Privacy Enforcement in Data Analysis Workflows

Yolanda Gil[1], William K. Cheung[2], Varun Ratnakar[1], Kai-kin Chan[2]

[1] Information Sciences Institute, University of Southern California
4676 Admiralty Way, Marina del Rey CA 90292, United States
{gil, varunr}@isi.edu
[2] Department of Computer Science, Hong Kong Baptist University,
Kowloon Tong, Hong Kong
{william, kkchan}@comp.hkbu.edu.hk

Abstract. Collaborative e-Science projects commonly require data analysis to be performed on distributed data sets which may contain sensitive information. In addition to the credential-based privacy protection, ensuring proper handling of computerized data for disclosure and analysis is particularly essential in e-Science. In this paper, we propose a semantic approach for enforcing it through workflow systems. We define privacy preservation and analysis-relevant terms as ontologies and incorporate them into a proposed policy framework to represent and enforce the policies. We believe that workflow systems with the proposed privacy-awareness incorporated could ease the scientists in setting up privacy polices that suit for different types of collaborative research projects and can help them in safeguarding the privacy of sensitive data throughout the data analysis lifecycle.

Keywords: Workflow generation, scientific workflows, privacy, trust.

1 Introduction

Trust and security were always central to the vision of the Semantic Web [1]. In a recent paper, Weitzner et al. [2] argue for a policy-aware infrastructure for the Web that ensures privacy and other social needs that would encourage people to share information freely. They also propose developing systems that are transparent and accountable [3] regarding their use of sensitive data from individuals and therefore can demonstrate their compliance with existing privacy laws.

The Web has always raised concerns for privacy data. There is concern about the wide availability of yellow pages and other directory information, and the fact that protected or sensitive information may become available over the Web perhaps unintentionally [4]. Of particular concern are record linkage techniques to cross-reference independent data sources and data mining algorithms that detect patterns, associate them with individuals, and reveal private or sensitive information about individuals that may violate basic privacy rights.

1.1 Motivation: Privacy in e-Science

Although privacy has broader interest and applicability, our research arises and focuses in the context of e-Science applications. Many areas in biomedical sciences envision benefit from clinical records (e.g., cabig.nci.nih.gov), phenotype information, and health history. In social and behavioral sciences, widely available on-line information can be integrated and analyzed to reveal significant patterns that emerge in specific communities, influential groups and individuals within a social network, and trends or events of interest. Much of this research is hindered because of the concern of individuals with their privacy and therefore their reluctance to allow the use of their personal data. Yet, many people would choose to give up their privacy for some greater good such as advancing medical research, especially when they are provided with mechanisms to protect the privacy of their data [6].

A variety of mechanisms are being investigated to ensure data privacy including secure data storage, data access control, auditing mechanisms, and securing lines of communication. Also, laws and policies for protecting and enforcing health information privacy will need to be formulated in order to determine how those technologies need to be used to implement the law. These mechanisms are important and necessary to control the access and release of data. However, they will not necessarily support the anticipated sophistication of people's wishes over the *fine-grained control* over the *uses* of their sensitive data, say, for clinical data analysis conducted by some third parties. Furthermore, the control is further complicated by the recent trend that the uses of sensitive data are no longer confined to the institution that collected or owns the data but *highly distributed* (e.g., cabig.nci.nih.gov).

1.2 Privacy Protection in Workflow Systems

In recent years, a variety of workflow systems have been developed to manage complex scientific analysis processes [5]. We see workflows as an artifact that captures, among other things, how data is being transmitted, pre-processed and analyzed, and for what purpose. Of particular concern for us is to enforce privacy protection in workflows by enabling workflow systems with privacy-awareness. Workflow systems can represent detailed models of the individual computations performed in the data, and be extended to express their privacy-related properties. In recent years, a variety of algorithms and approaches for privacy-preserving data analysis are being developed [8], where some transform data into privacy-preserved versions before putting together for subsequent analysis while some compute intermediate analysis results via a distributed and secure protocol. With these kinds of approaches, data sets can be processed and analyzed with well-defined guarantees as well as risks about the preservation of privacy of individuals. Thus, the already complex data analysis processes are now further complicated by the need and at the same time possibility to have data privacy protection integrated. The use of the semantic approach has been demonstrated to be effective in assisting users in creating and validating complex data analysis workflows, e.g., for large-scale earthquake data analysis [10], [11].

1.3 Our Contributions

We take a semantic approach to incorporate privacy awareness into workflow architectures and our implementation in the Wings/Pegasus workflow system [10]. The focus of our work to date has been on privacy policies that need to be addressed when workflows are designed and created. In particular, we show how the workflow systems could be extended to be able to *detect privacy policy violation* and to *provide corrective actions* for revising the workflows before the data analysis process can be safely executed.

The paper begins with ontological representations for privacy-relevant terms in data analysis workflows and illustrate how those ontologies could be used to describe workflow systems that incorporate traditional data analysis algorithms and privacy-preserving algorithms for analyzing sensitive data. To support automatic privacy policy enforcement in data analysis workflows, we propose a particular policy representation which has components describing applicable context, data usage requirement, privacy protection requirement, and corrective actions if the policy is violated. We present initial results on extending a workflow system to include representations of privacy policies that can be enforced by the system. We finalize with a discussion of related work and possible avenues for future research in this area.

2 Ontological Representations of Privacy-Relevant Terms in Data Analysis Workflows

Figure 1 depicts an ontology that contains core workflow concepts typically used to represent workflows and the extensions needed for describing privacy relevant concepts in data analysis workflows (shown in bold face). The classes shown in normal face are adopted from [10] for constructing workflows, including a *file ontology* for representing datasets, a *component ontology to* represent computations that correspond to steps in the workflow, and a *workflow ontology* to represent data-independent workflow templates. Unlike other scientific workflows that are composed of web services, Wings/Pegasus workflows being consider in this work are composed of codes that can be submitted for execution in a resource selected by the workflow system [10].

Privacy Preservation Ontology. This ontology includes a *PrivacyPreservation* class of privacy preservation methods that convert the input into privacy preserved forms. Privacy preservation methods can process on each attribute individually or the data set as a whole. *PrivacyPreservationPerAttribute* contains component types such as *Anonymization* (e.g., masking, substitution) and *PrivacyPreservationPerDataset* contains *Generalization* (e.g., k-anonymity [8]).

Data Analysis Ontology. It provides a separate taxonomy of data analysis methods. We consider here statistical data analysis algorithms that are widely used in many domains. In our ontology, *DataAnalysis* is the root class with subclasses like *Clustering* (e.g., Gaussian mixture model), *Classification*, etc.

Extensions of the file ontology. We extend the ontology to describe data protection up to *per-attribute* level. Some additional properties and classes added include:

- *hasAttribute* whose range captures data attributes described as *Attribute*.
- *hasAuthorizedUse* which refers to the intended use or purpose of the File.
- *Attribute* which models file attributes and has a property *protectedBy* (with sub-properties, e.g. *anonymizedBy*) to indicate the adopted privacy preservation method.
- special types of *File,* e.g., *DataSet* for raw data files and *Clusters* for clustering results which can go with data items (*ClustersWithDataItems)* or just per-cluster statistics (*ClustersWithStatistics*). The latter is needed when data privacy is an issue.

Extensions of workflow template ontology. Some properties added include:

- *hasPurpose* which refers to data analysis purpose.
- *hasOutputQuality* which refers to overall output quality descriptors, e.g. accuracy.

Extensions of component ontology. Some properties and subclasses added include:

- *hasParameterSet* which refers to the set of parameters needed by the component.
- *PPComponentType* which contains privacy preservation methods as its sub-classes, e.g., *Generalizer* (which in turn has sub-classes, e.g. *k-anonymity*)*,* and has a property *hasLevelOfProtection* for describing the level that its output is protected.
- *DAComponentType* which contains data analysis methods as its sub-classes, e.g., *Clustering* (which in turn has sub-classes, e.g. *GMM*), and has *supportPPType* and *supportDataType* to indicate its supporting types of privacy preservation and data.

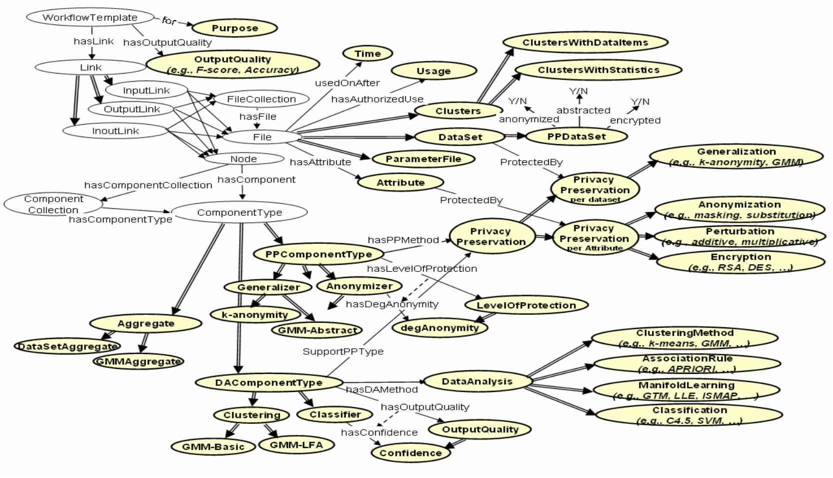

Fig. 1. An ontology for describing privacy aware data analysis workflows.

To illustrate how a domain-specific data analysis workflow can be described, we adopt a hypothetic *clinical data analysis* task. Like many other domains, clinical data can contain patents' personal identification and demographic information as well as sensitive ones including medical measurements, medical treatment, drug dosage, diagnosis, etc. We assume data collected from patient records archived at different clinic to (1) have the personal identification fields anonymized, (2) be generalized into groups based on their demographic information by k-anonymity and (3) be abstracted up to an agreed level of details based on the numerical medical attributes (e.g., by GMM [12]). Clustering is then carried out to identify patterns in different patient groups. Fig. 2 shows a related workflow together with a corresponding domain-specific workflow template created using Wings.

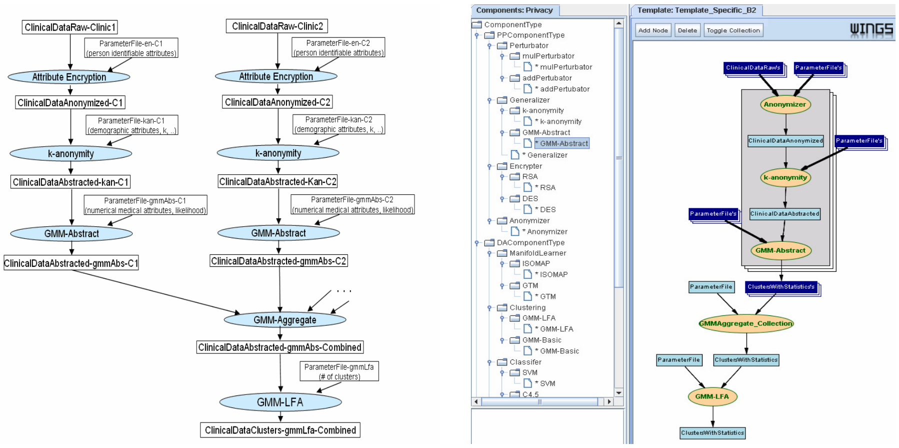

Fig. 2. A clinical data clustering workflow template represented in Wings.

3. A Privacy Policy Representation and Its Enforcement

We represent data privacy policies semantically based on the derived ontologies so as to support automatic policy enforcement in data analysis workflows via reasoning. Note that the privacy awareness being considered here is not conventional credential-based authentication and authorization. Instead we require a policy language that is flexible enough to describe conditions reflecting different relationships among data sets, components, and their privacy-relevant properties. Such a flexibility requirement naturally leads us to the use of rule-based representation. Note that other than expecting the users to specify the rules, rule-based policies carefully created by experts of the respective field can always be adopted.

In our current design, a policy representation contains four parts, namely (1) ***context***, (2) ***usage requirement***, (3) ***protection requirement***, (4) ***corrective action***. Informally, *context* specifies what workflows the policy applies. As we are dealing with data privacy, the context refers to some types of links, data or components where the policy is applicable. *Usage requirements* and *protection requirements* are for detecting policy violation within the context. Finally, *corrective actions* are suggestions for remedy of the policy violation, typically referring to the statements mentioned in the protection requirements. We further characterize requirements to be of *positive* and **negative** types. Positive requirements specify compliance conditions and policy violations occur when the conditions ARE NOT satisfied. Negative requirements specify non-compliance conditions and policy violations occur when the conditions ARE satisfied. As seen in the following, both types of requirements are essential for policy representation.

Context refers to the condition where the underlying policy is relevant. In other words, the policy applies only if this condition is satisfied.

- Example C1: *"Input files of a workflow containing medical images."*
  InLink(?l) ^ hasFile(?l, ?d) ^ hasAttribute(?d, ?a) ^ MedicalImage(?a)

Usage requirement refers to the non-amendable condition under which the use of data is required (+ve) or not allowed (-ve).

- Example UR1 (+ve): *"It is required that the purpose of the workflow should be equal to the authorized usage of the inputs that match the context."*
  for(?w, ?pw) ^ hasAuthorizedUse(?d, ?pw) ^ equal(?pw, ?pd)

Protection requirement refers to the *condition* when the use of data is required (+ve) or not allowed (-ve) with respect to data protection and analysis quality.

- Example PR3 (-ve): *"It is not allowed that the nodes that match the context have inputs with attributes in common."*
  hasAttribute(?d1, ?a1) ^ hasAttribute(?d2, ?a2) ^ equal(?a1, ?a2))

Corrective action refers to the remedies recommended to fix policy violation. For "usage requirement" violation, only a printed message stating the violating policy is expected as no remedy is possible. For "protection requirement" violation, a corresponding recommended action for fixing the violation will also be provided.

Policy Compliance Checking Via Reasoning We create 2 rules for each policy: a context component rule to locate where the policy applies and a requirement component rule to determine if non-compliance conditions occur within the context.

For the policy with a *negative* requirement, its context component rule and requirement component rule can simply be combined by conjunction and applied to a workflow description. Thus, the overall rule becomes [context rule] ^ [requirement rule] -> invalid (?l). Matched results will correspond to the policy violation situations.

For the policy with a *positive* requirement, the overall rule for detection problematic parts can be represented as [context rule] ^ not [requirement] -> invalid (?l). However, the rule becomes not a horn clause and thus cannot be easily represented using SWRL. Thus, instead of applying directly the overall rule, we apply the context rule first to the workflow and the matched results form a set with items of concern. Then, we applied the requirement rule to the set. The newly matched items are removed from the set in context and the remaining ones are the violation situations. This treatment works when the policies are free of conflicts among them. If there are some parts in the workflow with more than one policies applicable, policy conflicts will occur. We are currently investigating algorithms for policy conflict detection and resolution [15].

A Compliance Checking Walkthrough Given the workflow templates discussed in Section 4.1, two particular policy rules expressed in SWRL are considered:

General Policy G1:*"For all the inputs, it is required that the purpose of the workflow should be equal to the authorized usage of the inputs."*

context:	WorkflowTemplate(?w) ^ for(?w, ?l) ^ hasFile(?l, ?d)
usage:	+ve: for(?w, ?pw) ^ hasAuthorizedUse(?d, ?pdl) ^ equal(?pw, ?pd)
protection:	NULL
correction:	prompt [workflow and data purpose mismatch]

Domain Specific Policy S1: *"For data that contain dosage information, it is not allowed that they are not first anonymized before being used for analysis."*

context:	hasLink(?w, ?l) ^ hasFile(?l,?d) ^ hasAttribute(?d, ?a) ^ Dosage(?a)
	hasDestinationNode(?l, ?n) ^ hasComponent(?n, ?c) ^ DAComponent(?c)
usage:	NULL
protection:	-ve: anonymized(?d, ?aVal) ^ equal(?aVal, false)
correction:	prompt [add an anonymization step right after (?d) found at (?l)]

Suppose a researcher creates a simple workflow template that takes directly all the raw clinical datasets and feeds them into a basic GMM clustering component to perform a clinical study. The workflow system would find that policy G1 applies and is respected. However, policy S1 is fired as the aggregate dataset fed to the GMM-basic was found not to be anonymized. Fig. 3 shows the detection of the violation of the policy S1. The workflow in Fig. 2 complies with all these policies. In [13] we describe an interactive scenario where users would be assisted during workflow construction to create workflows that comply with a set of privacy policies.

4 Related Work

To the best of our knowledge, there has not been prior work on extending workflow systems with data privacy awareness. Some policy frameworks like KAoS [14] and Rei [15] have recently been proposed for security and privacy on the Semantic Web. To contrast with KAoS and Rei, our data privacy policies in data analysis workflows need to refer to properties of data, components, etc. In addition, the policies of concern are not credential-based ones as those in KAoS and Rei. Also, the policies we use not only are aimed to detect violations but also to suggest corrective actions in terms of how to fix the causes of violation.

5 Conclusions

In this paper, we motivated the need for a new type of privacy policies that constrain processing on data. We described our initial work on a semantic approach to represent privacy policies relevant to data analysis. We argued the validity of the approach by showing how privacy-preserving data analysis processes can be defined using ontologies, and how the ontologies can be combined with a policy framework to represent the policies. We discussed how those policies can be applied via examples. Future work includes conflict detection algorithms for the proposed policy framework and incorporation of the policy enforcement module in the Wings system.

Acknowledgement

This research was supported in part by the Air Force Office of Scientific Research (AFOSR) through grant FA9550-06-1-0031. The second author was a Visiting Scholar at ISI, USC while this paper was written, and he thanks Hong Kong Baptist University for a sabbatical leave.

References

1. Berners-Lee, T., Hall, W., Hendler, J., O'Hara, K., Shadbolt, N., Weitzner, D. "A Framework for Web Science." Foundations and Trends in Web Science, Vol 1, No 1 (2006)
2. Weitzner, D.J., Hendler, J., Berners-Lee, T., and Connolly, D.: Creating a Policy-Aware Web: Discretionary, Rule-Based Access for the World-Wide Web. In Web and Information Security, E. Ferrari and B. Thuraisingham (Eds), IRM Press (2005)
3. Weitzner, D.J., Abelson, H., Berners-Lee, T, Hanson, C., Hendler, J., Kagal, L., McGuinness, D.L., Sussman, G.J., Waterman, K.K.: Transparent Accountable Data Mining: New Strategies for Privacy Protection. Technical Report, MIT-CSAIL-TR-2006-007, MIT (2006)
4. Sweeney, L.: Finding Lists of People on the Web. ACM Computers and Society, 34(1) (2004)
5. Taylor, I.J., Deelman, E., Gannon, D., and Shields M.S. (eds.). Workflows for e-Science. Springer Verlag (2006)
6. Mandl, K.D., Szolovits, P., and Kohane, I.S.: Public Standards and Patients' Control: How To Keep Electronic Medical Records Accessible But Private". British Medical Journal, Vol. 322, No. 7281 (2001) 283-287
7. Deelman E. and Gil, Y.: Final Report of the NSF Workshop on the Challenges of Scientific Workflows. (http://vtcpc.isi.edu/wiki/images/3/3a/NSFWorkflowFinal.pdf) (2006)
8. Sweeney, L.: k-Anonymity: A Model For Protecting Privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002; 557-570.
9. Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M.: Tools for Privacy Preserving Distributed Data Mining. ACM SIGKDD Explorations, 4(2) (2003) 19-26
10. Gil, Y., Ratnakar, V., Deelman, E., Mehta, G. and Kim, J.: Wings for Pegasus: Creating Large-Scale Scientific Representations of Computational Workflows. Proceedings of the 19th Annual Conference on Innovative Applications of Artificial Intelligence (2007)
11. Kim, J., Gil, Y., and Ratnakar, V.: Semantic Metadata Generation for Large Scientific Workflows. In Proceedings of the Fifth International Semantic Web Conference (2006)
12. Cheung, W.K., Zhang, X., Wong, H., Liu, J., Luo, Z. And Tong, F: Service-oriented Distributed Data Mining. IEEE Internet Computing, 10(4) (2006) 44-54.
13. Cheung, W.K., and Gil., Y.: Towards Privacy Aware Data Analysis Workflows for e-Science. Proceedings of 2007 Workshop on Semantic e-Science (SeS2007), held in conjunction with the Twenty-Second Conference of the Association for the Advancement of Artificial Intelligence (2007).
14. Bradshaw J., Uszok, A., Jeffers, R., Suri, N., Hayes, P., Burstein, M., Acquisti, A., Benyo, B., Breedy M., Carvalho, M., Diller, D., Johnson, M., Kulkarni, S., Lott, J., Sierhuis, M. and Van Hoof, R.: Representation and Reasoning For DAML-based Policy and Domain Services in KAoS and Nomads. In Proceedings of the Second International Joint Conference on Autonomous Agents and Multiagent Systems. ACM Press, New York (2003) 835-842.
15. Kagal, L. Finin, T., and Joshi, A.: A Policy Language for Pervasive Systems. In Proceedings of the Fourth IEEE International Workshop on Policies for Distributed Systems and Networks, (2003) 63-76