EPSRC Engineering and Physical Sciences Research Council

ORCHID

# Provenance Analytics

**Amir Sezavar Keshavarz**

Web and Internet Science Research Group
Electronics and Computer Science
University of Southampton

## Overview

Goal:

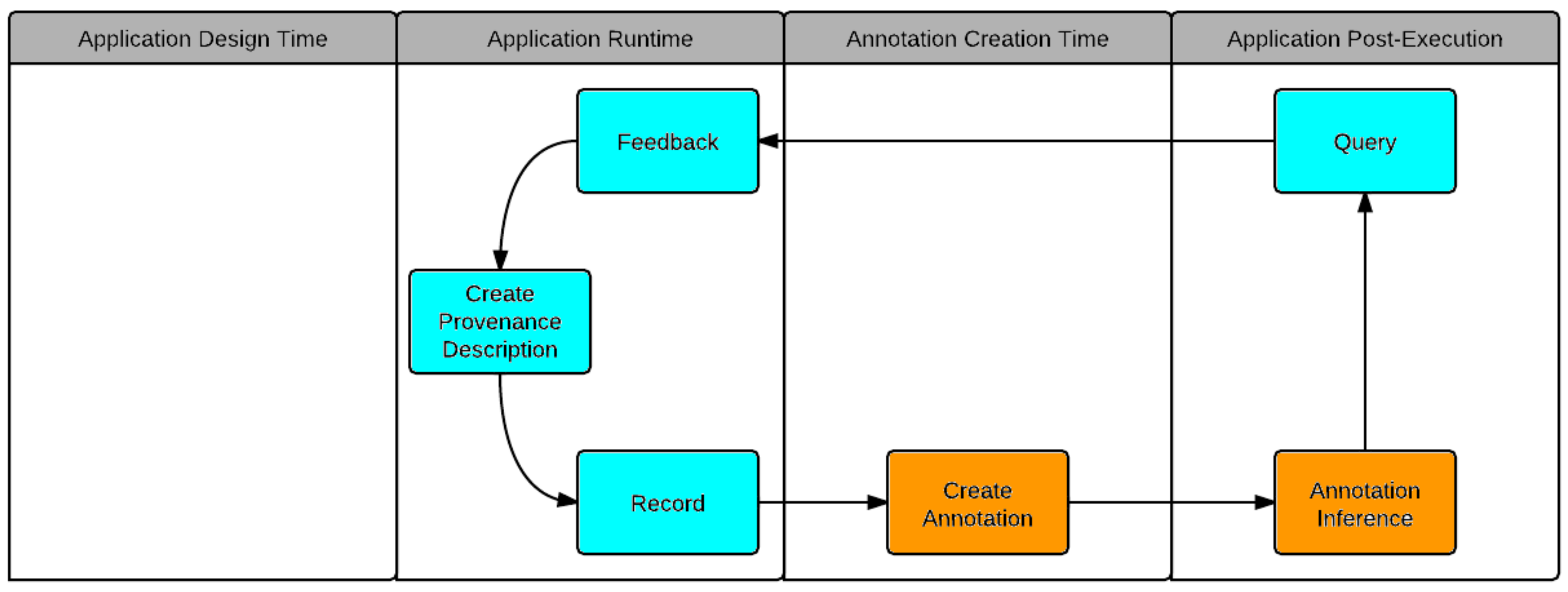
- **To introduce a mechanism for domain specific interpretation of provenance of data**
- **Provenance of data is generated in "Application Runtime" stage based on requirements pre-defined in "Application Design Time" stage. Annotation (data about data) can be created in "Annotation Creation Time" stage and new annotations can be inferred in "Application Post-Execution" stage. New annotations can be inferred based on existing annotations, provenance of data, and other external data that has not been generated in earlier stages.**

Provenance analytics is a solution consisting of:

- **Annotation level: Annotation is utilised as a generic mechanism to enable users to attach any information to the elements of a provenance graph.**
- **Inference level: New annotations are inferred based on existing annotations and information that the provenance graph provides**
- **Annotation propagation framework and provenance graph traversal**

Instantiations of the framework: trust and error to propagate and infer trust and error values

## Annotation Life Cycle



Application Design Time | Application Runtime | Annotation Creation Time | Application Post-Execution

Feedback

Query

Create Provenance Description

Record

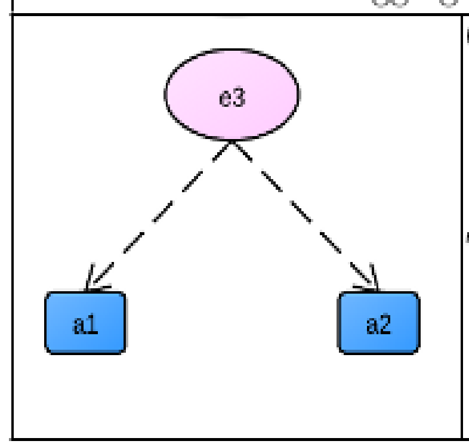
Create Annotation

Annotation Inference

Annotation life cycle shows different stages involving in creation and inference of new annotations

## Annotation Inference Framework

Aggregation-Rule



**Given that**

$hasAnnotation(e3, a1)$

$hasAnnotation(e3, a2)$

$\ldots$

**Then**

$hasAnnotation(e3, a3)$

and

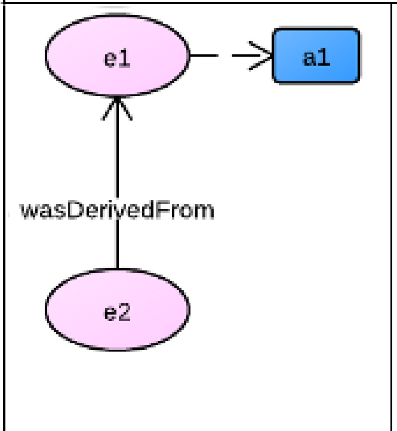
$a3 = F_{Aggregation}(G, e3, [a1, a2, ...])$

- Annotation inference is a form of inference, that given a provenance graph with some annotations, infers new annotations for the same graph.
- These inferences are defined by a set of rules:

***1.Relation rules*** are about propagating annotations over relations.

***2.Aggregation and transformation rules*** aggregate all new annotations into one or transform the type of annotation.

Derivation-Forward-Propagation-Rule



**IF**

$G \vdash$

$entity(e1, attrs1)$

$entity(e2, attrs2)$

$wasDerivedFrom(d1, e2, e1, attrs3)$

$hasAnnotation(e1, a1)$

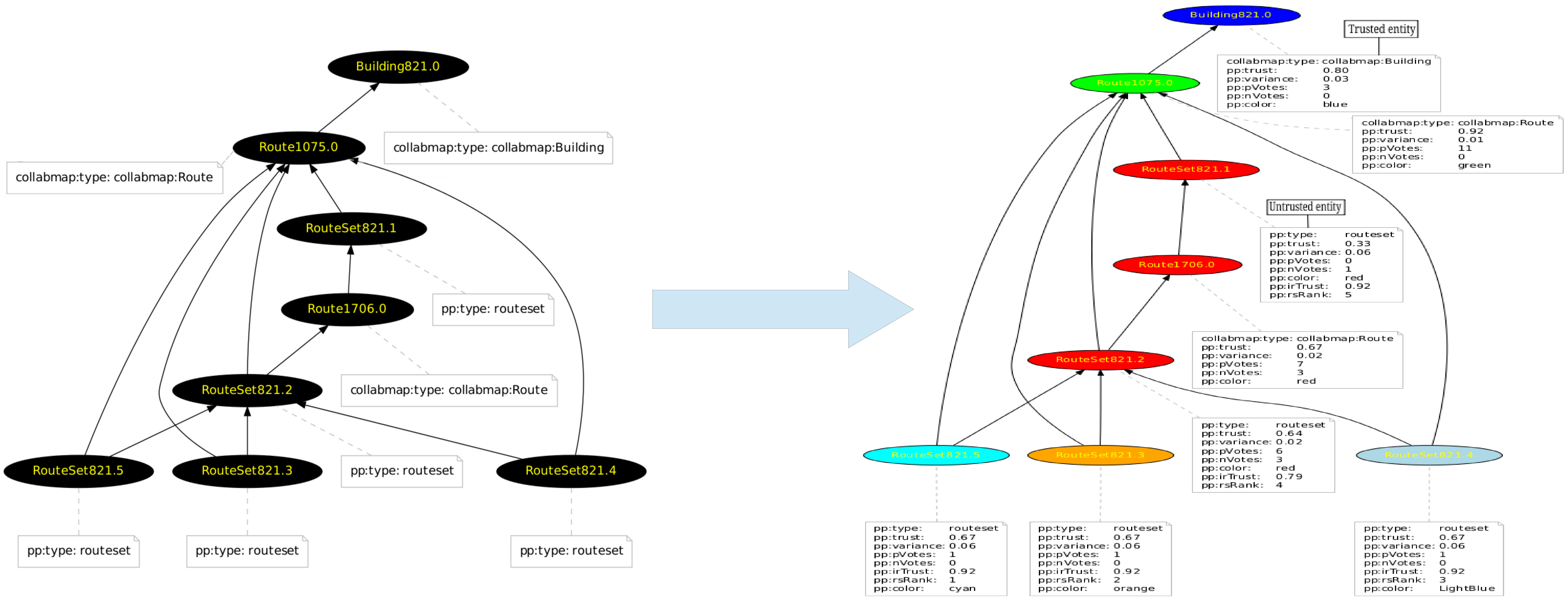
**THEN** there exists a2 such that

$hasAnnotation(e2, a2)$

and

$a2 = F_{Derivation}(G, d1, e1, e2, a1)$

## CollabMap Provenance Graph – Before and After Provenance Analytics



```
collabmap:type: collabmap:Building
pp:trust:       0.80
pp:variance:    0.03
pp:pVotes:      3
pp:nVotes:      0
pp:color:       blue
```

```
collabmap:type: collabmap:Route
pp:trust:       0.92
pp:variance:    0.01
pp:pVotes:      11
pp:nVotes:      0
pp:color:       green
```

```
pp:type:     routeset
pp:trust:    0.33
pp:variance: 0.06
pp:pVotes:   0
pp:nVotes:   1
pp:color:    red
pp:irTrust:  0.92
pp:rsRank:   5
```

```
collabmap:type: collabmap:Route
pp:trust:       0.67
pp:variance:    0.02
pp:pVotes:      7
pp:nVotes:      3
pp:color:       red
```

```
pp:type:     routeset
pp:trust:    0.64
pp:variance: 0.02
pp:pVotes:   6
pp:nVotes:   3
pp:color:    red
pp:irTrust:  0.79
pp:rsRank:   4
```

```
pp:type:     routeset
pp:trust:    0.67
pp:variance: 0.06
pp:pVotes:   1
pp:nVotes:   0
pp:irTrust:  0.92
pp:rsRank:   1
pp:color:    cyan
```

```
pp:type:     routeset
pp:trust:    0.67
pp:variance: 0.06
pp:pVotes:   1
pp:nVotes:   0
pp:irTrust:  0.92
pp:rsRank:   2
pp:color:    orange
```

```
pp:type:     routeset
pp:trust:    0.67
pp:variance: 0.06
pp:pVotes:   1
pp:nVotes:   0
pp:irTrust:  0.92
pp:rsRank:   3
pp:color:    LightBlue
```

## Results

CollabMap application

- **CollabMap is a crowdsourcing application to get users to augment existing maps, provided by Google Maps and panoramic views from Google Street Views, by drawing evacuation routes.**
- **Over 5,000 provenance graphs, around 9,700 nodes, and 220,000 relations**

Applying the propagation framework and trust instantiation on CollabMap data to compute

- **Trust value for buildings, routes, and route sets**
- **The total number of positive and negative votes of each user for buildings, routes, and route sets**
- **Another notion of trust for each route set based on its included routes**

## Future Work

- Propagate annotations over following relations to be more compliant with W3C PROV specification
  - **Delegation, Communication, Bundle**
- Privacy instantiation to be applied for agentSwitch application
  - **To propagate and infer new privacy labels for derived information from existing information which have privacy labels**
  - **Application in auditing to identify any leakage of private information, in online or pseudo-online applications to enforce privacy policy**
- Evaluation of the framework
  - **To demonstrate the framework can be efficiently instantiated**
    - **Assess performance (time) and scalability**
    - **Scalability is defined as the ability of the framework to handle and accommodate large provenance graph and many large provenance graphs**
  - **To demonstrate the framework is useful for being instantiated**
    - **It is useful if it is possible to develop different instantiations based on it**

UNIVERSITY OF Southampton