Towards Privacy Aware Data Analysis Workflows for e-Science

William K. Cheung

Department of Computer Science
Hong Kong Baptist University
Kowloon Tong, Hong Kong
william@comp.hkbu.edu.hk

Yolanda Gil

Information Sciences Institute
University of Southern California
4676 Admiralty Way, Marina del Rey, CA 90292
gil@isi.edu

Abstract

e-Science is getting more distributed and collaborative and data privacy quickly becomes a major concern, especially when the data contain sensitive information. Existing data access policies for privacy management are too restrictive for supporting the large variety of data analysis needs in e-Science. In this paper, we argue the need of a new type of policies that govern data privacy based on the type of processing done on the data. A semantic workflow approach is proposed to address the challenge. Data analysis processes are described as workflows. Ontologies for data analysis and privacy preservation describe the functionalities and the privacy attributes of the processes, as well as process-constraining privacy policies. We give some examples of related policies with their potential fields for application explained. Also, we present via a case study on distributed data clustering to illustrate how the approach could be integrated with a workflow system to make it privacy aware.

Introduction

Data privacy is important in e-science, especially when distributed and collaborative data analysis processes are involved. It is not difficult to find scenarios where distributed data analysis and data privacy protection are both needed at the same time. For example, one can analyze individuals' clinical data like brain images by gaining access to related remote sources for disease diagnosis (Beltrame et al. 2006), where the patients' identity has to be kept strictly confidential. Other than the subjects' identity threat, the scientists themselves may have privacy concerns on their scientific findings as data sets, preliminary results and data analysis processes can now be easily and widely shared in e-Science collaborations (Deelman and Gil 2006). These privacy concerns are important even though the advantages of sharing data to facilitate collaborative scientific research are well understood (NIH 2004). Thus, the need for having privacy management support in e-Science is immediate.

Existing data access policies offer a very basic privacy protection mechanism, where a user can access a data source if his/her certificates and credentials satisfy the access policies defined for that data source. An alternative approach that is less restrictive is to apply privacy preserving techniques to the data before releasing it, where sensitive information like personal identity or medical background are properly hidden through anonymization and partitioning (Samarati 2001). In addition, one can design new data analysis algorithms that are privacy preserving, a fast-growing area for the past few years (Chris et al. 2003). These algorithms can preserve data privacy through techniques like secure multiparty computation (Lin et al 2005) and data generalization (Zhang & Cheung 2005), and yet can perform reasonable data analysis.

These existing mechanisms alone are too restrictive for many applications, especially in e-Science. Consider the case of a cancer patient signing a release form for their medical records. He or she may be not only willing but eager to allow access to for medical research purposes as long as it is anonymized. However, they suspect that if the record is released it could be used by insurance companies to design more profitable insurance rates that would raise his or her medical expenses. Given the choice to release the data or not without any say on the use of the data, the patient may decline to allow the use of its record. Clearly, a more flexible privacy mechanism would allow the patient to specify a policy on how the data will be processed, not just on the blanket release of the data per se. Consider another case, where a cancer research laboratory has collected treatment protocol data for thousands of patients over several decades. The lab is happy to share its data with medical researchers in other fields to analyze the relationship of cancer with liver transplantation or with heart failure, but would not want other competing research groups to use the data in competing cancer research quests. Existing approaches would not allow the lab to specify

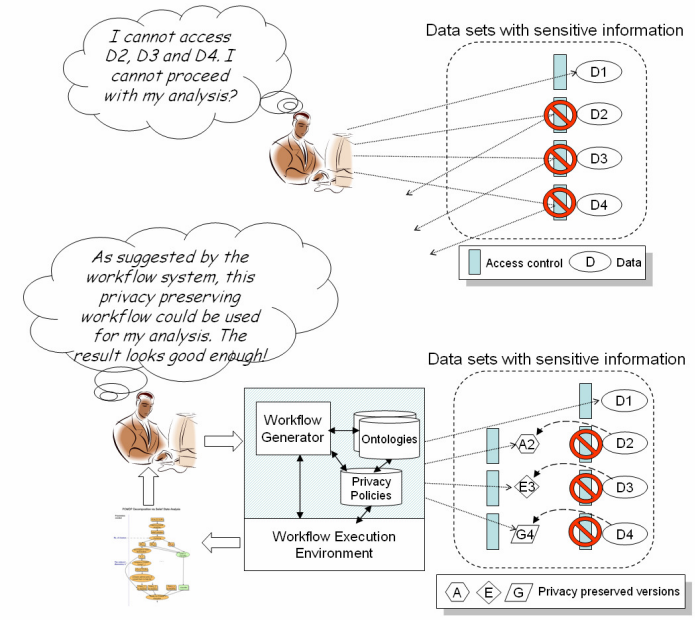

Figure 1. Data analysis using (a) traditional access control, (b) privacy-aware workflow systems.

this kind of policy concerning the use of the data, only to specify who would have access to it.

The goal of our work is to investigate *a new kind of privacy protection policies that constrain the type of processing on the data*, rather than the access to the data. That is, instead of defining policies to specify who can access a data set and how much of it can be accessed, our goal is to define policies that specify *what can be done with the data*. This would allow a more flexible approach to privacy that covers data processing in addition to the existing data access techniques.

The key idea in our approach to privacy protection is the use of *workflows* to describe the type of processing done to a dataset, and to express policies that can be used to control the creation and execution of workflows. Workflows have recently emerged as a useful paradigm to represent and manage complex computations in many scientific applications (Deelman and Gil 2006). *We propose to extend workflow systems to be privacy-aware*, so that they can be given privacy policies defined in terms of types of analysis and data handling performed by the workflow system. Figure 1 illustrates a traditional access control approach (a) and contrasts it with a privacy-aware workflow system (b). A data access control policy would either enable or disable a user's access solely based on their credentials and certificates. In this example, the user is only able to access D1 but not D2, D3, or D4. In contrast, a privacy-aware workflow system would enable the expression of additional kinds of privacy policies that would enable access based on the type of analysis done as expressed in the workflow. The workflow system could assist the user in modifying the analysis in order to satisfy the privacy policies stated. In this framework we can selectively specify privacy policies for the same data set that would allow it to be accessible for certain types of workflows (analyses) and not for others.

This paper describes our work to date in defining a new class of privacy policies that apply at the workflow (data processing) level. To define these workflow-level policies, one needs to address the following major issues:

- how to properly represent policies in terms of data analysis processes, data privacy concepts and related workflow constructs (ontological issue),
- how to automatically enforce those policies in data analysis process management within the workflow system (policy enforcement issue), and
- how to provide provenance regarding the privacy of the data as well as their data analysis history so that the system can justify its use of the data (provenance issue).

In this paper, we describe how a semantic approach can be adopted to address those issues in the context of workflow. In particular, we show how semantic web technology can be used to describe data analysis workflows as well as their data privacy requirements. Also, we explain how the privacy related ontology could be used together with some policy framework for representing privacy polices for controlling data analysis process creation and execution. Examples of possible privacy polices enabled by the introduced privacy awareness are provided. In addition, a case study is presented to illustrate how the proposed approach can be used to govern a particular distributed data mining process.

Motivation

Managing privacy in data analysis processes in e-Science has two different aspects of concern, namely *data* privacy and *process* privacy. The former one concerns the privacy of *data sources* as well as *data products* created during data analysis processes. The latter one concerns the privacy of knowledge captured in *data analysis processes*. In this paper, we mainly focus on protection of data privacy in the context of workflow.

Our Goal: Privacy-Aware Data Analysis

Instead of specifying policies to control data access, we set policies on types of data analysis that can be applied to the data. The following are some higher level expressions of the policies that we are targeting where notions like purposes of analysis, types of analysis, characteristics related to data privacy and analysis accuracy are involved to govern data analysis processes.

Example 1 *Patient medical images* should not be released for analysis except for the *purpose of supporting a particular medical image analysis project* and the images have to be *encrypted* if they are transmitted via untrusted networks.

Example 2 Given the *purpose of medical diagnosis*, any *classification* performed on *clinical data* must provide the *confidence level* for each data item and have its *overall accuracy* reaching a particular *level of quality standard*.

Example 3 Data containing *drug dosage information* should not be released for *any analysis* except for the *purpose of public health care study*, and the data should not contain *any personal identification attribute* and have to be *properly anonymized* before they can be used.

For the three examples provided, terms in italics reveal the need of a vocabulary to describe workflow-relevant concepts about data privacy and data analysis, and the remaining non-italic portions correspond to the constructs for describing policies in existing data access control frameworks. With a similar analogy, our proposed approach (to be shown in the later sections) also involves two parts, namely ontological description of privacy and data analysis, and the adoption of a policy framework with the ontological description integrated.

Related Research on Privacy Preserving Data Analysis

While restricting access to the data could be found to restrict to support various kind of data analysis, one could adopt the approach of restricting information in the data so that they are (a) free of *identifiers* that would permit linkages to any target individual and (b) free of *content* that would create unacceptably high risks of individual identification. For example, one may allow a set of data to be released and analyzed as far as fields related to personal information are anonymized.

In the literature, techniques for releasing data without disclosing sensitive information have been proposed for various applications. For example, cryptography-based techniques have been found useful in private data communication in untrusted networks (Stalling 2005). Techniques like anonymization (Samarati 2001) and microaggregation (Domingo-Ferrer & Mateo-Sanz 2002) have been found useful in applications like statistical disclosure control. Also, there has been recent research interest in developing data mining algorithms which are privacy preserving with underlying techniques including secure multiparty computation (Lin, Clifton & Zhu 2005), random data perturbation (Kargupta et al. 2005) and data generalization (Cheung et al. 2006).

Before we proceed, it is worth mentioning that our concern is not only limited to the identity threat. In fact, one could generalize the target to be hidden from individuals' identity to some important data attributes or experiment runs which will depend on the particular application and situation at hand.

Related Research on Policy Governed Data Analysis

As an alternative approach for privacy protection in data analysis, policies of data usage can be adopted for governing data analysis processes (Weitzner et al. 2006). For example, a research lab wants some of their on-line data and analysis tools to be only used for the purpose of demonstrating the system's analysis capability and thus posts a related data usage policy. In case the data set is later on found to be used (say together some other data sources) for re-identification disclosure of the subjects who provide the data, the one doing that will be accountable for the consequence. In addition, it will be even more appealing if such policy-violating data analysis processes can be caught early on and be stopped before they are actually executed.

While there has been work found in the literature for representing and reasoning about privacy policies (Bradshaw et al. 2003, Kagal, Finin & Joshi 2003), only conventional security concepts like authentication, authorization and encryption have been considered. We envision that privacy preserving data analysis techniques will soon get more mature and widely accepted. The family of privacy policies will need to be further enriched with the additional privacy related semantics being properly represented and reasoned.

Approach

Our approach to develop a privacy-aware data analysis framework is to extend existing workflow systems to incorporate privacy policies that control the type of data analysis done on the data. Modeling data analysis processes as workflows, also called scientific workflows, is common in e-Science (Deelman & Gil 2006, *Ludascher et al.* 2006, Oinn *et al.* 2006, Wassermann *et al.* 2006). However, in the literature, workflow systems possessing data privacy awareness are still lacking. Conventional workflow systems were designed with the primary objectives of providing component abstraction, interconnectivity and reliable execution in mind. In e-Science, data oriented and user (scientist) oriented perspectives have been stressed. Examples include the use of visual programming environments for constructing the data analysis processes, e.g., Taverna (Oinn *et al*. 2006), Kepler (Ludascher et al. 2006) and Sedna (Wassermann et

al. 2006), and the adoption of a template oriented approach for workflow creation, e.g., Wings (Gil et al. 2007).

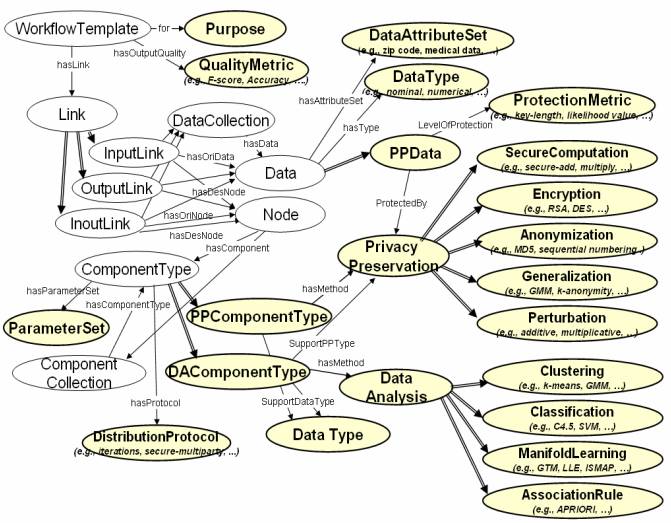

Figure 2. An ontology for describing privacy aware data analysis workflows.

Regarding data privacy control, most workflow systems are designed to support only the conventional data access polices for privacy control. If other types of privacy polices are to be respected, they can only be managed manually.

We propose a semantic approach for data privacy management in workflow systems. In particular, we first derive ontologies to describe fundamental concepts of data privacy and data analysis, and to integrate them into the ontological description of workflows to characterize their privacy related properties. Then, we argue that a new type of privacy policies can be specified using the derived ontologies and some policy description framework so that policy compliance test for data analysis workflows could be performed automatically via metadata reasoning. In the following, instead of providing a comprehensive view of every aspect of data privacy protection (e.g., user authentication, data/workflow access control, etc.), we focus on those more related to privacy and data analysis.

Building Blocks

Figure 2 depicts an ontology that contains most of the important concepts needed for describing privacy aware data analysis workflows. Note that the classes shown in normal face are specifically proposed as building blocks for describing essential constructs in workflows. They include WorkflowTemplate, Link, Node, Data and ComponentType which are adopted from Gil et al. 2007 and thus further details will not be provided. The classes in bold face as well as their related properties are building blocks for modeling privacy awareness as well as extensions of the workflow related ontologies constructed for embracing privacy awareness.

Privacy preservation ontology. This ontology enables us to describe workflow components that perform privacy preservation techniques. PrivacyPreservation is the root class, and the possible subclasses and their instances include

- **Encryption**, {e.g., RSA, DES, RC4}
- **Anonymization**, {e.g., MD5, sequential numbering}
- **Generalization**, {e.g., Gaussian mixture model, k-anonymity}
- **Perturbation**, {e.g., additive, multiplicative}
- **SecureComputation**, {e.g., secure-add, secure-multiply}

Data analysis ontology. This ontology gives the taxonomy of workflow components that process the data. We consider here statistical data analysis algorithms (Hastie, Tibshirani & Friedman 2003) that are widely used in many domains, but domain-specific analysis types would be part of the ontology as well (Cannataro et al. 2004). DataAnalysis is the root class, and the possible subclasses and their instances include:

- **Clustering**, {e.g., k-means, Gaussian mixture model)
- **Classification**, {e.g., C4.5, support vector machine}
- **Manifold Learning**, {e.g., GTM, ISOMAP, LLE}
- **Association Rule**, {e.g., Apriori}

Note that what being described is a very simple one. Data analysis is a mature area and a more comprehensive data analysis ontology which can describe the available tools (e.g., Bernstein, Provost & Hill 2005) should be needed.

Extensions of workflow related ontologies. Given the privacy preservation ontology and the data analysis ontology being available, the workflow template ontology is extended with privacy awareness as follow:

Two properties are added as the metadata of the workflow:

- **for** whose range captures the purpose of the workflow, e.g., medical diagnosis, public health study, and the purposes can be further described by a domain ontology specific to the application.
- **hasOutputQuality** whose range captures the semantic and metric of the workflow's output quality, e.g. F-score, classification accuracy, which again can be further described using an additional ***analysis quality ontology***.

Two properties are added as the metadata of the Data class (which can be the source or intermediate data products):

- **hasDataType** whose range can take values of *numerical*, *nominal*, *relational*, *semi-structured*, etc.
- **hasDataAttributeSet** which describes the schema of the data attributes which should further be described

by the domain specific ontology needed by **hasPurpose**.

Two properties and two subclasses are added to ComponentType:

- **hasParameterSet** whose range captures the set of parameters needed for configuring a particular component which is especially common for data analysis component.
- **hasProtocol** which is added to ComponentCollection and has its range referring to the distributed computing protocol needed among the components in the collection, e.g., iterative protocol for split-and-merge, secure-multiparty computation protocol.
- **PPComponentType** and **DAComponentType** which are subclasses of ComponentType specialized in privacy preservation and data analysis respectively. The latter one, in addition, carries three new properties. Two of them are supportPPType and supportDataType for characterizing what kind of the data its instance can process and the other one is hasOutputQuality which is semantically the same as that of the WorkflowTemplate class.

A new class called **PPData** is introduced as a subclass of Data to represent privacy preserved data types and it has two properties:

- **ProtectedBy** whose range captures the privacy preservation technique being adopted, e.g. encryption, anonymization.
- **LevelofProtection** whose range captures the level of protection possessed by the data. While there doesn't exist so far a universal way to specify the level of protection, different privacy preservation techniques are associated with their ways of specifying the level. For example, for encryption, the key length is a good indicator of the protection level. For anonymization, the size of the smallest indistinguishable set is a good indicator. Related knowledge can be further described in the privacy preservation ontology.

Some of the concepts related to data and data analysis components are highly related to the specification of Predictive Model Markup Language (PMML: http://sourceforge.net/projects/pmml). PMML is an XML-based language which can be used for describing data mining models to facilitate their exchange across different platforms. Because of its goal, elements for describing details of data structures (like matrices) and mining model schema are included. To contrast with the goal of PMML, ours is situated at a higher level for data analysis validation and leaves the detailed interoperability issue to the lower level of the stack of the workflow system.

Privacy Policies for Data Analysis Workflows: Evaluation of some existing policy frameworks

Given the ontologies derived in the previous section, we can have a better idea what privacy awareness can be incorporated for data analysis can have in addition to the conventional user authentication and authorization. This implies opportunities for new types of privacy policies for data analysis to be derived and at the same time challenges regarding how the polices are going to be represented and reasoned together with the workflow's metadata for policy compliance checking.

KAoS (Bradshaw et al. 2003) and Rei (Kagal, Finin & Joshi 2003) are two representative projects that make use of Semantic Web technology to specify privacy related policies. The former one was proposed in the context of multiagent systems and the latter one was proposed in the context of pervasive computing. Their latest versions follow RDF-Scheme based syntax and support four types of policies including positive authorization, negative authorization, positive obligation, and negative obligation (Tonti et al. 2003). Authorization refers to the notion that some action is permitted or not. Obligation refers to the notion that some action has to and should not be done given a certain condition.

Both projects provide ontologies for specifying policies with the concepts of like resource, actors, actions, context, policy, control, etc. included and are expected to be further extended in application specific way. Take KAoS as an example. If ones want to describe privacy policies which can support also the notion of privacy awareness presented in the previous section, they can map the ComponentType class to the Action class in KAoS and so that PPComponentType and DAComponentType will be referred to the privacy preservation actions and data analysis actions respectively. The properties adhering to them as metadata become the context of those actions in the terminology of KAoS.

To contrast with our need of policies for controlling data privacy, KAoS primarily focuses on real-time communication and interaction among software agents, and thus related dynamics are carefully modeled in them. For data analysis workflows in e-Science, those constructs are of less relevance since a relatively reliable distributed computing environment is mostly assumed. Instead, the concern, which is also the theme of this paper, is more on the data, the overall accuracy of the analysis and the privacy issues within an execution. Also, as numerical measures are often involved in the context of privacy awareness, ways to formally incorporate them in the policy conditions and have an engine that can reason upon them obviously is an open issue. Last but not the least, the enriched set of privacy preservation and data analysis

techniques open up opportunities for new privacy policies to be specified (as to be detailed in the following sub-section) and at the same time impose new challenges on how those policies can be automatically enforced. More effort in formalizing and addressing the underlying challenges in a disciplined manner is essential.

Examples of privacy policies needed in data analysis workflow

To help better illustrate the new set of privacy policies needed for data analysis workflows, we present the following examples and relate them with some possible real scenarios. Some of them are similar to what we show in the Motivation section but with further contextual details included using the vocabulary provided by the ontologies we described in the previous section.

- Policies governing privacy in data transmission

Policy 1: Patient medical images (data) should not be the input of any data analysis component or the output of the overall workflow except for the purpose of medical diagnosis, and the images have to be encrypted using RSA with a key length of at least 128 bits before transmitted.

As the analysis is related to medical image comparison, privacy preservation techniques like generalization and perturbation will not be suitable or the image quality will be degraded. The encryption approach fits well to the situation as stated in Policy 1. An example of a research initiative where distributed images are shared is The Biomedical Informatics Research Network (http://www.nbirn.net/).

- Policies governing privacy in data source selection

Policy 2: Data containing drug dosage information should not be the input of any data analysis component or the output of the overall workflow except for the purpose of public health care study, and the data should not contain person identification attributes and should be generalized by at least 5-anonymity.

Health care data are known to be sensitive, especially when they touch on critical issues like mental health. As healthcare data analysis mostly only focuses on trends and patterns and less on particular cases, the generalization approach fits well. Policy 2 or similar ones are relevant to the analysis conducted at health related organization like National Survey of Family Growth (http://www.cdc.gov/nchs/nsfg.htm) and National Institute of Health (http://www.nih.gov/).

Policy 3: Data set satisfying k-anonymity as required for its privacy protection should not be the input of any data analysis component if the component carries more one input data sources and they have overlapping attributes.

This policy exemplifies the need to handle the potential of information bleaching via combining multiple privacy preserved data. While more rigorous explanation on why Policy 3 is essential is a bit out of the tone of this paper, only the intuitive idea is provided. There are possibilities that data items in different input sources (e.g., clinical records and on-line phone books) may in fact refer to the same entities. If the association of the multiple data sources can be derived via the overlapping attributes, one could learn more about those entities by studying their group attributes in the multiple sources. In general, more research attention is needed to handle situations with multiple sources of information aggregated, even if each of them is privacy preserved. The case study to be presented in the next section will further elaborate this point.

- Policies governing privacy in intermediate data products

Policy 4: Given that the purpose of the analysis may lead to critical action taking (e.g., patient isolation), the confidence level of the final output should be available and be higher than an associated threshold or the workflow execution should be considered unsatisfactory and halted.

This policy governs the analysis output whose quality may vary depending on the data analysis algorithm itself (e.g., a training phase is needed during the analysis) and the imprecision of the input data, possibly caused by privacy preservation. As the quality measure will not be known during the workflow creation phase, this type of policies can only be enforced during the workflow execution.

Case Study – Distributed Data Clustering

Suppose that there is a clinical trial study which is related to mental health. Data are collected from patients at different clinics and contain a range of medical measurements, drug dosage, as well as patient related demographic information. The set of attributes are not identical for the data sets collected from different clinics but all of them are supposed to have the personal identification fields removed before releasing (Policy CS1). Also, the data are restricted from further analysis unless the patients' data are further generalized into groups, each satisfying k-anonymity and being represented by only the first and second order statistics (Policy CS2).

Clustering has been a common technique for discovering patterns in datasets. So, one of the analysis tasks under this study is to apply clustering to the combined dataset to identify patient groups with similar medical measurements under a certain amount of drug dosage. The researcher uses

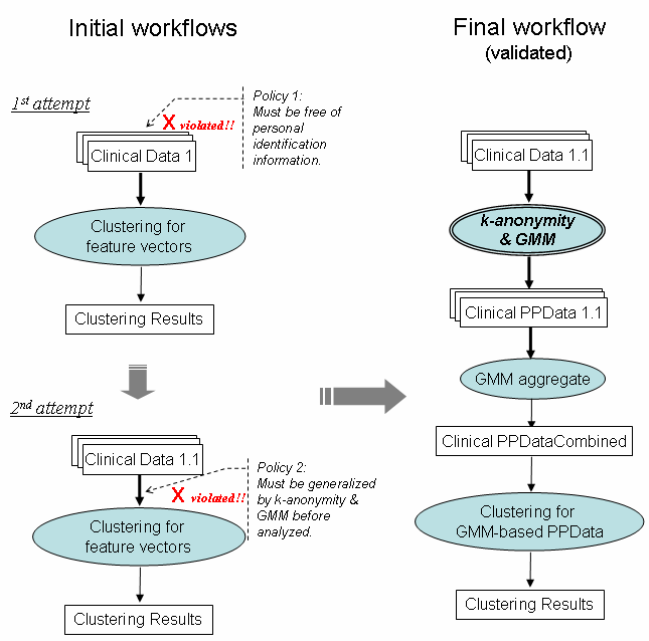

Figure 3. Creation of a distributed data clustering workflow.

a privacy aware workflow system as what we have described. Also, we assume that the data provided by the hospitals and clinics have their metadata accurately tagged.

The researcher creates a workflow template and puts directly all the data together and feeds them into a clustering component. The workflow system applies the policies to the data analysis workflow and finds that one data source still contains the patient name as revealed in the metadata (DataAttributeSet). This violates Policy CS1 and thus the system prompts the user to transform the dataset to address this issue.

The workflow system finds that Policy CS1 is respected this time but violation of Policy CS2 is detected instead as the data fed into the clustering component are not privacy protected as required. The researcher learns from the system about Policy CS2 and thus feeds the data first to a correct generalization-based privacy preserving components before going into the clustering component. With both policies satisfied, the researcher forgets that the original clustering component he selected does not support privacy preserved data. The workflow system detected the mismatch with the help of the metadata captured in instances of PPData (via ProtectedBy) and DAComponentType (via SupportPPType), and prompts the researcher about the mismatch problem. As suggested by the system, the researcher switches the clustering component to one that can support clustering of PPData abstracted as GMMs (Cheung et al 2006). The system eventually finds the data analysis flow valid (as shown in Figure 3) and thus executes it.

Later on, it is brought to the attention of the researcher that one patient's identity is revealed. The researcher uses the workflow system's provenance function enabled by the metadata and manages to trace back to the clustering execution just explained. After careful investigation, it is found that there are some unexpected cases where the same patient went to several clinics and the complaining patient is one of those cases. As the same patient falls into different groups at different hospitals, their intersections can thus be uniquely characterized and thus the k-anonymity property no longer holds. From this, the researcher learns that Policy CS2 is insufficient and needs to be revised. Instead, it should be the combined dataset that needs to satisfy the k-anonymity instead of only those before the combination. The researcher simply needs to modify the policy so that the workflow system can avoid similar incidents from happening again in the future.

The data analysis task being described in this case study is a relatively simple one and the goal is to show how a workflow system with privacy awareness embedded would operate. As can be read from the case study, the policies involved apply directly to some particular links in the workflow and the corresponding implementation should not be a big challenge. Whether there are needs for privacy policies to be described in a more holistic sense, and how can these global constraints be decomposed and propagated to the corresponding parts of the workflow for privacy controls are interesting issues worth future investigation.

Conclusion and Future Work

In this paper, we motivated the need for managing privacy in data analysis workflows so that a new type of privacy policies that constrain processing on data can be supported. We also 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 analysis-relevant terms can be defined in ontologies, and how they can be combined within a policy framework to represent the policies. Finally, we discussed how those policies can be applied via various examples, potential areas of application and a detailed case study. We believe that workflow systems with the proposed privacy-awareness incorporated could ease the scientists in setting appropriate privacy polices that suit for different types of collaborative research projects and at the same time can help them in safeguarding the privacy of sensitive data throughout the data analysis lifecycle.

We are currently implementing the approach by extending the Wings framework. The ontologies and policies shown in the paper will be represented in OWL and SWRL so that the policy enforcement can be carried out by the workflow system with the help of OWL reasoners. How to design a privacy policy framework which suits best for data analysis

is no doubt an open research issue. Also, as hinted in the Case Study section of the paper, a related open issue is to gain further understanding on the full spectrum of the policies needed to be represented in the policy framework.

Extending the focus from merely data privacy to also process privacy is another important direction of investigation. Recently, the need of sharing experimental processes among scientific have been identified to be important in further facilitating collaboration knowledge discovery in empirical science. Related collaborative process sharing tools (e.g., myExperiment.org) have been built to ease the corresponding sharing management. However, workflow systems with the policy-based enforcement as described in this paper incorporated for controlling workflow provenance sharing are still lacking and should form an important compliment of what we discussed in this paper.

Acknowledgements

This research was supported in part by the Air Force Office of Scientific Research (AFOSR) through grant FA9550-06-1-0031.

References

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. 2003. 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*, 835--842, New York, ACM Press.

Beltrame, F., Canesi, B., Molinari, E., Porro, I., Schenone, A., Torterolo, L. 2006. *Book of Abstracts of Enabling Grids for E-sciencE User Forum*. (http://egee-intranet.web.cern.ch/egee-intranet/User-Forum/2006/)

Bernstein, A., Provost, F. and Hill S. 2005. Toward Intelligent Assistance for a Data Mining Process: An Ontology-Based Approach for Cost-Sensitive Classification. *IEEE Transactions on Knowledge and Data Engineering*, 17(4), 503-518.

Cannataro, M., Comito, C., Schiavo, F.L. and Veltri, P. 2004. Proteus, a Grid based Problem Solving Environment for Bioinformatics: Architecture and Experiments. *IEEE Computational Intelligence Bulletin*, 3(1), 7-18.

Cheung, W.K., Zhang, X., Wong, H., Liu, J., Luo, Z. And Tong, F. 2006. Service-oriented Distributed Data Mining, *IEEE Internet Computing*, 10(4), 44-54.

Clifton, C., Kantarcioglu, M., Vaidya, J., Lin, X. and Zhu, M. 2003. Tools for Privacy Preserving Distributed Data Mining. *ACM SIGKDD Explorations*, 4(2), 19-26.

Deelman E. and Gil, Y. 2006. Final Report of the NSF Workshop on the Challenges of Scientific Workflows. (http://vtcpc.isi.edu/wiki/images/3/3a/NSFWorkflowFinal.pdf)

Domingo-Ferrer, J., and Mateo-Sanz, J.M. 2002. Practical Data-oriented Microaggregation for Statistical Disclosure Control. *IEEE Transactions on Knowledge & Data Engineering*, 14(1), 189-201.

Gil, Y., Ratnakar, V., Deelman, E., Mehta, G. and Kim, J. 2007. Wings for Pegasus: Creating Large-Scale Scientific Representations of Computational Workflows. To appear in *Proceedings of the 19th Annual Conference on Innovative Applications of Artificial Intelligence (IAAI).*

Hastie, T., Tibshirani, R. and Friedman J.H. 2003. The *Element of Statistical Learning*, Springer-Verlag.

Kagal, L. Finin, T., and Joshi, A. 2003. A Policy Language for Pervasive Systems. In *Proceedings of the Fourth IEEE International Workshop on Policies for Distributed Systems and Networks*, 63-76.

Kargupta, H., Datta, S., Wang, Q., and Sivakumar, K. 2005. Random Data Perturbation Techniques and Privacy-preserving Data Mining. *Knowledge and Information Systems*, 7(4), 387-414, New York, Springer Verlag.

Lin, X., Clifton, C., and Zhu, M. 2005. Privacy-preserving clustering with distributed EM mixture modeling. *Knowledge and Information Systems*, 8(1), 68--81, New York, Springer-Verlag.

Ludascher, B., Altintas, I., Berkley, C., Higgins, D., Jaeger, E., Jones, M., Lee, E.A., Tao, J., and Zhao, Y. 2006. Scientific Workflow Management and the Kepler System. *Concurrency Computation: Practice & Experience*, 18(10), 1039-1065.

NIH. 2004. *Data Sharing Workbook*, National Institutes of Health. (http://grants2.nih.gov/grants/policy/data_sharing/data_sharing_workbook.pdf)

Oinn, T., Greenwood, M., Addis, M., Alpdemir, M.N., Ferris, J., Glover, K., Goble, C., Goderis, A., Hull, D., Marvin, D., Li, P., Lord, P., Pocock, M.R., Senger, M., Stevens, R., Wipat, A. and Wroe, C. 2006. Taverna: Lessons in Creating a Workflow Environment for the Life Sciences, *Concurrency Computation: Practice & Experience*, 18(10), 1067-1100.

Samarati, P. 2001. Protecting Respondents' Identities in Microdata Release. *IEEE Transactions on Knowledge and Data Engineering*, 13(6), 1010-1027.

Singh M.P. and Vouk, M.A. 1996. Scientific Workflows: Scientific Computing meets Transactional Workflows. In *Proceedings of the NSF Workshop on Workflow and Process Automation in Information Systems: State-of-the-Art and Future Directions*, SUPL28--34.

Stalling, W. 2005. *Cryptography and Network Security: Principles and Practices* (4th ed.). Prentice Hall.

Tonti, G., Bradshaw, J., Jeffers, R., Montanari, R., Suri, N. and Uszok A. 2003. Semantic Web Languages for Policy Representation and Reasoning: A Comparison of KAoS, Rei, and Ponder. In *Proceeding of 2nd International Semantic Web Conference*, *Lecture Notes in Computer Science*, 2870, 419-437, Springer-Verlag.

Wassermann, B., Emmerich, W., Butchart, B., Cameron, N., Chen, L. and Patel, J. 2006. Sedna: A BPEL-based Environment for Visual Scientific Workflow Modeling. In *Workflows for eScience - Scientific Workflows for Grids*, Taylor, I.J., Deelman, E., Gannon, D., and Shields M.S. (eds.), Springer Verlag.

Weitzner, D.J., Abelson, H., Berners-Lee, Tim, Hanson, C., Hendler, J., Kagal, L., McGuinness, D.L., Sussman, G.J., Waterman, K.K. 2006. Transparent Accountable Data Mining: New Strategies for Privacy Protection, Technical Report, MIT-CSAIL-TR-2006-007, MIT.