Towards Privacy Aware Data Analysis Workflows for e-Science

William K. Cheung
Department of Computer Science
Hong Kong Baptist University
William@comp.hkbu.edu.hk

Yolanda Gil
Information Sciences Institute
And Department of Computer Science
University of Southern California
gil@isi.edu
Outline

• Motivation: e-Science meets Privacy
• Current privacy protection techniques
• Privacy in workflow systems
• Enforcing privacy in workflow systems
  – Representation of privacy policies
  – Policy compliance checking
• Conclusions
Motivation: e-Science meets Privacy

• Tremendous advantages of use of medical records in medical studies [Hodge et al 99]
  – Phenotype use for genomics [Dugas et al 02]
  – Biomedical imaging [www.nbirm.net]
  – Cancer research [cabig.nci.nih.gov]

• Privacy of medical records:
  – Tradeoff with quality of treatment [Simons et al 05]
  – Incentives: first access to new treatments [Kohane and Altman 05]
  – Altruism [Mandl et al 01]

• Giving up privacy for pre-specified uses (eg study)
  – Not for insurance purposes, employers, other studies
Privacy and the Open Web

• Mounds of sensitive data about individuals is readily available in the open web
  – Can combine with private sources of data [Dan et al 05]
• Open web already contains sensitive information that should not be available and violates privacy acts [Sweeney 04]
• Data mining and record linkage techniques aggregate information and identify individuals [Sweeney 97]
OECD’s Eight Principles of Fair Information Practices

<table>
<thead>
<tr>
<th>Principle</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection limitation</td>
<td>The collection of personal information should be limited, should be obtained by lawful and fair means, and, where appropriate, with the knowledge or consent of the individual.</td>
</tr>
<tr>
<td>Data quality</td>
<td>Personal information should be relevant to the purpose for which it is collected, and should be accurate, complete, and current as needed for that purpose.</td>
</tr>
<tr>
<td>Purpose specification</td>
<td>The purposes for the collection of personal information should be disclosed before collection and upon any change to that purpose, and its use should be limited to those purposes and compatible purposes.</td>
</tr>
<tr>
<td>Use limitation</td>
<td>Personal information should not be disclosed or otherwise used for other than a specified purpose without consent of the individual or legal authority.</td>
</tr>
<tr>
<td>Security safeguards</td>
<td>Personal information should be protected with reasonable security safeguards against risks such as loss or unauthorized access, destruction, use, modification, or disclosure.</td>
</tr>
<tr>
<td>Openness</td>
<td>The public should be informed about privacy policies and practices, and individuals should have ready means of learning about the use of personal information.</td>
</tr>
<tr>
<td>Individual participation</td>
<td>Individuals should have the following rights: to know about the collection of personal information, to access that information, to request correction, and to challenge the denial of those rights.</td>
</tr>
<tr>
<td>Accountability</td>
<td>Individuals controlling the collection or use of personal information should be accountable for taking steps to ensure the implementation of these principles.</td>
</tr>
</tbody>
</table>

- Consented collection
- Right to know and correct personal info
- Quality control
- Safe storage
- Access control to DBs
- Limited disclosure
OECD’s Eight Principles of Fair Information Practices

<table>
<thead>
<tr>
<th>Principle</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Collection limitation</td>
<td>The collection of personal information should be limited, should be obtained by lawful and fair means, and, where appropriate, with the knowledge or consent of the individual.</td>
</tr>
<tr>
<td>Data quality</td>
<td>Personal information should be relevant to the purpose for which it is collected, and should be accurate, complete, and current as needed for that purpose.</td>
</tr>
<tr>
<td>Purpose specification</td>
<td>The purposes for the collection of personal information should be disclosed before collection and upon any change to that purpose, and its use should be limited to those purposes and compatible purposes.</td>
</tr>
<tr>
<td>Use limitation</td>
<td>Personal information should not be disclosed or otherwise used for other than a specified purpose without consent of the individual or legal authority.</td>
</tr>
<tr>
<td>Security safeguards</td>
<td>Personal information should be protected with reasonable security safeguards against risks such as loss or unauthorized access, destruction, use, modification, or disclosure.</td>
</tr>
<tr>
<td>Openness</td>
<td>The public should be informed about privacy policies and practices, and individuals should have ready means of learning about the use of personal information.</td>
</tr>
<tr>
<td>Individual participation</td>
<td>Individuals should have the following rights: to know about the collection of personal information, to access that information, to request correction, and to challenge the denial of those rights.</td>
</tr>
<tr>
<td>Accountability</td>
<td>Individuals controlling the collection or use of personal information should be accountable for taking steps to ensure the implementation of these principles.</td>
</tr>
</tbody>
</table>

- Collection for a purpose
- Use only for authorized purpose
- Accountability throughout these principles
Some Examples

• **Example 1**
  – People may allow any study to *use* their data if they will be stored only in *aggregated, non-identifiable form*.

• **Example 2**
  – Patients may allow their data being *used for medical research* but not *for other reasons* (e.g. *investigating more profitable insurance rates*).

• **Example 3**
  – Scientists may allow their data to be *used for cross-disciplinary research investigation* but not *for competing research quests on similar research areas*.

• **Example 4**
  – Patients may allow their data to be *used* but not *if anonymity is compromised above a threshold k* (e.g. *k-size clusters*).
Ensuring Privacy Principles are Respected

• What technologies could enforce privacy policies concerning the *use* of sensitive information?
  – Limit the use of sensitive data to those authorized during collection

• What technologies could ensure accountability of underlying systems?
  – On-demand examination by audit or by user
Computational Workflows

- Interdependent sets of computations
- Dependencies are data flow
- Computations can be submitted for execution in various remote resources
- Input data may be obtained from remote data repositories
- New data products may be stored in remote data repositories
Data Analysis/Mining Processes as Workflows
Workflows in Wings/Pegasus for Seismic Hazard Analysis [Gil et al IAA-07]

- Input data: a site and an earthquake forecast model
  - thousands of possible fault ruptures and rupture variations, each a file, unevenly distributed
  - \( \sim 110,000 \) rupture variations to be simulated for a given site
- 8043 application nodes in the workflow instance generated by Wings
- 24,135 nodes in the executable workflow generated by Pegasus, including:
  - data stage-in jobs, data stage-out jobs, data registration jobs
- Executed in USC HPCC cluster, 1820 nodes w/ dual processors) but only < 144 available
  - Including MPI jobs, each runs on hundreds of processors for 25-33 hours
  - Runtime was 1.9 CPU years
- Significant contribution to create a more accurate seismic hazard map for SoCal
  - First integration of multiple physics-based models
  - Currently fine-tuning and cross-validating models
- Provenance records of workflow creation and execution

Yolanda Gil  gil@isi.edu
Approach

• Extend current workflow systems to be privacy-aware

- Ontologies for privacy preservation and data analysis processes
- Policy rules governing data processing and analysis
- Workflow systems with policies checking capability

Yolanda Gil  gil@isi.edu
Privacy-Preserving Workflows

**P1:** No personal ID information can leave the data source

**P2:** Sensitive data must be $k$-anonymized

Centralized workflow not compliant with policies

- **Loc1**
  - ParameterFile
  - dataSet
  - dataSet
  - dataSet

- **Loc2**
  - Aggregation
  - ParameterFile
  - dataSet

- **Loc3**
  - Analysis
  - ClustersWithDataitems

Distributed workflow compliant with policies

- **Loc1 .. Loc n**
  - Anonymization
  - ParameterFile
  - dataSet

- **Loc2**
  - Abstraction
  - ParameterFile
  - ClustersWithStatistics

- **Loc3**
  - Analysis
  - GMMAggregate
A Semantic Workflow Approach

- Ontologies for privacy preservation and data analysis processes
- Policy rules governing data processing and analysis
- Workflow systems with policies checking capability

Yolanda Gil  gil@isi.edu
A Semantic Approach to Policy-Aware Workflow Systems

- **Need to describe process declaratively**
  - Input, output and intermediate files (data/parameters)
  - Analysis components
  - Process flow (links)
  - Data movement
  - Execution conditions

- **Need to enforce policies**
  - Policy reasoning
    - *E.g., constraints on data usage, data movement, types of processing, execution, …*
  - Policy enforcement within system

- **Need to record process and compliance**
  - Provenance trails
Workflow Components for Privacy
Preserving Data Analysis

- **Objective 1:** Data free of *identifiers* linking to any target individual.
  - **Anonymization** (e.g., “{Alice, id111, i1, i2}” -> “{X, *, i1, i2}”)
    - *Method:* Recode or mask data attributes [Samarati 2001]
    - *Applied to:* Association Rule Mining [Lakshmanan et al. 2005]

- **Objective 2:** Data free of *content* leading to high risks of individual identification.
  - **Perturbation** (e.g., “(1,0), (1,1), (0,1)” -> “(2,-1), (1,-1),(1,0)”)
    - *Method:* Add random perturbation to data
    - *Applied to:* Regression, Classification [Du et al. 2003, Liu et al. 2006]
  - **Generalization** (e.g., “(1,0), (1,1), (0,1)” -> “3 , (0.67,0.67)”)
    - *Method:* Abstract data using local statistics
    - *Applied to:* Clustering, Manifold Learning [Klusch et al. 2003]
A Semantic Workflow Approach

Ontologies for privacy preservation and data analysis processes

Workflow systems with policies checking capability

Policy rules governing data processing and analysis

Yolanda Gil  gil@isi.edu
Examples of Privacy Policies for Data Analysis

• **Policy 1**  *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.

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

• **Policy 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 data that contain dosage information, it is required that they are first abstracted by k-anonymity before being further analysed.”

context:

hasLink(?w, ?l) \^ hasFile(?l, ?d) \^ hasAttribute(?d, ?a) \^ Dosage(?a) \^ hasDestinationNode(?l, ?n) \^ hasComponent(?n, ?c) \^ DAComponent(?c)

usage:

NULL

protection:

+ve: abstracted(?d, ?dgVal) \^ equal(?dgVal, true) \^ abstractedBy(?d, ?m) \^ k-anonymity(?m)

correction:

prompt [add a k-anonymity step right after (?d) found at (?l) ]
Policy Languages/Frameworks

Classification

Well-defined Semantics

- RBAC
- ACL
- Ponder

No Formal Semantics

- Kaos
- Rei
- XACML
- P3P

- PSPL
- SD3, RT
- PeerTrust
- Cassandra
- Protune
- PeerAccess

Centralized Evaluation

Distributed Policies, Centralized Evaluation

Distributed Evaluation

Extracted from ESWC'06 Tutorial “Semantic Web Policies: Where are we and What is Missing” by Piero A. Bonatti and Daniel Olmedilla

Rolanda Gil  gil@isi.edu
A Semantic Workflow Approach

Ontologies for privacy preservation and data analysis processes

Policy rules governing data processing and analysis

Workflow systems with policies checking capability

Yolanda Gil  gil@isi.edu
A Workflow for Clustering Clinical Data in Wings
Policy Compliance Checking Via Reasoning
Initial workflows

1st attempt

Clinical Data 1

\[ \text{Clustering for feature vectors} \]

Clustering Results

\[ \times \text{violated!!} \]

2nd attempt

Clinical Data 1.1

\[ \text{Clustering for feature vectors} \]

Clustering Results

\[ \times \text{violated!!} \]

Policy 1: Must be free of personal identification information.

Final workflow (validated)

Clinical Data 1.1

\[ \text{k-anonymity & GMM} \]

GMM aggregate

Clinical PPData 1.1

Clinical PPDataCombined

\[ \text{Clustering for GMM-based PPData} \]

Clustering Results

Policy 2: Must be generalized by k-anonymity & GMM before analyzed.
I cannot access D2, D3 and D4. I cannot proceed with my analysis?
The workflow system suggests this privacy preserving workflow for my analysis. The result looks good enough!
Conclusions

• Policy-aware workflow systems that can
  – Represent privacy policies that concern the use of sensitive information
  – Analyze tradeoffs between privacy and accuracy
  – Enforce privacy policies at all levels of the workflow creation and execution process
  – Accountability through detailed provenance records