Crafting a Balance between Big Data Utility and Protection in the Semantic Data Cloud

Yuh-Jong Hu    Kua-Ping Cheng    Ya-Ling Huang
{hu, 99753025, 99753026}@cs.nccu.edu.tw

Emerging Network Technology (ENT) Lab.
Department of Computer Science
National Chengchi University, Taipei, Taiwan

June-12-2013

International Conference on
Web Intelligence, Mining, and Semantics (WIMS’13)
Motivations

1. How to effectively collect and analyze complex big data, including structured and unstructured, is hot but the related privacy issue does not arise much attention.

2. Statistical Disclosure Control (SDC) for microdata protection has been well-established so this is a good starting point.

3. How to achieve a balance between big data utility and privacy protection through the combination of SDC and Semantic Web techniques?

4. Solving a complex big data utility and protection problem requires a multi-disciplinary approach, including statistics and computer science.
Motivations

1. How to effectively collect and analyze complex big data, including structured and unstructured, is hot but the related privacy issue does not arise much attention.

2. Statistical Disclosure Control (SDC) for microdata protection has been well-established so this is a good starting point.

3. How to achieve a balance between big data utility and privacy protection through the combination of SDC and Semantic Web techniques?

4. Solving a complex big data utility and protection problem requires a multi-disciplinary approach, including statistics and computer science.
Motivations

1. How to effectively collect and analyze complex big data, including structured and unstructured, is hot but the related privacy issue does not arise much attention.

2. Statistical Disclosure Control (SDC) for microdata protection has been well-established so this is a good starting point.

3. How to achieve a balance between big data utility and privacy protection through the combination of SDC and Semantic Web techniques?

4. Solving a complex big data utility and protection problem requires a multi-disciplinary approach, including statistics and computer science.
Motivations

1. How to effectively collect and analyze complex big data, including structured and unstructured, is hot but the related privacy issue does not arise much attention.

2. Statistical Disclosure Control (SDC) for microdata protection has been well-established so this is a good starting point.

3. How to achieve a balance between big data utility and privacy protection through the combination of SDC and Semantic Web techniques?

4. Solving a complex big data utility and protection problem requires a multi-disciplinary approach, including statistics and computer science.
Research Goals

1. How can we provide semantic metadata markup services for structured data to establish a semantic data cloud?

2. How can we provide data integration and protection services within an outsourcing homogeneous data source for effective microdata analysis without fear of illegal data disclosure?

3. How can we apply data exchange and protection services across outsourcing heterogeneous data sources to have effective microdata sharing and analysis without fear of illegal data leakage?

4. How can we design and implement semantics-enabled policy of SDC for data protection while enforcing data analysis?
Research Goals

1. How can we provide semantic metadata markup services for structured data to establish a semantic data cloud?

2. How can we provide data integration and protection services within an outsourcing homogeneous data source for effective microdata analysis without fear of illegal data disclosure?

3. How can we apply data exchange and protection services across outsourcing heterogeneous data sources to have effective microdata sharing and analysis without fear of illegal data leakage?

4. How can we design and implement semantics-enabled policy of SDC for data protection while enforcing data analysis?
Research Goals

1. How can we provide semantic metadata markup services for structured data to establish a semantic data cloud?

2. How can we provide data integration and protection services within an outsourcing homogeneous data source for effective microdata analysis without fear of illegal data disclosure?

3. How can we apply data exchange and protection services across outsourcing heterogeneous data sources to have effective microdata sharing and analysis without fear of illegal data leakage?

4. How can we design and implement semantics-enabled policy of SDC for data protection while enforcing data analysis?
Research Goals

1. How can we provide semantic metadata markup services for structured data to establish a semantic data cloud?

2. How can we provide data integration and protection services within an outsourcing homogeneous data source for effective microdata analysis without fear of illegal data disclosure?

3. How can we apply data exchange and protection services across outsourcing heterogeneous data sources to have effective microdata sharing and analysis without fear of illegal data leakage?

4. How can we design and implement semantics-enabled policy of SDC for data protection while enforcing data analysis?
Contributions

1. Propose concepts of a semantic big data analysis pipeline to enable automated data analysis, protection, and interpretation services.

2. Semantics-enabled policies, as a combination of ontologies and rules, are represented and enforced for big data in the statistical databases.

3. Provide transparent SDC selection techniques for data users on solving a data analysis and protection of the statistical databases.

4. Preliminary results are discovered on crafting a balance between data utility and protection through enforcing semantics-enabled policies.
Contributions

1. Propose concepts of a semantic big data analysis pipeline to enable automated data analysis, protection, and interpretation services.

2. Semantics-enabled policies, as a combination of ontologies and rules, are represented and enforced for big data in the statistical databases.

3. Provide transparent SDC selection techniques for data users on solving a data analysis and protection of the statistical databases.

4. Preliminary results are discovered on crafting a balance between data utility and protection through enforcing semantics-enabled policies.
Contributions

1. Propose concepts of a semantic big data analysis pipeline to enable automated data analysis, protection, and interpretation services.

2. Semantics-enabled policies, as a combination of ontologies and rules, are represented and enforced for big data in the statistical databases.

3. Provide transparent SDC selection techniques for data users on solving a data analysis and protection of the statistical databases.

4. Preliminary results are discovered on crafting a balance between data utility and protection through enforcing semantics-enabled policies.
Contributions

1. Propose concepts of a semantic big data analysis pipeline to enable automated data analysis, protection, and interpretation services.

2. Semantics-enabled policies, as a combination of ontologies and rules, are represented and enforced for big data in the statistical databases.

3. Provide transparent SDC selection techniques for data users on solving a data analysis and protection of the statistical databases.

4. Preliminary results are discovered on crafting a balance between data utility and protection through enforcing semantics-enabled policies.
Semantics-enabled policies are composed of ontologies and rules, where ontologies are used for describing the concepts of data analysis and protection, and rules are used for enforcing the principles of data analysis and protection.

Semantics-enabled policies, ACP, DHP, and DRP are respectively correspond to, query restriction, data manipulation, and output perturbation for microdata protection.

- Access Control Policy (ACP) provides restricted Pattern-Based Queries (PBQs) through Datalog rules.
- Data Handling Policy (DHP) provides data usage conditions matching between data owners’ privacy preferences and users’ usage context.
- Data Releasing Policy (DRP) describes what are available SDC methods with de-identifiable PII are disclosed for analysis but data privacy is preserved.
Semantics-enabled Policies

1. Semantics-enabled policies are composed of ontologies and rules, where ontologies are used for describing the concepts of data analysis and protection, and rules are used for enforcing the principles of data analysis and protection.

2. Semantics-enabled policies, **ACP**, **DHP**, and **DRP** are respectively correspond to, query restriction, data manipulation, and output perturbation for microdata protection.
   - Access Control Policy (ACP) provides restricted Pattern-Based Queries (PBQs) through Datalog rules.
   - Data Handling Policy (DHP) provides data usage conditions matching between data owners’ privacy preferences and users’ usage context.
   - Data Releasing Policy (DRP) describes what are available SDC methods with de-identifiable PII are disclosed for analysis but data privacy is preserved.
Semantics-enabled policies are composed of ontologies and rules, where ontologies are used for describing the concepts of data analysis and protection, and rules are used for enforcing the principles of data analysis and protection.

Semantics-enabled policies, ACP, DHP, and DRP are respectively correspond to, query restriction, data manipulation, and output perturbation for microdata protection.

- Access Control Policy (ACP) provides restricted Pattern-Based Queries (PBQs) through Datalog rules.
- Data Handling Policy (DHP) provides data usage conditions matching between data owners’ privacy preferences and users’ usage context.
- Data Releasing Policy (DRP) describes what are available SDC methods with de-identifiable PII are disclosed for analysis but data privacy is preserved.
Semantics-enabled policies are composed of ontologies and rules, where ontologies are used for describing the concepts of data analysis and protection, and rules are used for enforcing the principles of data analysis and protection.

Semantics-enabled policies, **ACP**, **DHP**, and **DRP** are respectively correspond to, query restriction, data manipulation, and output perturbation for microdata protection.

- **Access Control Policy (ACP)** provides restricted Pattern-Based Queries (PBQs) through Datalog rules.
- **Data Handling Policy (DHP)** provides data usage conditions matching between data owners’ privacy preferences and users’ usage context.
- **Data Releasing Policy (DRP)** describes what are available SDC methods with de-identifiable PII are disclosed for analysis but data privacy is preserved.
Semantics-enabled Policies

1. Semantics-enabled policies are composed of ontologies and rules, where ontologies are used for describing the concepts of data analysis and protection, and rules are used for enforcing the principles of data analysis and protection.

2. Semantics-enabled policies, ACP, DHP, and DRP are respectively correspond to, query restriction, data manipulation, and output perturbation for microdata protection.

- Access Control Policy (ACP) provides restricted Pattern-Based Queries (PBQs) through Datalog rules.
- Data Handling Policy (DHP) provides data usage conditions matching between data owners’ privacy preferences and users’ usage context.
- Data Releasing Policy (DRP) describes what are available SDC methods with de-identifiable PII are disclosed for analysis but data privacy is preserved.
Automated Big Data Analysis Pipeline [32]
Semantics of a super-peer data cloud is described as the policy ontology, including modular concepts of SPD.

Semantics-enabled policies perform data integration within an SPD.

Semantics-enabled policies are unified to fulfill data exchange across SPDs.
Semantics of a super-peer data cloud is described as the policy ontology, including modular concepts of SPD.

Semantics-enabled policies perform data integration within an SPD.

Semantics-enabled policies are unified to fulfill data exchange across SPDs.
Semantics of Super-Peer Domain (SPD) Cloud

1. Semantics of a super-peer data cloud is described as the policy ontology, including modular concepts of SPD.
2. Semantics-enabled policies perform data integration within an SPD.
3. Semantics-enabled policies are unified to fulfill data exchange across SPDs
Policy Ontology for Super-Peer Domain Cloud
DEFINITION OF ACP ONTOLOGY

ACP describes the concept of data usage access control in the super-peer of an SPD.
**SPECIFICATION OF ACP RULE**

\[
\text{Request}(\forall r) \land \text{hasCondition}(\forall r, \forall c) \land \text{Condition}(\forall c) \\
\land \text{hasCondition}(\forall \text{avp}, \forall \text{ac}) \land \text{Condition}(\forall \text{ac}) \\
\land \text{AccessVerifyPolicy}(\forall \text{avp}) \land \text{sameAs}(\forall \text{ac}, \forall c) \\
\land \text{empower}(\forall \text{avp}, \forall \text{qt}) \land \text{QueryType}(\forall \text{qt}) \\
\rightarrow \text{isEmpowered}(\forall r, 1) \land \text{hasQueryType}(\forall r, \forall \text{qt}) \quad (1)
\]
Ontology for Data Handling Policy (DHP)

**Definition of DHP Ontology**

DHP describes the concept of semantic metadata markup services and decides which data owners’ privacy preferences match which data users’ usage context.
Rule for Data Handling Policy (DHP)

**Specification of DHP Rule**

Request(?r) ∧ isEmpowered(?r, 1) ∧ hasCondition(?r, ?c)
∧ Condition(?c) ∧ DataPolicy(?dp) ∧ Condition(?dc)
∧ hasCondition(?dp, ?dc) ∧ sameAs(?c, ?dc) ∧ hasSQL(?dp, ?s)
→ sqwrl:select(?s) ← (2)
Definition of DRP Ontology

DRP describes the concept for which part of PII attributes are allowed to disclose for analysis and still ensures the privacy principles.
Ontology for Data Releasing Policy (DRP)(Conti.)

**Definition of DRP Ontology**

- hasData.Request(), hasData⁻.Data().
- hasQueryType.Request(), hasQueryType⁻.(QueryType(PBQs)).
- hasPartOf.Data(), hasPartOf⁻.ID(), hasPartOf⁻.Name(), ...
- hasPartOf⁻.ZIP(), hasPartOf⁻.Cholesterol().
- hasSubClassOf.DataAttribute(),
- hasSubClassOf⁻. Identifiers(),
- hasSubClassOf⁻. Quasi⁻identifiers(),
- hasSubClassOf⁻. Confidential(),
- hasPartOf. Identifiers(), hasPartOf⁻. ID(id.), ...
- hasPartOf. Confidential(), hasPartOf⁻. Disease().
Ontology for Data Releasing Policy (DRP)(Conti.)

**Definition of DRP Ontology**

- hasSubClassOf.DataType(),
- hasSubClassOf.Categorical(),
- hasSubClassOf.Continuous().
- hasContinuous.Cholesterol(), hasContinuous.Continuous().
- hasCategorical.ID(), hasCategorical.Categorical().
  ...
- hasCategorical.Doctor(), hasCategorical.Categorical().
- canApply.SDC(generalization), canApply.Categorical().
  ...
- canApply.SDC(top-coding), canApply.Continuous().
## Specification of DHP Rules

Request(\(?r\)) \& hasData(\(?r, ?d\)) \& Data(\(?d\))
\& hasPartOf(\(?d, ?pod\)) \& hasQueryType(\(?r, PBQ\))
\& sqwrl: makeSet(\(?rs, ?pod\)) \& sqwrl: groupBy(\(?rs, ?r\))
\& Quasi-identifiers(\(?qui\)) \& hasPartOf(\(?qui, ?qpod\))
\& sqwrl: groupBy(\(?qs, ?qui\)) \& sqwrl: contains(\(?rs, ?qs\))
\& Confidential(\(?c\)) \& hasPartOf(\(?c, ?dc\))
\rightarrow sqwrl: selectDistinct(\(?qui, ?gpod\)) ← (3)
Rules for Data Handling Policy (DHP)(Conti.)

**Specification of DHP Rules**

Request(?r) ∧ hasData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?b) ∧ selected(?r, ?b) ∧ hasContinuous(?b, ?con) ∧ Continuous(?con) ∧ SDC(?sdc) ∧ canApply(?sdc, ?con) → sqwrl: select(?b, ?sdc) ← (4)

**Specification of DHP Rules**

Request(?r) ∧ hasData(?r, ?d) ∧ Data(?d) ∧ hasPartOf(?d, ?b) ∧ selected(?r, ?b) ∧ hasCategorical(?b, ?cat) ∧ Categorical(?con) ∧ SDC(?sdc) ∧ canApply(?sdc, ?cat) → sqwrl: select(?b, ?sdc) ← (5)
Specification of DHP Rules

Request(\(r\)) \land \text{hasData}(\(r, d\)) \land \text{Data}(\(d\))
\land \text{hasPartOf}(\(d, b\)) \land \text{selected}(\(r, b\))
\land \text{hasContinuous}(\(b, con\)) \land \text{Continuous}(\(con\))
\land \text{SDC}(\(sdc\)) \land \text{canApply}(\(sdc, con\))
\rightarrow \text{sqwr1} : \text{select}(\(b, sdc\)) ← (4)

Specification of DHP Rules

Request(\(r\)) \land \text{hasData}(\(r, d\)) \land \text{Data}(\(d\))
\land \text{hasPartOf}(\(d, b\)) \land \text{selected}(\(r, b\))
\land \text{hasCategorical}(\(b, cat\)) \land \text{Categorical}(\(con\))
\land \text{SDC}(\(sdc\)) \land \text{canApply}(\(sdc, cat\))
\rightarrow \text{sqwr1} : \text{select}(\(b, sdc\)) ← (5)
Rules for Data Handling Policy (DHP)(Conti.)

**Specification of DHP Rules**

\[
\text{Request}(?r) \land \text{hasData}(?r, ?d) \land \text{Data}(?d) \\
\land \text{hasPartOf}(?d, ?b) \land \text{select}(?r, ?b) \land \text{isHandled}(?b, 1) \\
\land \text{hasPartOf}(?d, ?a) \land \text{notSelected}(?r, ?a) \\
\rightarrow \text{canUse}(?r, ?a) \land \text{canUse}(?r, ?b) \leftarrow (6)
\]
Semantic Data Analysis and Protection

- Improve the situation, where SDC enforcement is obliged to original data providers and a data analytics user lacks the flexibility to choose suitable SDC methods.
- Seek a balance between a data owner’s right for privacy protection and a data user’s need for data analytics through transparency of SDC methods releasing.
- Semantics-enabled Data Releasing Policy (DRP) calls for which SDC methods and ensures maximum data utility with a tolerable data disclosure risk.

![Data Utility vs Disclosure Risk Diagram]

- Original Microdata (SBQ)
- Risk Tolerable Line
- Released and Protected Microdata (PBQ)
- No-Released Microdata (Deny Access)
Semantic Data Analysis and Protection

- Improve the situation, where SDC enforcement is obliged to original data providers and a data analytics user lacks the flexibility to choose suitable SDC methods.
- Seek a balance between a data owner’s right for privacy protection and a data user’s need for data analytics through transparency of SDC methods releasing.
- Semantics-enabled Data Releasing Policy (DRP) calls for which SDC methods and ensures maximum data utility with a tolerable data disclosure risk.

![Diagram showing the relationship between Data Utility and Disclosure Risk]
Semantic Data Analysis and Protection

- Improve the situation, where SDC enforcement is obliged to original data providers and a data analytics user lacks the flexibility to choose suitable SDC methods.
- Seek a balance between a data owner’s right for privacy protection and a data user’s need for data analytics through transparency of SDC methods releasing.
- Semantics-enabled Data Releasing Policy (DRP) calls for which SDC methods and ensures maximum data utility with a tolerable data disclosure risk.
A Three-Tier SDC Prototyping System
Major Papers Cited

- Privacy protection for big data:
  [10] [37] [45]
- Statistical Disclosure Control (SDC):
  [1] [12] [16] [29]
- Privacy-aware access control policy:
  [2] [5] [6] [31] [46] [47]
Conclusion and Future Works

Preliminary Results

1. Semantics-enabled policies, ACP, DHP, and DRP, are proposed and verified through query restriction, manipulation, and output perturbation, which can ensure the privacy protection principles.

2. Supporting a simple but not yet optimal balance between data utility and protection through policies call for SDC methods.

Future Work

1. Establish a distributed R + Hadoop/MapReduce environment to provide big data deep analysis without violating personal privacy.

2. Design and implement an automated big data analysis pipeline system through Semantic Web Services.

3. The ultimate goal is to craft an optimize balance between data utility and protection in the automated big data analysis life cycle.
## Preliminary Results

1. Semantics-enabled policies, ACP, DHP, and DRP, are proposed and verified through query restriction, manipulation, and output perturbation, which can ensure the privacy protection principles.

2. Supporting a simple but not yet optimal balance between data utility and protection through policies call for SDC methods.

## Future Work

1. Establish a distributed R + Hadoop/MapReduce environment to provide big data deep analysis without violating personal privacy.

2. Design and implement an automated big data analysis pipeline system through Semantic Web Services.

3. The ultimate goal is to craft an optimize balance between data utility and protection in the automated big data analysis life cycle.
Conclusion and Future Works

Preliminary Results

1. Semantics-enabled policies, ACP, DHP, and DRP, are proposed and verified through query restriction, manipulation, and output perturbation, which can ensure the privacy protection principles.

2. Supporting a simple but not yet optimal balance between data utility and protection through policies call for SDC methods.

Future Work

1. Establish a distributed R + Hadoop/MapReduce environment to provide big data deep analysis without violating personal privacy.

2. Design and implement an automated big data analysis pipeline system through Semantic Web Services.

3. The ultimate goal is to craft an optimize balance between data utility and protection in the automated big data analysis life cycle.
Conclusion and Future Works

**Preliminary Results**

1. Semantics-enabled policies, ACP, DHP, and DRP, are proposed and verified through query restriction, manipulation, and output perturbation, which can ensure the privacy protection principles.

2. Supporting a simple but not yet optimal balance between data utility and protection through policies call for SDC methods.

**Future Work**

1. Establish a distributed R + Hadoop/MapReduce environment to provide big data deep analysis without violating personal privacy.

2. Design and implement an automated big data analysis pipeline system through Semantic Web Services.

3. The ultimate goal is to craft an optimize balance between data utility and protection in the automated big data analysis life cycle.
Conclusion and Future Works

Preliminary Results

1. Semantics-enabled policies, ACP, DHP, and DRP, are proposed and verified through query restriction, manipulation, and output perturbation, which can ensure the privacy protection principles.

2. Supporting a simple but not yet optimal balance between data utility and protection through policies call for SDC methods.

Future Work

1. Establish a distributed R + Hadoop/MapReduce environment to provide big data deep analysis without violating personal privacy.

2. Design and implement an automated big data analysis pipeline system through Semantic Web Services.

3. The ultimate goal is to craft an optimize balance between data utility and protection in the automated big data analysis life cycle.
System Demo and Q&A

(Loading wims13demo.mp4)
R. N. Adam and C. J. Worthmann.

A. C. Ardagna et al.
A privacy-aware access control system.

A. P. Bernstein and L. M. Haas.
Information integration in the enterprise.

M. Bezzi et al.
Modeling and preventing inferences from sensitive value distribution in data release.

A. P. Bonatti.
Datalog for security, privacy and trust.

S. Cabuk et al.
Towards automated security policy enforcement in multi-tenant virtual data centers.

D. Calvanese et al.
Logical foundations of peer-to-peer data integration.
D. Calvanese et al.
Data management in peer-to-peer data integration systems.

D. Calvanese and G. D. Giacomo.
Data integration: A logic-based perspective.

A. Cavoukian and J. Jonas.
Privacy by design in the age of big data, 2012.

S. Ceri et al.
What you always wanted to know about Datalog (and never dared to ask).

V. Ciriani et al.
Microdata protection.

C. Cliffton et al.
Privacy-preserving data integration and sharing.

Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act.
M. Cox and D. Ellsworth.
Application-controlled demand paging for out-of-core visualization.

J. Domingo-Ferrer et al.
Risk-utility paradigms for statistical disclosure limitation: How to think, but not how to act - discussion: A science of statistical disclosure limitation?

C. Dwork.
Differential privacy.

C. Dwork.
A firm foundation for private data analysis.

A. Eberhart et al.
Semantic technologies and cloud computing.

T. Eiter et al.
*Rules and Ontologies for the Semantic Web*.

R. Faigin et al.
Data exchange: Semantics and query answering.

S. Foresti.
*Preserving Privacy in Data Outsourcing.*

P. Haase et al.
Semantic technologies for enterprise cloud management.

Y. A. Halevy.
Answering queries using views: A survey.

Y. J. Hu et al.
Semantic legal policies for data exchange and protection across super-peer domains in the cloud.

Law-aware semantic cloud policies with exceptions for data integration and protection.
In *International Conference on Web Intelligence, Mining and Semantics (WIMS’12)*. ACM Press, June 2012.

Semantics-enabled policies for information sharing and protection in the cloud.
In *Proc. of 3rd Int. Conf. on Social Semantics*, LNCS 6984, Oct. 2011.
A semantic privacy-preserving model for data sharing and integration.

A. Hundepool et al.
*Statistical Disclosure Control*.

A. Inam et al.
A hybrid approach to private record linkage.
In *24th International Conference on Data Engineering (ICDE)*, pages 496–505. IEEE, 2008.

G. Karjoth and M. Schunter.
A privacy policy model for enterprises.

A. Labrinidis et al.
Challenges and opportunities with big data.

M. Lenzerini.
Data integration: A theoretical perspective.

J. Madhavan et al.
Web-scale data integration: You can only afford to pay as you go.

J. Manyika et al.
Big data the next frontier for innovation, competition, and productivity.

D. Martin et al.
OWL-S: Semantic markup for web service.

C. A. Mora et al.
Top ten big data security and privacy challenges.

M. Morgenstern.
Security and inference in multilevel database and knowledge-base systems.

A. Nash and A. Deutsch.
Privacy in GLAV information integration.

J. M. O’Connor and A. K. Das.
SQWRL: a query language for OWL.

R. Popp and J. Poindexter.
Countering terrorism through information and privacy protection technologies.

K. Schwab et al.
Personal data: The emergence of a new asset class.

F. J. Sequeda et al.
Survey of directly mapping SQL databases to the semantic web.

L. Sweeney.
K-annonumity: a model for protecting privacy.

O. Tene and J. Polonetsky.
Privacy in the age of big data: A time for big decisions.

S. D. C. d. Vimercati et al.
Access control policies and languages in open environments.

J. D. Weitzner et al.
Creating a policy-aware web: Discretionary, rule-based access for the world wide web.