Identifying and Addressing Privacy Leakage from Query Logs: An Accountability Approach

Massachusetts Institute of Technology Computer Science and Artificial Intelligence Lab 32 Vassar Street, Cambridge, MA 02139

Instead of modifying query logs to prevent privacy and identity leakage, we propose to allow users and query log providers to explicitly specify what data in the logs is available for research and what researchers can do with this data. We suggest that instead of restricting access to the logs, researchers be allowed to use the logs for certain purposes and in a specific manner as defined by privacy policies. Researchers can be monitored to ensure appropriate usage of the dataset by auditing their transaction logs and can subsequently be held accountable for any privacy breaches they cause by their failure to conform to the policy.

Categories and Subject Descriptors: D.2.7 [Software Engineering]: Distribution and Maintenance—*documentation*; H.4.0 [Information Systems Applications]: General; I.7.2 [Text Processing]: Document Preparation *languages*; *photocomposition*

General Terms: Documentation, Languages

Additional Key Words and Phrases: Document preparation, publications, typesetting

1. INTRODUCTION

The use of query logs for studying search engines and for improving information retrieval on the Web is invaluable. This use, however, is currently restricted as it might sacrifice user privacy and expose a significant amount of private and identifying information. Query logs - whether from Microsoft adCenter, Live Search, or Live Search SDK - alone do not lead to leakage of private information or to identifying individual searchers. However, cross referencing these logs with data (public or otherwise known) or using data mining algorithms to find patterns in these logs could lead to a breach of privacy. It might not only be possible to associate these logs with individuals but also to uncover sensitive information about these individuals.

We believe efforts to address information policy issues such as online privacy have been overly dominated by access restriction and privacy-preserving algorithms such as anonymization, generalization, and perturbation. An alternative is to emphasize the design of systems that provide greater information accountability as judged against rules governing appropriate use, rather than information security and access restriction. In a world where information is ever more easily copied and improperly passed on even by authorized users, and where automated correlations and inferences across multiple databases can uncover information even when it has not been explicitly revealed, accountability must become a primary means by which society addresses issues of appropriate use.

Several privacy-preserving data mining techniques such as anonymization, generalization, and perturbation are used to prevent privacy leakage in publicly available datasets such as query logs. These techniques, however, have vulnerabilities that can be used for breaching privacy. Sweeney deanonymized medical records by cross referencing them with a voter database [Sweeney 1997]. Similarly Narayanan demonstrated that a little

2 · MIT CSAIL

data about an individual Netflix subscriber can be used to easily identify this subscriber's records from the Netflix anonymized dataset [Arvind Narayanan and Vitaly Shmatikov 2006]. Backstrom et al. showed that users cannot expect privacy from anonymized social networks where each individual's identity is replaced by a random user ID but the connections between individuals are revealed [Backstrom et al. 2007]. Another privacy-preserving data mining technique is data perturbation, where the original data is perturbed such that the original data record cannot be recovered but the patterns in the data can be mined. This technique also has vulnerabilities that can lead to privacy problems as highlighted by Liu [Liu et al. 2007].

Instead of modifying query logs to prevent privacy and identity leakage, we propose to allow users and query log providers to explicitly specify what data in the logs is available for research and what researchers can do with this data. We suggest that instead of restricting access to the logs, researchers be allowed to use the logs for certain purposes and in a specific manner as defined by privacy policies. Researchers can be monitored to ensure appropriate usage of the dataset by auditing their transaction logs and can subsequently be held accountable for any privacy breaches they cause by their failure to conform to the policy. Privacy policies will be specified in terms related to privacy preserving techniques such as anonymization, degree of anonymity, clustering, cluster size, abstraction, gaussian mixture model, etc. Different attributes of the log (time-stamp, query string, url, etc) can have different privacy preserving requirements. For example, the query string can require anonymity of degree 5 whereas the time-stamp can require gaussian mixture model abstraction. By agreeing to this policy, researchers are obliged to utilize the query log according to the privacy requirements and are also obliged to capture each transaction they perform on this query log data. Our approach does not prevent privacy or identity leakage from query logs but allows violators of applicable privacy policies associated with these logs to be identified and held accountable.

Transparency and accountability will make any privacy breaches visible to all concerned. This visibility alone will not guarantee policy compliance. The vast majority of legal and social rules that form the fabric of our societies are not enforced perfectly or automatically, yet somehow most of us follow most of the rules most of the time. We do so because social systems built up over thousands of years encourage us to do so, and often make compliance easier than violation. For those comparatively rare cases where rules are broken, we are all aware that we may be held accountable through a process that looks back through the records of our actions and assesses these actions against the rules. Augmenting systems with policy awareness will help us bring these information systems more in line with the transparency and accountability we rely on in other arenas where human interaction is governed by social rules.

2. FUNDAMENTAL POLICY ISSUES IN QUERY LOG RESEARCH AND PRI-VACY

Pro's: QL research conducted in open, peer-reviewed context can help -advance the stateof-the-art in information retrieval -support impartial, peer-reviewed research assessing the effectiveness and potential bias of commercial search engines

Con: increase in privacy risk

3. THE INFORMATION ACCOUNTABILITY APPROACH

Background about IA approach

Our goal is to develop a technical proof-of-concept that can be the basis for monitoring and enforcing privacy rules in various query log research contexts. In particular, our prototype will demonstrate how information accountability techniques can be address privacy challenges when it is not possible or practical to protect privacy by the various access limitation approaches discussed in section one. We believe that information accountability with respect to query logs usage will emerge from four basic thrusts: privacy-preserving ontology, audit logging, a policy language framework, and accountability reasoning tools. We propose to use Semantic Web techniques for defining privacy preserving techniques as well as the policy language to enable progressively larger overlapping communities to develop shared vocabularies in a step-by-step, bottom-up fashion. Perfect global interoperability of these ontologies and policies is unlikely, but that is not a fatal flaw. Just as human societies learn to cope with overlapping and sometimes contradictory rules, especially when jurisdictional boundaries are crossed, so too will policy-aware systems be able to develop at least partial interoperability.

Main components of our accountability framework include:

Privacy-preserving ontology: Terms in this ontology will be used to define privacy requirements of query log usage as well as to describe transactions made by researchers over these query logs. We propose to extend Gil's privacy preserving ontology for data analysis in workflows [Gil et al. 2007] and modify it for use with query logs.

Transaction logs: Researchers will have to assume responsibility for recording information usage events that may be relevant to the current or future assessment of accountability to the set of privacy policies associated with the logs. A number of fundamental questions must be answered about logs, however: what information should be kept and what discarded? How will the logs themselves be secured?

Policy Language Framework: We have developed AIR, a policy language based on Semantic Web technologies, for defining privacy policies pertaining to the misuse of data [Hanson and Kagal 2007]. We will adapt this language and its associated reasoner in order to support the privacy-preserving ontology and the privacy requirements of query logs.

Policy Reasoning Tools: We will develop an accountable system that will assist users in seeking answers to questions such as: Is this piece of query log data allowed to be used for the specified purpose ? Has this data been modified and used according privacy requirements ? Is a string of inferences permissible for use in a given context, depending on the provenance of the data and the applicable rules.

4. REVIEW OF CURRENT WORK

-privacy preserving datamining -problem of assessing privacy leakage in the open world of the Web

This proposal aims to integrate Gil's initial privacy preserving ontology with the TAMI framework in order to investigate the requirements of query log privacy policies and their impact on the AIR policy language and accountability mechanisms developed within TAMI. Gil et al use semantic web technologies to add privacy awareness into workflow systems such as e-Science that perform analysis on distributed data sets [Gil et al. 2007]. Gil's current research proposes a set of terms for preserving privacy but lacks a reasoning or enforcement mechanism. We believe that TAMI fills this gap and that accountability is the

4 · MIT CSAIL

appropriate technique for managing privacy violations of query log data.

OSERA is an open source project whose goal is to provide an open source architecture for the Federal government (http://www.osera.gov/). Within this project, researchers have modeled the Privacy Act of 1974 and developed a privacy ontology. The US Privacy Act ontology permits the inferring of allowed disclosures based on disclosure target, authorization, and intent. This ontology though interesting is not directly applicable to query logs.

Policy-Aware Web is a project that investigates rule-based policy management for the Web (http://policyawareweb.org). The PAW framework exploits the inherently decentralized and open nature of the Web by allowing policies, meta-policies, and policy languages to be combined, extended, and otherwise handled in the same scalable, modular manner as are any Web resources [Kagal et al. 2006]. It integrates a Semantic Web rules language (N3) with a theorem prover designed for the Web (Cwm) and makes it is possible to use the Hypertext Transport Protocol (HTTP) to provide a scalable mechanism for the exchange of rules and, eventually proofs, for access control on the Web. This project was a precursor to TAMI and allowed us to explore several of our initial research ideas with respect to policies, policy languages, rules, and representation of policies.

Contextual integrity is a logical framework that is aimed at understanding privacy expectations and their implications [Barth et al. 2006]. It is able to express and reason over policies of transmission of personal information. These policies focus on whom the information is about, how it is transmitted, and past and future actions by both the subject and the users of the information. The model appears to be able to capture many notions of privacy policies in legislation such as those in HIPAA, COPPA, and GLBA. This theoretical work has promising results but does not provide a practical enforcement mechanism or tools for policy compliance.

5. THREAT MODELS AND EXAMPLES OF HOW AIR COULD ENCOURAGE COMPLIANCE AND PROVIDE ACCOUNTABILITY

A. Internal abuse (ie. students in an IR lab run an identity theft ring) B. Privacy leakage from published aggregate research results

6. SUMMARY

7. EXPECTED OUTCOMES

REFERENCES

- ARVIND NARAYANAN AND VITALY SHMATIKOV. 2006. How to Break Anonymity of the Netflix Prize Dataset. *New Scientist* 2577 (November).
- BACKSTROM, L., DWORK, C., AND KLEINBERG, J. 2007. Wherefore Art Thou R3578X? Anonymized Social Networks, Hiddne Patterns, and Structural Steganography. In *World Wide Web Conference*.
- BARTH, A., DATTA, A., MITCHELL, J., AND NISSENBAUM, H. 2006. Privacy and Contextual Integrity: Framework and Applications. In *IEEE Symposium on Security and Privacy*.
- GIL, Y., CHEUNG, W., RATNAKAR, V., AND KIN CHAN, K. 2007. Privacy Enforcement in Data Analysis Workflows. In *Privacy Enforcement and Accountability with Semantics Workshop at International Semantic Web Conference.*
- HANSON, C. AND KAGAL, L. 2007. AIR Policy Language. http://dig.csail.mit.edu/TAMI/ 2007/amord/air-specs.html.
- KAGAL, L., BERNERS-LEE, T., CONNOLLY, D., AND WEITZNER, D. 2006. Using Semantic Web Technologies for Policy Management on the Web. In 21st National Conference on Artificial Intelligence (AAAI 2006).

5

- LIU, K., GIANNELLA, C., AND KARGUPTA, H. 2007. A Survey Of Attack Techniques On Privacy-Preserving Data Perturbation Methods. Springer.
- SWEENEY, L. 1997. Weaving technology and policy together to maintain confidentiality. Journal of Law, Medicine and Ethics, 25(2-3):98-110.