A Policy Driven Semantic Approach to Data Usage Management

Anupam Joshi, Tim Finin, Karuna Joshi, Madan Oberoi
CSEE Department, UMBC
Baltimore, MD 21250
{joshi, finin, kjoshi1}@umbc.edu

***Abstract*— As the amount of information available on the web has increased, several privacy and security issues around the use of such information have arisen. Government (and private) entities are able to gather and analyze data from several disparate sources with ease. This ability to do large scale analytics of publicly accessible data leads to significant privacy concerns, especially when done by governments. The converse is also true, with concerns about data being shared by individuals and organizations to the web and the cloud. Our work develops a semantically rich, policy driven approach to address the privacy, security and usage concerns around such data.**

Introduction and Motivation

In today's highly networked information infrastructure, a significant amount of information is accessible publicly over the web. Such information is gathered by a variety of government and private entities. This information, gathered from a variety of sites, can be linked together and analyzed to make inferences about entities of interest. While the expectations of privacy vary with culture and country, it appears that often citizens are relatively more comfortable with commercial companies mining their personal information rather than law enforcement agencies collecting and mining this data across information sources. One concern in particular is that Law Enforcement or Counter Intelligence agencies often use such public information to "fish" for potential suspects [1, 2, 3]. Similar concerns about data aggregation have also been expressed recently about companies (such as Google, Facebook, etc) that provide a platform with a variety of applications that are commonly used.

A related issue is the problems being faced by cloud/web based service platforms. These have the promise to significantly lower the cost and increase the effectiveness of many data storage, access, and analysis tasks. However, reluctance of individuals and organizations to share data because of privacy, confidentiality, and usage concerns is preventing their adoption. Within the past year for instance, the federal government in the US has mandated that data centers be consolidated, and that a set fraction of the federal IT tasks be done using (public) clouds [32]. A key barrier to this however is the reluctance of the CIOs to let data go outside the organization, since they cannot ensure that the cloud/web based provider will be able to meet the organization, as well as legal/statutory constraints on sharing and usage that they have to enforce.

Our research has sought to address this issue by using machine understandable and semantically rich descriptions of the a) data, b) policies governing access, usage and privacy, and c) the query context

Related Work

The TAMI (Transparent Accountable Data-mining Initiative) project attempts to address issues of transparency, accountability in context of personal privacy by changing the perspective from controlling or preventing access to encouraging appropriate use of accessed data and inferring when data is misused by investigating the audit logs [10]. Our proposed work is closely related as it relies on logs to figure out whether obligations are met. However, unlike TAMI, our model does enforce privacy policies but does so on the end use data produced as a result of the query instead of the initial data dump required.

Kagal, Hanson and Weitzner [11] have discussed providing explanations associated with the derivation of a policy decision in the form of a list of reasons, called dependencies by them, using semantic web technologies. This kind of explanations will help the user as well as database owner agencies to understand how the results were obtained, thereby increasing trust in the policy decision and enforcement process. Our model will provide similar justifications about query decisions.

A lot of work has been done to develop machine interpretable policy frameworks [12], [13]. Rein (Rei and N3) [14] is a distributed framework for describing and reasoning over policies in the Semantic Web. It supports N3 rules [15], [16] for representing interconnections between policies and resources and uses the CWM forward-chaining reasoning engine [17], to provide distributed reasoning capability over policy networks. AIR [18] is a policy language that provides automated justification support by tracking dependencies during the reasoning process. It uses Truth Maintenance System [19] to track dependencies. Policies and data are represented in Turtle [20], whereas the reasoning engine is a production rule system [21] with additional features for improved reasoning efficiency such as goal direction. Rei and AIR consider rules defined over attributes of classes in the domain including users, resources, and the context. Though our initial prototype uses OWL to describe privacy policies, we plan to use AIR in the future to take advantage of its built-in justification feature.

Letouzey et al [22] have discussed existing security models by defining the security policy through logically distributing RDF data into SPARQL views and then defining dynamic security rules, depending on the context, regulating SPARQL access to views. Kagal and Pato [23] have explored the use of semantic privacy policies, justifications for data requests, and automated auditing to tackle the privacy concerns in sharing of sensitive data. Their architecture evaluates incoming queries against semantic policies and also provides a justification for permitting or denying access, which helps requesters formulate

privacy-aware queries. Currently our conceptual model does not restrict the query language to be used, but we plan to use SPARQL for better integration with Semantic Web data sources.

FRAMEWORK

The basis of our approach is the use of policies that describe the data, along with the constraints on that data (who can access it, under what circumstances, for what use etc.) that the individual or the organization providing that data wishes to associate with it. Another element of our approach is articulating the context in which the query is made. The context of the query minimally includes who is asking for the information, and for what purpose. More generally, it includes an identification of the person or entity which initiated the query, their role in some (predefined) hierarchy which the data store understands, the group(s) to which they belong, and the intended use of the information. In this sense, we capture the concepts associated with usage [6] and group based controls [7]. In order to address privacy concerns, organizations that collect personal data during their routine business prepare and publish privacy policies to assure their clients. These privacy policies determine the way, modalities, quantum, time period after which, conditions/situation under which, and with whom such personal information can be shared. We note that these policies are generally not machine interpretable or formal policies. However, by making them machine interpretable, we can reason over these policies, and the query context, to decide if the data can be shared. An important feature of the approach is the system of automatic periodic audit to check whether the privacy policies were correctly enforced or not, and identify cases of exception. This is particularly useful in cases where information is shared with 'after-access' obligations, for instance those that maintain that the data would only be used for the stated purpose. The audit component helps to assure the database owners that their privacy policies are being complied with by the user who queried for the data.

A similar approach is used to handle the case of using services on the web or the cloud to store data and perform computations (such as analytics) on it. The claim is that by removing complexity and management issues from the user end, a lower total cost of ownership and greater efficiencies can be realized by cloud based services. Many organizations however face a major barrier to adopting such systems -- they have complex internal policies, as well as legal and statutory constraints on how they handle their data that must be enforced. Such policies are today enforced on internal resources (like data centers) controlled by the organization. For instance, a policy might say that the data must be stored under a certain jurisdiction. When acquiring remote cloud services, it today requires significant human intervention and negotiation -- people have to check whether a provider's service attributes ensure compliance with their organization's constraints. This can get very complex if the provider is composing services using components provided by third parties distributed across the web.

Another concern that organizations have for cloud based services is with security and privacy of the data on the cloud. Since most of the cloud based services allow multiple users at the same time (multi-tenancy), organizations are reluctant to use cloud services for their business critical applications. A semantically rich policy-based framework that manages the cloud data access and security permissions can help elevate these concerns.

Our approach includes a methodology to address the lifecycle issue for virtualized services delivered from the web or the cloud [30], including elements related to data management. This lifecycle provides ontologies [31] to describe data, services and their attributes. In particular, we use semantically rich descriptions of the requirements, constraints, and capabilities that are needed at each phase of the lifecycle [29]. Policies can be described using the same ontology terms so that compliance checks can be automated. This methodology is complementary to previous work on ontologies, e.g., OWL-S, for service descriptions in that it is focused on automating processes needed to procure services on the cloud.

We realize the overall model using OWL (Web Ontology Language) [8] as our semantic description language for the data and query context using ontologies that we have developed [28]. We use Jena [9] as our reasoning infrastructure, and Jena Rules are used to describe policies.

We have developed and implemented a cloud storage service prototype to demonstrate and evaluate our methodology. We used Semantic Web technologies such as OWL, RDF, and SPARQL to develop this tool. The prototype allows cloud consumers to discover and acquire disk storage on the cloud by specifying the service constraints, security policies and compliance policies via a simple user interface. This prototype was developed as part of our collaboration with National Institute of Standards and Technology (NIST).

IMPLEMENTATION

We use a smart cloud broker based approach to address the problem of encouraging the use of web/cloud services. When acquiring web or cloud based services, the consumer organization identifies the technical and functional specifications that a service needs to fulfill. In addition, they specify the organizational policies and legal constraints relating to data usage and management, and security/privacy policies for the service. Service compliance policies such as required certifications, standards to be adhered to, etc. are also identified. Depending on the service cost and availability, a consumer may be amenable to compromise on the service quality. Once the consumers have identified and classified their service needs, they issue a Request for Service (RFS) to a cloud broker service. This RFS uses the ontologies we have developed [30,31] to specify elements of the service acquisition process, as well as security and usage constraints.

The broker engine queries various service providers to match the service domain, data type, compliance needs, functional, and technical specifications; and returns the result with the service providers in priority order. If a consumer finds the exact service meeting their constraints, they can begin consuming the service. Otherwise, the consumer and the service provider will have to negotiate on the service

constraints and policies to be met. Service acceptance is usually guided by the Service Level Agreements (SLA) that the service provider and consumer agree upon. A side effect of the negotiation process is that machine understandable SLAs specified in our ontology are automatically generated [32], and can be used for monitoring compliance.

At times, the service provider will need to combine a set of services or compose a service from various components delivered by distinct service providers in order to meet the consumer's requirements. Hence, service negotiation also includes the discussions that the main service provider has with the other component providers. When the services are provided by multiple providers (composite service), the primary provider interfacing with the consumer is responsible for composition of the service.

For the information gathering aspect of the data usage management problem, a compliance checker, similar in concept to the broker above, is used. In our prototypes we have focused on a centralized entity. In ongoing efforts, we are investigating methods to distribute this component. Our ontology describes the notion of hierarchical position level, group, and use. We have adopted description logics (DL), specifically OWL, and associated inferring mechanisms to develop the model and policies. The requester information consists of his position in the hierarchy, his group membership and use for which information is being sought. In our system this information is represented in N3 [15] using the NAT ontology we have developed. The *Nat* ontology defines various properties such as '*belongs_to_hierarchyLevel',* '*has_designation'* and *'belongs_to_group'* that can be used to represent the requester details. FOAF [25] is used to allow individuals to describe personal information about themselves and their relationships. This information is used to determine whether the requester has the permission to access the query result based on data owner's (or provider's) privacy policies. The reasoning engine performs reasoning over this information and privacy policies. Our system uses the Jena Semantic Web framework [26] [27] for reasoning over the context data and the policies. These reasoners are used to infer additional facts from the existing knowledge base coupled with ontology and rules. The instance of such reasoner with a ruleset can be bound to a data model and used to answer queries about the resulting inference model. In our system, the reasoning engine uses the *Nat* ontology and the FOAF ontology to represent the requester information, and privacy policies represented in the Jena rule language to generate an inference model. This inference model is used to decide whether the information can be released to requester.

CONCLUSION

The model described above addresses the usage management and control concerns in a multi-user and multi-database owner environment. It addresses both the data gathering issues (where information is gathered from multiple sites and combined to make inferences) and the cloud/web service issue (where data has to be shared with a service provider on the web).

REFERENCES

[1] U.S. General Accounting Office, "Data Mining: Federal Efforts Cover a Wide Range of Uses", (GAO-04-548), May 2004, at 3, 27-64, http://www.gao.gov/new.items/d04548.pdf.

[2] Department of Homeland Security, "Report to Congress on the Impact of Data Mining Technologies on Privacy and Civil Liberties", 7 (2007), http://www.dhs.gov/xlibrary/assets/privacy/privacy_rpt_datamining_2007.pdf.

[3] Department of Homeland Security, "Report to Congress on the Impact of Data Mining Technologies on Privacy and Civil Liberties" 8 (2006), http://www.dhs.gov/xlibrary/assets/privacy/privacy_data_%20mining_%20report.pdf;

[4] Conference Report Cantigny Conference Series, "Counterterrorism Technology and Privacy", McCormick Tribune Foundation, 2005, http://www.mccormickfoundation.org/publications/counterterrorism.pdf

[5] Gio Wiederhold, "Mediators in the Architecture of Future Information Systems", IEEE Computer, March 1992, pages 38-49.

[6] Jaehong Park, Ravi Sandhu "The $UCON_{ABC}$ usage control model", ACM Transactions on Information and System Security (TISSEC) Volume 7 Issue 1, February 2004ACM New York, NY, USA

[7] R. Krishnan, R. Sandhu, J. Niu, and W. H. Winsborough. "A conceptual framework for group-centric secure information sharing". In ASIACCS '09: Proceedings of the 4th International Symposium on Information,Computer, and Communications Security, pages 384–387, New York, NY, USA, 2009. ACM.

[8] W3C, "OWL Web Ontology Language", February 2004, http://www.w3.org/TR/owl-features/

[9] "VOID – Vocabulary of Interlinked Datasets", http://semanticweb.org/wiki/VoiD

[10] Daniel J. Weitzner, Harold Abelson, Tim Berners-Lee, Chris Hanson, James Hendler, Lalana Kagal, Deborah L. McGuinness, Gerald Jay Sussman, K. Krasnow Waterman, "Transparent Accountable Data Mining: New Strategies for Privacy Protection", MIT CSAIL Technical Report-2006-007, http://www.w3.org/2006/01/tami-privacy-strategies-aaai.pdf

[11] Lalana Kagal, Chris Hanson, Daniel Weitzner, "Using Dependency Tracking to Provide Explanations for Policy Management", IEEE Policy: Workshop on Policies for Distributed Systems and Networks, June 2008

[12] Tim Moses. eXtensible Access Control Markup Language TC v2.0 (XACML), February 2005.

[13] Sushil Jajodia, Pierangela Samarati, V. S. Subrahmanian, and Elisa Bertino. "A unified framework for enforcing multiple access control policies" In Proceedings of ACM SIGMOD International Conference on Management of Data, pages 474–485. ACM Press, 1997.

[14] Lalna Kagal and Tim Berners-lee. "Rein : Where policies meet rules in the semantic web", Technical report, Laboratory, Massachusetts Institute of Technology, 2005.

[15] TimBerners-Lee and Dan Connolly, "Notation3 (N3): A readable RDF syntax", Technical report, 2008.

[16] Tim Berners-Lee, Dan Connolly, Eric Prud'hommeaux, and Yosi Scharf, "Experience with n3 rules", In Rule Languages for Interoperability, 2005.

[17] Tim Berners-Lee, "Cwm - a general purpose data processor for the semantic web".

[18] Lalana Kagal, Chris Hanson, and Daniel Weitzner, "Using dependency tracking to provide explanations for policy management", In Proc. IEEE Workshop on Policies for Distributed Systems and Networks, pages 54–61, Washington, DC, 2008. IEEE Computer Society.

[19] Jon Doyle, "Truth maintenance systems for problem solving", Technical report, Cambridge, MA, USA, 1978.

[20] D. Beckett, "Turtle - Terse RDF Triple Language", Technical report, 2007.

[21] D. A. Waterman and F. Hayes-Roth, editors, "Pattern-Directed Inference Systems", 1978.

[22] Gabillon, A. Letouzey, L. Univ. de la Polynesie Francaise, Faaa, French Polynesia, "A View Based Access Control Model for SPARQL",

Network and System Security (NSS), 2010 4th International Conference, September 2010, Melbourne.

[23] Lalana Kagal_ and Joe Pato, "Preserving Privacy Based on Semantic Policy Tools", IEEE Security & Privacy Magazine Special Issue on: "Privacy-Preserving Sharing of Sensitive Information" August 2010, http://dig.csail.mit.edu/2010/Papers/IEEE-SP/db-privacy.pdf

[24] Mathew Cherian, "A Semantic Data Federation Engine", MIT Masters Thesis Jan 2011

[25] "The Friend Of A Friend (FOAF) Project ",http://www.foaf-project.org/

[26] "Jena – Semantic Web Framework for Java", http://jena.sourceforge.net/

[27] Carroll et al, "Jena: implementing the semantic web recommendations", ACM, pages 74-83, 2004

[28] Madan Oberoi, Pramod Jagtap, Anupam Joshi, Tim Finin and Lalana Kagal, "Information Integration and Analysis: A Semantic Approach to Privacy", Proc. Third IEEE International Conference on Information Privacy, Security, Risk and Trust (PASSAT), Boston, MA, Oct 2011

[29] Karuna Joshi, "Automation of Service Lifecycle on the Cloud by Using Semantic Technologies", Proceedings of tenth International Semantic Web Conference, Part II, pp 285-292, Bonn, Oct 2011

[30] K. Joshi , T. Finin , Y. Yesha, "Integrated Lifecycle of IT Services in a Cloud Environment", in Proceedings of The Third International Conference on the Virtual Computing Initiative (ICVCI 2009), Research Triangle Park, NC, October 2009

[31] K. Joshi, OWL Ontology for Lifecycle of IT Services on the Cloud, 2010, http://ebiquity.umbc.edu/ontologies/itso/1.0/itso.owl

[32] US Federal Cloud Computing Initiative, http://www.info.apps.gov/node/2, retrieved on Feb 28 2012