References

1. Burstein, A.J.: Amending the ECPA to enable a culture of cybersecurity research. Harvard Journal of Law and Technology **22**(1) (2008)
2. Dittrich, D., Bailey, M., Dietrich, S.: Towards community standards for ethical behavior in computer security research. Technical report, Stevens CS department (April 2009)
3. Burstein, A.J.: Conducting cybersecurity research legally and ethically. In: USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET '08). (April 2008)
4. Mozilla Labs: Test pilot `https://testpilot.mozillalabs.com/`.
5. United States Department of Health and Human Services: Protection of human subjects. Title 45 Code of Federal Regulations, Pt 46. 1998ed.
6. Garfinkel, S.: IRBs and security research: Myths, facts and mission creep. In: Proceedings of the 1st Conference on Usability, Psychology, and Security. (2008)
7. Gunsalus, C., Bruner, E., Burbules, N., Dash, L., Finkin, M., Goldberg, J., Greenough, W., Miller, G., Pratt, M.: Mission creep in the IRB world. Science **312** (2006)
8. Sweeney, L.: k-anonymity: A model for protecting privacy. International Journal of Uncertainty Fuzziness and Knowledge Based Systems **10**(5) (2002) 557–570
9. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: -Diversity: Privacy Beyond k-Anonymity. ACM Transactions on Knowledge Discovery from Data **1**(1) (2007)
10. Li, N., Li, T., Venkatasubramanian, S.: t-closeness: Privacy beyond k-anonymity and l-diversity. In: Proceedings of the IEEE International Conference on Data Engineering (ICDE). (April 2007) 106–115
11. Narayanan, A., Shmatikov, V.: De-anonymizing social networks. In: 30th IEEE Symposium on Security & Privacy. (2009)
12. Schoen, S.: What information is "personally identifiable"? (2009) `http://www.eff.org/deeplinks/2009/09/what-information-personally-identifiable`.
13. Barbaro, M., Zeller, T.: A face is exposed for AOL searcher no. 4417749. (2009) `http://www.nytimes.com/2006/08/09/technology/09aol.html`.
14. Narayanan, A., Shmatikov, V.: Robust de-anonymization of large sparse datasets. In: IEEE Symposium on Security and Privacy. (2008) 111–125
15. Specht v. Netscape Communications Corp. 150 F. Supp. 2d 585 (2001) `http://www.nysd.uscourts.gov/courtweb/pdf/D02NYSC/01-07482.pdf`.
16. Google: Terms of service (April 2007) `http://www.google.com/accounts/TOS`.
17. Facebook: Statement of rights and responsibilities (August 2009) `http://www.facebook.com/terms.php`.
18. MySpace: Terms of Service (2009) `http://www.myspace.com/index.cfm?fuseaction=misc.terms`.
19. Heller, M.: The tragedy of the anticommons: Property in the transition from Marx to markets. Harvard Law Review **111** (1998) 621–2462
20. Facebook: Statistics (2009) `http://www.facebook.com/press/info.php?statistics`.
21. MySpace: Statistics (2009) `http://www.myspace.com/statistics`.
22. Felt, A., Evans, D.: Privacy protection for social networking APIs. 2008 Web 2.0 Security and Privacy (W2SP08) (2008)
23. Singh, K., Bhola, S., Lee, W.: xBook: Redesigning Privacy Control in Social Networking Platforms. Proceedings of the 18th USENIX Security Symposium (2009)
24. Gross, R., Acquisti, A., Heinz III, H.: Information revelation and privacy in online social networks. In: Proceedings of the 2005 ACM workshop on Privacy in the electronic society, ACM (2005) 80
25. Jim, T., Swamy, N., Hicks, M.: Defeating script injection attacks with browser-enforced embedded policies. In: Proceedings of the 16th International Conference on World Wide Web. (2007) 601–610

26. Zeller, W., Felten, E.W.: Cross-site request forgeries: Exploitation and prevention. Technical report, Princeton University (October 2008) http://www.freedom-to-tinker.com/sites/default/files/csrf.pdf.
27. Symantec Inc.: Symantec global Internet security threat report: Trends for 2008. http://www.symantec.com/business/theme.jsp?themeid=threatreport (April 2009)
28. Provos, N., McNamee, D., Mavrommatis, P., Wang, K., Modadugu, N.: The ghost in the browser: Analysis of Web-based malware. In: Proceedings of the 2007 Workshop on Hot Topics in Understanding Botnets (HotBots). (April 2007)
29. Provos, N., Mavrommatis, P., Rajab, M.A., Monrose, F.: All your iFRAMEs point to us. In: Proceedings of the 17th Usenix Security Symposium. (July 2008) 1–15
30. Moshchuk, A., Bragin, T., Gribble, S.D., Levy, H.M.: A crawler-based study of spyware on the web. In: Proceedings of the 2006 Network and Distributed System Security Symposium (NDSS). (February 2006)
31. Flickr: Flickr apis terms of use (2009) http://www.flickr.com/services/api/tos/.
32. Google: University research program for google search - terms of use - google research (2009) http://research.google.com/university/search/terms.html.