Tag-based Information Flow Analysis for Document Classification in Provenance

Jyothsna Rachapalli, Murat Kantarcioglu and Bhavani Thuraisingham

The University of Texas at Dallas

TaPP 2012 : 4th USENIX Workshop on the Theory and Practice of Provenance
Boston, MA, Jun 14, 2012 - Jun 15, 2012

Introduction

- **We propose a tagging mechanism :**
  - As a stepping stone towards our vision of secure provenance
  - To track the flow of sensitive information in a provenance graph
  - To automate the process of document classification

Motivation

- A crucial aspect of Intelligence/Health-care domain is :
  - manage and protect sensitive information effectively and efficiently
- Confidential data leaks are one of the worst kinds of leaks :
  - can cause serious damage to organizations
- Therefore, it becomes imperative to detect and manage such leaks efficiently

Open Provenance Model (OPM)

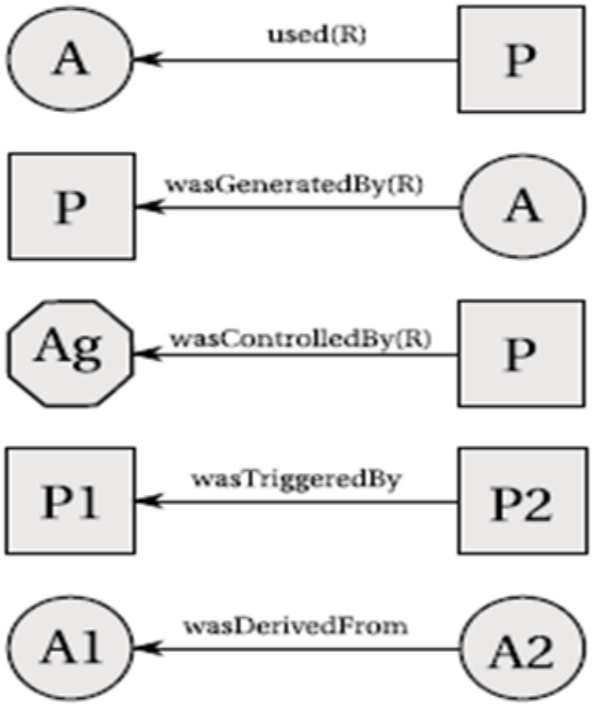
- Tagging mechanism comprises of inference rules based on OPM
- Open Provenance Model
  - is a general model of provenance
  - defines provenance in a precise, technology-agnostic manner
  - designed to allow provenance information to be exchanged between systems facilitating interoperability, by means of a shared provenance model
  - models provenance as a directed acyclic graph that captures causal relationships
- We build our tagging mechanism based on OPM,
  - to make it independent of any specific domain or technology such as databases, workflows or distributed systems

Open Provenance Model

- OPM graph consists nodes and dependencies
  - Artifacts
  - Agents
  - Processes
- The dataflow oriented view comprises of artifacts and "was derived from" edges connecting them.
- We propose a property called tag
  - to annotate an artifact

subject:	an artifact
property:	http://openprovenance.org/property#tag
value:	an Integer
meaning:	Represents Artifact rating or priority

Causal Dependencies



Tag-based Mechanism

- We state two inference rules based on dataflow oriented view
  - In Single Source Derivation, an artifact is known to be derived from only one other artifact
  - In Multi-Source Derivation, an artifact is known to be derived from more than one artifact
- We then propose an implementation scheme using
  - Web Ontology Language (OWL)
  - Semantic Web Rule Language (SWRL).
  - Query Language -SPARQL

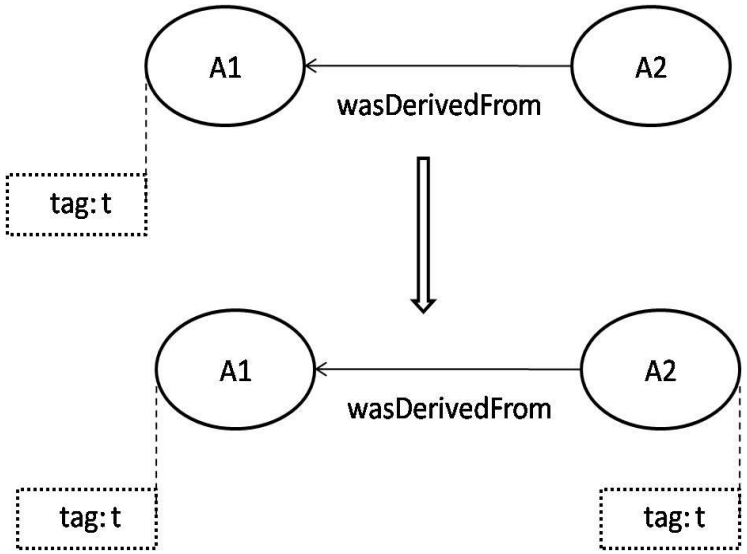
Tag Propagation Rules

Single Source Derivation:

- If Artifact A1 is annotated with tag "t" and artifact A2 was derived from A1
  - then artifact A2 is annotated with tag t as well

- Artifact(?a1)^Artifact(?a2)^tag(?a1,?t) ^wasDerivedFrom(?a2, ?a1) --> tag(?a2, ?t)

Tag Propagation



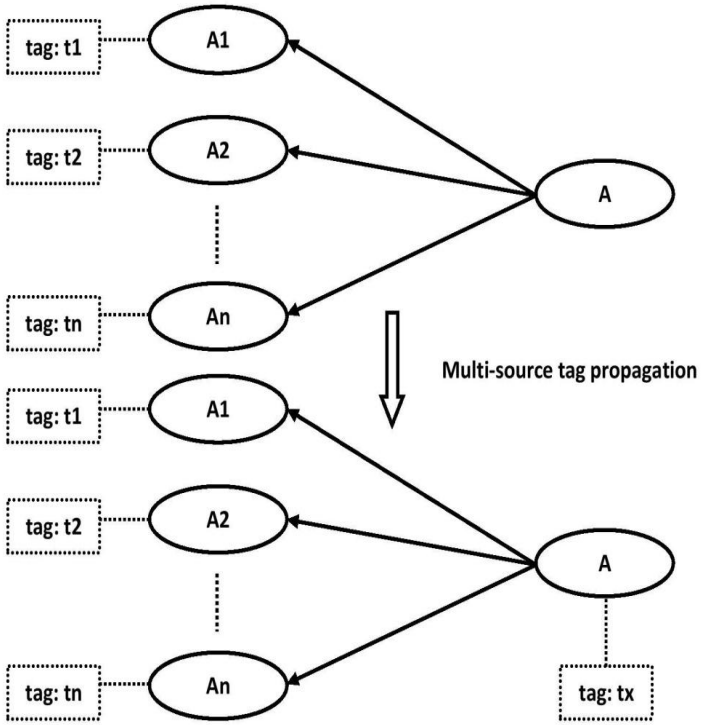
Tag Propagation Rules

Multi-Source derivation

- If artifact A was derived from artifacts A1, A2, A3,..., An, annotated with tags "t1", "t2", "t3", ..., "tn",
  - then artifact A is annotated with tag **tx**, (highest priority tag)

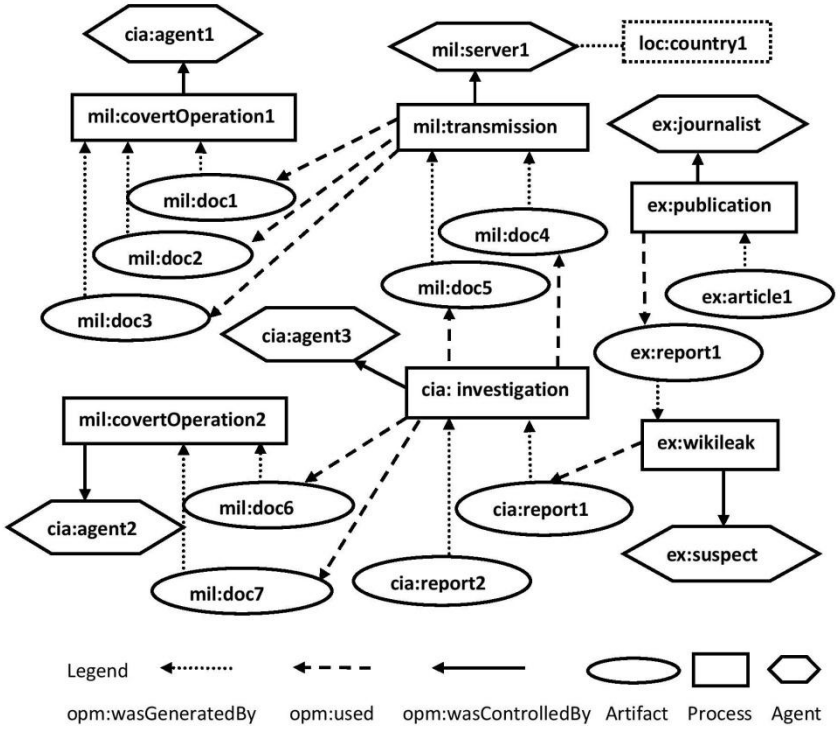
- Artifact(?a1)^Artifact(?a2)^tag(?a1,?t1)^ tag(?a2,?t2)^wasDerivedFrom(?a2,?a1)
  ^swrlb:greaterThan(?t1, ?t2)
  --> tag(?a2, ?t1)

Tag propagation



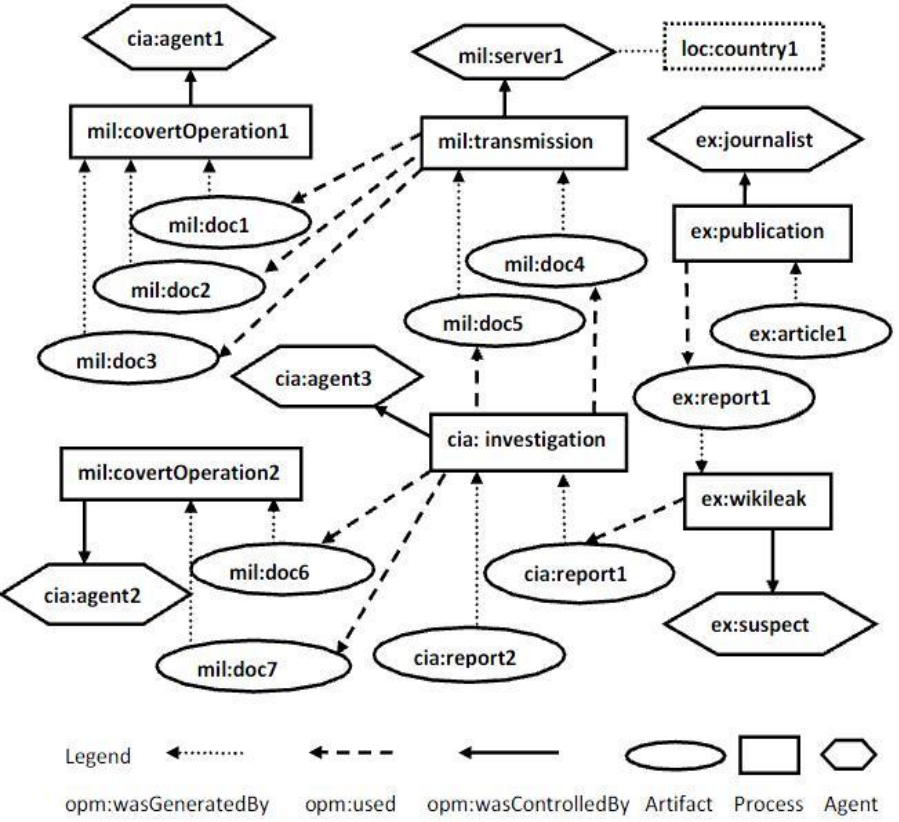
Intelligence Domain Usecase

Classification	Tag Value
Top-secret	4
Secret	3
Confidential	2
Unclassified	1

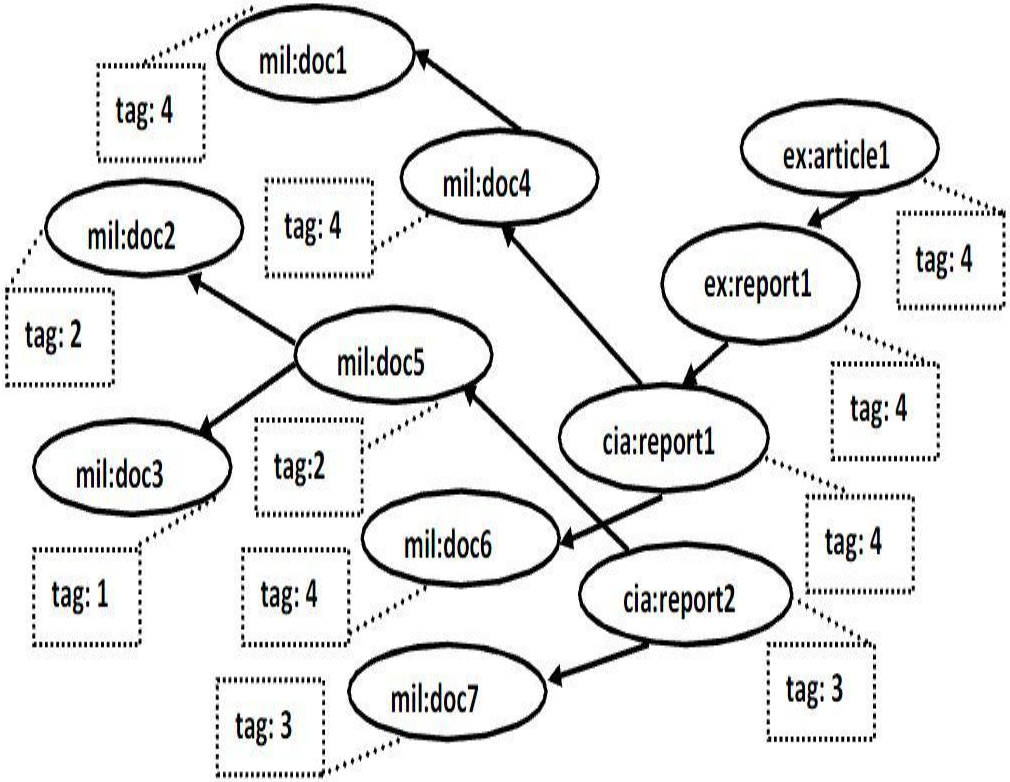


Tag Propagation – Intelligence Use-case

Intelligence Use-case



Data-flow view



Sample SPARQL Queries

Description	SPARQL Queries
• List all the topsecret documents	SELECT ?x WHERE {?x opm:tag "4" }
• List all the topsecret documents generated within time interval t1 and t2.	SELECT ?x WHERE {?x opm:tag "4" . ?x opm:time ?t . FILTER {(?t >= t1) && (?t <= t2) }}
• What is the classification level of mil:doc1	SELECT ?y WHERE{mil:doc1 opm:tag ?y}
• List all the documents from which cia:report1 was derived along with their classification levels	SELECT ?y ?z WHERE{cia:report1 opm:wasDerivedFrom ?y . ?y opm:tag ?z.}

Conclusion

- We proposed a tag-based mechanism to track the flow of sensitive/valuable information
  - default tag propagation rules for OPM; can be overridden with custom rules if needed
- It can help with decisions such as:
  - Providing Access Control
  - Sanitization
  - Scoping provenance : Recording fine-grained provenance information for data labeled with high priority tags and recording minimum required provenance information for data labeled with low priority tags