Addressing Data Reuse Issues at the Protocol Level

Oshani Seneviratne
Decentralized Information Group
MIT Computer Science and Artificial Intelligence Lab
Cambridge MA, USA
oshani@csail.mit.edu

Lalana Kagal
Decentralized Information Group
MIT Computer Science and Artificial Intelligence Lab
Cambridge MA, USA
lkagal@csail.mit.edu

***Abstract*—We propose a novel way of managing how data on the Web is used with an infrastructure that enables accountability on the Web at the protocol level. We propose a protocol, HTTPA (Accountable Hyper Text Transfer Protocol), which requires that the data producer and the data consumer come to an agreement before an HTTP transaction takes place. This process makes both parties accountable for the agreement they had entered into, especially when reusing the data that was transferred. In HTTPA, the data consumer expresses her intentions of access and usage, and the data producer expresses her usage restrictions. The data transfer only happens when the intentions match the restrictions and the transfer along with the agreement is logged. This protocol cannot prevent the unauthorized reuse of data, but rather it can be used to develop accountability mechanisms that will identify violators allowing them to be held accountable for data they inappropriately consumed and served.**

***Keywords*-security; data privacy; authorization; authentication; accountability**

I. Introduction

Most discussions of Internet privacy, both policy and technology, tend to assume Alan Westin's perspective [1], which defines privacy as the ability for people to determine for themselves "when, how, and to what extent, information about them is communicated to others". This assumes that there are major privacy risks from unauthorized access to information. This focus on controlling information access has been found to be flawed [2]. The reality is that, even when the information is within reasonable bounds of security, it can leak outside these privacy boundaries violating the initial restrictions imposed on the data, as many social media outlets on the Web provide an easy medium for information dissemination at an unprecedented level. The technology press is filled with announcements by social networking sites about their new privacy controls, i.e. new ways for users to define access rules; followed by embarrassment when the choices prove to be inadequate or too complex for people to deal with [3], [4]. For example, Facebook's changes to its privacy settings in spring 2010 made news that highlighted how convoluted their privacy policy has become [5]. Tools such as a "Terms of Service Tracker" [6] have led to visualizations of how Facebook is sharing more private data than ever before [7]. Also, Facebook's Open Graph Protocol's "like" button has led to possible privacy violations ranging from exposure of browsing habits of people on medical sites to pornographic sites being shared with an unanticipated audience [8].

Even when access control systems are successful in restricting access to particular users, they are ineffective as privacy protection for systems like the World Wide Web, where it is easy to copy or aggregate information. These days, it is also possible to infer sensitive information such as social security numbers (SSN) [9], political affiliations [10], and even sexual orientation [11] from publicly available information. Another problem with using up-front access control systems is that it is the users' responsibility to define and maintain their privacy policies in every domain they participate in.

A pure notice and choice model is also not an adequate framework for privacy protection. The choice to whether opt-in or opt-out becomes meaningless and "user choice" is becoming a way for the industry to shift blame to users whenever a privacy breach happens. Many websites publish privacy policies which are often very verbose, and rarely do users have the time to read them or understand what they really mean. A typical user will click through the privacy policy statements without completely understanding the risks involved. In a pure access restriction system, those who obtain access to the data, legitimately or not, can use the data without restriction. An example for this, is the controversial whistle-blowing site *wikileaks*. This website exposes sensitive data with the aim of making governments and large businesses more transparent and accountable. Their claim is that no-one has been intentionally harmed so far because of the data published on the site. However, due to the sensitive nature of the data published on the site, it is possible for nations, if not individuals, to get harmed at some level and diplomatic relations to deteriorate. In a recent memo, several U.S. agencies have issued a warning [12] saying that the documents published on the site "does not alter the classified status or automatically result in declassification of the documents". Further, the memo states that "classified information, whether or not already posted on public websites or disclosed to the media remains classified and unauthorized federal employees should not look at

leaked classified data". This *usage restriction* is inherently faulty because there can be no enforcement (unless the employees are only accessing the website from their work computers where their web browsing is monitored), nor can employees be held accountable for accepting the restrictions imposed on the sensitive data.

Therefore, in addition to enforcing privacy policies through restricted access, which does not seem to work well in the current Web landscape, we suggest also using "information accountability" . Weitzner et al define information accountability in terms of usage–when information has been used, it should be possible to determine whether the usage was appropriate, identify the violators and hold them accountable [13]. In our accountability research, we focus on helping users conform to policies by making them aware of the usage restrictions associated with the data [14], [15] and helping them understand the implications of their actions and of violating the policy, thus encouraging transparency and accountability in how user data is collected and used. Lampson argues that to be practical, accountability needs an eco-system that makes it easy for senders to become accountable and the receivers to demand it [16]. It is our belief that HTTPA will provide this eco-system.

II. Motivating Scenarios

Users are increasingly finding their information such as personal profiles, friends, and interests spread across multiple social networking sites and accessed by all sorts of people, many of whom they did not originally intend to share their data with. As social media is becoming central to many things ranging from recruiting to personal relationships, the ability to grant and restrict access to personal data is becoming critical. The ubiquity of the Web, the ability to connect data from external sites to the social networking sites, and the amount of time people spend interacting with social media are both advancing our freedoms and enabling novel invasions of privacy. It is our belief that users should be aware of and ideally be in control of information about them on the Web.

In the scenarios described below, we take a policy-centric view on privacy on the Social Web (i.e. the Web landscape that facilitates online social interaction), where policies capture the permissions such as access control, obligations such as terms-of-use, licensing, and other data-handling settings that allow a user to control their interactions with other users. In particular, policies apply privacy settings to the profile and social media frameworks to consistently manage the user expectations of privacy and other obligations. This allows individuals and businesses on the Social Web to share information without any fear of violating user privacy or any regulations within the purview of the intention of use of their audience. We draw few examples from the Social Web to illustrate the importance of having the protocol described in this paper for transferring private data on the Web. These examples show how the intentions of data access will be matched up with the usage restrictions imposed on the social data of an individual.

In the following scenarios assume that Alice is a user of an imaginary social networking site called 'SocialBook'. Alice communicates with SocialBook using our protocol, and both parties have specified their intensions and usage restrictions using the RMP (Respect My Privacy) ontology [15]. The Provenance Tracker 'TrustMe' is a third party entity trusted by both Alice and SocialBook.

A. Upstream Usage Restriction Management

Suppose Alice wants to upload some pictures on SocialBook. The default settings on her smart Web client is set with the usage restriction that any HTTP payload carrying data with MIME type such as 'image', or subtypes such as 'image/[bmp,gif,jpeg,png,x-ico,x-tiff]' will only be posted/uploaded if the recipient acknowledges the full ownership of the content to her. However, it appears that SocialBook has extremely draconian terms of service that if uploaded to SocialBook, the data becomes the property of SocialBook. Alice's client examines these two policies, and informs Alice about the mismatch, which then prompts Alice to either stop posting her pictures or to notify SocialBook for the potential terms of use mismatch. In the latter case, TrustMe gets a notification of the handshake that happened between the parties. If SocialBook decides to modify the terms of use, it will send another request which Alice accepts and the data will be transferred.

B. Downstream Usage Restriction Management

Alice has a photo on SocialBook with a usage restriction specifying that the photo cannot be used for any *commercial* purposes. An employee from a large advertising company, Bob, accessed that photo. Bob's smart client confirmed with SocialBook and was logged on TrustMe that the intention of accessing the photo was *non-commercial*, and that he will honor the corresponding usage restriction that Alice has imposed on the photo. However, few weeks later, Alice found out that Bob had used her photo in an online advertisement for his company. Through her Web client Alice complains to TrustMe by giving the URI of her photo that Bob had allegedly used. Alice in her compliant also says that Bob's advertisement had used her photo, and that it is of *commercial-use*. TrustMe verifies that Bob had accessed the photo by looking up the accountability logs. Then it looks up the original usage restriction that Bob agreed to, verifies that it had indeed violated Alice's terms of use, and sends a takedown request to Bob with a proof detailing the violation.

III. HTTPA in a Nutshell

In HTTPA, before every data transfer, the provider and the consumer have to agree on the usage restrictions associated

with the data, and the intentions for data access. This is facilitated by a trusted third party "Provenance Tracker" in an "intentions and usage restrictions handshake". The sender/data producer will evaluate to what extent the usage restrictions match the data consumer's intentions. If they match, the data consumer is granted access to the data; else she is notified of the mismatched components.

The protocol's success hinges on the following crucial components given below[1]:

A. Authentication

Authentication is important in the protocol, not just for access control, but also to find the identity of the users who accessed resources should their owners claim that someone violated their usage restrictions on those resources. HTTPA will use the WebID protocol [17] to manage authentication.

B. Usage Restrictions Management

HTTPA uses the RMP [15] to describe the usage restrictions and the intentions associated with the data. Some of the terms included are: No Ownership Transfer, No Commercial/Employment/Financial/Medical/Insurance use of the data.

C. Handshake

HTTPA breaks away from the traditional client-sever model of HTTP transactions, to allow clients to act as servers, and vice versa. The sender (server/data provider) conveys usage restrictions, and the receiver (client/data consumer) notifies her intentions on the data. In the current implementation, we define two HTTP Headers: 'X-UsageRestrictions' and 'X-Intentions' for these purposes. If any one of the parties do not agree with the other party's usage restrictions/negotiations, further negotiations can be carried out using the 'X-Negotiate' header.

D. Provenance Trackers and Logging

Provenance Trackers are essentially special Web servers that are delegated to handle logging to enable provenance in HTTPA transactions. They are trusted by both parties involved in the data transfer, and the party initiating the transaction can designate the provenance trackers. The logs kept at the provenance trackers have several characteristics: they are immutable except by protocol components, encrypted, secure, readable only by trusted parties involved in the HTTPA transaction, and have all the records pertaining to a particular data transfer and usage such as what data was accessed, the specified intent of access, and the agreed upon usage restrictions.

[1]Due to the space constraints, we will not go in to detailed explanations of each of the components.

E. Accountability Tracking

If a user finds that she was wronged because someone else misused her data by violating the usage restrictions associated with the data, she can take recourse by producing a provenance trail with the help of the provenance tracker.

IV. Related Work

Various machine readable approaches to describing privacy policies have been proposed over many years. P3P (Platform for Privacy Preferences) protocol [18] was developed at the W3C with the intention of communicating the privacy policies of websites to the user-agents who connect with them. The recommendation allows website operators to express their data collection, use, sharing, and retention practices in a machine-readable format. A user-agent can retrieve a machine readable privacy policy from the Web server and respond appropriately (for e.g. display symbols or prompt the user for action). However, P3P has several limitations: a complicated language to express policies, inability to express preferences on third party data collection, and to specify multiple privacy policies for one Web page [19]. These limitations have prevented P3P from wide adoption. Unlike in P3P, both parties have a say in the data transfer in our protocol. Also, our work attempts to bring down the complexity barrier by making the usage restrictions and the intentions expression simpler with the help of smart clients.

FTC endorsed a 'Do not Track proposal' [20] recently to facilitate consumer choice about online tracking, and there are already several implementations that support this proposal. One of the most compelling technical implementations describes sending the user's intention of not to track online browsing behavior in an HTTP header [21]. Although this approach works for this specific use case, it seems very limited for general purpose usage restrictions matching with intentions. Also, the communication described in their proposal through HTTP Headers is mono-directional, whereas our protocol allows bi-directional communication enabling both parities to engage in a dialogue.

The Simple Policy Negotiation for Location Disclosure proposal [22] describes a system that lets a user have a dialogue with a website that uses her location data before disclosure. Their proposal has many similarities to our accountable data transfer protocol, such as implementing a simple standard for transmitting policy information just-in-time. However, their domain is limited to geo-location data, and the their standard does not handle provenance tracking.

V. Conclusion

HTTPA addresses the limitations of current privacy work and provides the infrastructure to build more privacy-aware systems. The requestor, on data access, will convey what her intention for the data access is. The data provider will determine the compliance/non-compliance of the intention

sent by the requestor with the usage restrictions associated with the resources that are being accessed. Their negotiation is being logged by a trusted third party called 'Provenance Trackers' to ensure accountability. If usage restrictions are compliant with the intentions, the data access request will be successful. If it is non-compliant, an explanation as to why the data cannot be transferred will be conveyed to the requestor. The recipient of the data will be held accountable for the usage restrictions she accepted upon the data transfer. In other words, recipients cannot argue after the fact that they did not know the expectations of the data server: for retention or for use of information. Similarly, users cannot claim after the fact that the data server was deceptive or that they had not been informed. This enables market and regulatory forces to punish users who misuse data. We believe that government organizations, academic institutions, and businesses will be the early adopters of this accountable Web protocol with usage restriction management within their intranets. On the longer run, in a similar vein in which the growth of e-commerce Web sites led to the massive adoption of HTTPS, we envision that HTTPA will be accepted by the larger Web community, as privacy problems slowly cripple the growth of the Web.

ACKNOWLEDGMENTS

The authors would like to thank Tim Berners-Lee, Hal Abelson, Joe Pato, Mike Speciner and other members of the Decentralized Information Group for their valuable input on this topic. This material is based upon work supported by the National Science Foundation under Award No. CNS-0831442 and by the Air Force Office of Scientific Research under Award No. FA9550-09-1-0152.

REFERENCES

[1] A. Westin, "Privacy and freedom (Fifth ed.). New York, U.S.A.: Atheneum," 1968.

[2] L. Kagal and H. Abelson, "Access control is an inadequate framework for privacy protection," in *W3C Privacy Workshop*, 2010. [Online]. Available: http://www.w3.org/2010/api-privacy-ws/papers/privacy-ws-23.pdf

[3] The Local, "Headmaster fired after Facebook pic scandal," 2009. [Online]. Available: http://www.thelocal.se/20148/20090618

[4] GigaOm, "Is facebook beacon a privacy nightmare?" [Online]. Available: http://gigaom.com/2007/11/06/facebook-beacon-privacy-issues

[5] G. Gates, "Facebook privacy: A bewildering tangle of options." [Online]. Available: http://www.nytimes.com/interactive/2010/05/12/business/facebook-privacy.html

[6] T. T. of Service Tracker (A project of the Electronic Frontier Foundation), "Facebook privacy policy." [Online]. Available: http://www.tosback.org/policy.php?pid=39

[7] M. McKeon, "Facebook privacy vizualizer." [Online]. Available: http://mattmckeon.com/facebook-privacy

[8] M. Tuffield, "Nhs.uk allowing google, facebook, and others to track you." [Online]. Available: http://mmt.me.uk/blog/2010/11/21/nhs-and-tracking

[9] PC World, "Researchers expose security flaw in social security numbers," 2009.

[10] J. Lindamood, R. Heatherly, M. Kantarcioglu, and B. Thuraisingham, "Inferring private information using social network data," in *World Wide Web Conference poster paper*, 2009.

[11] C. Jernigan and B. Mistree, "Gaydar: Facebook friendships reveal sexual orientation," 2009.

[12] "Wikileaks Access Warning." [Online]. Available: http://edition.cnn.com/2010/US/12/03/wikileaks.access.warning

[13] D. Weitzner, H. Abelson, T. Berners-Lee, C. Hanson, J. Hendler, L. Kagal, D. McGuinness, G. Sussman, and K. K. Waterman, "Transparent Accountable Inferencing for Privacy Risk Management," in *AAAI Spring Symposium on The Semantic Web meets eGovernment*, March 2006.

[14] O. Seneviratne, L. Kagal, and T. Berners-Lee, "Policy aware content reuse on the web," in *ISWC2009 - International Semantic Web Conference*, October 2009. [Online]. Available: http://dig.csail.mit.edu/2009/Papers/ISWC/policy-aware-reuse/paper.pdf

[15] T. Kang and L. Kagal, "Enabling privacy-awareness in social networks," in *Intelligent Information Privacy Management Symposium at the AAAI Spring Symposium 2010*, March 2010. [Online]. Available: http://dig.csail.mit.edu/2010/Papers/Privacy2010/tkang-rmp/paper.pdf

[16] B. Lampson, "Usable security: how to get it," *Communications of the ACM*, Jan 2009. [Online]. Available: http://portal.acm.org/citation.cfm?id=1592761.1592773

[17] "Webid protocol," *WebID 1.0 - Web Identification and Discovery*. [Online]. Available: http://getwebid.org/spec/drafts/ED-webid-20100809/index.html

[18] L. F. Cranor, "Web privacy with platform for privacy preferences," *Oreilly Books*, Jan 2002.

[19] "Pretty poor privacy: An assessment of p3p and internet privacy," *Electronic Privacy Information Center*, June 2000. [Online]. Available: http://epic.org/reports/prettypoorprivacy.html

[20] Federal Trade Commission Staff Report, "Protecting consumer privacy in an era of rapid change - a proposed framework for businesses and policymakers," 2010. [Online]. Available: http://www.ftc.gov/os/2010/12/101201privacyreport.pdf

[21] J. Mayer and A. Narayanan, "Do Not Track - Universal Web Tracking Opt Out." [Online]. Available: http://donottrack.us

[22] E. Wilde, "Simple policy negotiation for location disclosure," *w3.org*. [Online]. Available: http://www.w3.org/2010/policy-ws/papers/03-Doty-Wilde-Berkeley.pdf