Temporal Semantics for the Open Provenance Model

Jan Van den Bussche
Hasselt University, Belgium

joint work with Natalia Kwasnikowska (Hasselt)
and Luc Moreau (Southampton)

prov•e•nance |ˈprävənəns|

noun

the place of origin or earliest known history of something *:*

an orange rug of Iranian provenance.

• the beginning of something's existence; something's origin *: they try to understand the whole universe, its provenance and fate.*

See note at **ORIGIN** .

• a record of ownership of a work of art or an antique, used as a guide to authenticity or quality *: the manuscript has a distinguished provenance.*

ORIGIN late 18th cent.: from French, from the verb ***provenir 'come or stem from,'*** from Latin ***provenire***, from ***pro- 'forth' + venire 'come.'***

Provenance in computing

"Data provenance:" *Where does this piece of data come from?*

"Workflow provenance, Process provenance:" *What happened?*

scientific databases, computational science, operating systems, debugging, workflow management

- need for a common data model for provenance information

$\Rightarrow$ Open Provenance Model

The Open Provenance Model (OPM)

Consensus data model

Scientific computing community

OPM v1.1 specification published July 2010
[Luc Moreau et al., *Future Generation Computer Systems*]

W3C Provenance Working Group started 2011

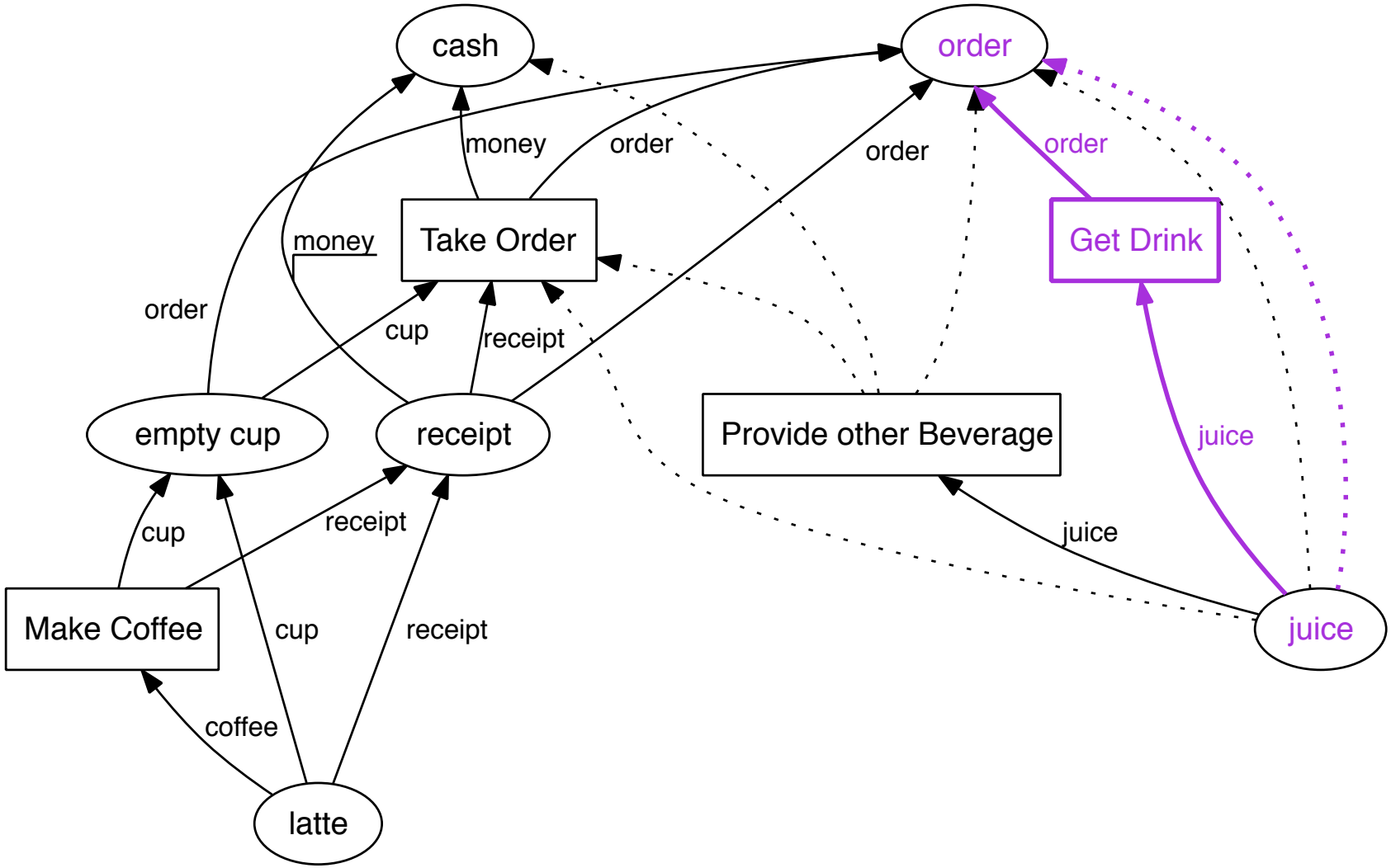
OPM graph

Directed graph

Two kinds of nodes: **processes** and **artifacts**

Four kinds of edges:

$P \xrightarrow{r} A$	"P used A in role r" *meaning: P could not have completed without A*
$A \xrightarrow{r} P$	"A was generated by P in role r" *meaning: A could not have existed without P*
$P_1 \rightarrow P_2$	"P_1 was informed by P_2" *meaning: P_1 could not have completed without P_2*
$A_1 \rightarrow A_2$	"A_1 was derived from A_2" *meaning: A_1 could not have existed without A_2*


cash
order
money
order
order
order
money
Take Order
Get Drink
order
cup
receipt
empty cup
receipt
Provide other Beverage
juice
cup
receipt
juice
Make Coffee
cup
receipt
juice
coffee
latte

Inference rules for "multi-step" edges

0. **if** $X \to Y$ or $X \xrightarrow{r} Y$
   then $X \xrightarrow{*} Y$

1. **if** $A \xrightarrow{*} B \xrightarrow{*} C$
   then $A \xrightarrow{*} C$

2. **if** $A \xrightarrow{*} B \xrightarrow{*} P$
   then $A \xrightarrow{*} P$

3. **if** $P \xrightarrow{*} A \xrightarrow{*} B$
   then $P \xrightarrow{*} B$

4. **if** $P_1 \xrightarrow{*} A \xrightarrow{*} P_2$
   then $P_1 \xrightarrow{*} P_2$

A critique on the OPM spec

Only syntax, no (formal) semantics

Inference rules just a syntactic edge-adding game; in what sense are they sound? Are they complete?

Multi-step edges cannot be asserted in the OPM graph; lack of support for levels of granularity

Difference in meaning between single-step and multi-step edges?

What is correct reasoning?

There is a rule:
if $P_1 \xrightarrow{*} A \xrightarrow{*} P_2$
then $P_1 \xrightarrow{*} P_2$

But there is no rule:
if $A_1 \xrightarrow{*} P \xrightarrow{*} A_2$
then $A_1 \xrightarrow{*} A_2$

Need for a formal semantics

Our work

Define an improved version of the OPM data model

Provide a temporal semantics

Investigate soundness, completeness, of inference rules

OPM graphs, take 2

Directed graph, two kinds of nodes (processes and artifacts)

Seven kinds of edges:

kind	precise	imprecise
generated-by	$A \xrightarrow{r} P$	$A \rightarrow P$
used	$P \xrightarrow{r} A$	$P \rightarrow A$
derived-from	$A \xrightarrow{r} B$	$A \rightarrow B$
informed-by	—	$P_1 \rightarrow P_2$

Temporal semantics

Set $Vars$ of **temporal variables**:

- create(A) for each artifact A
- begin(P) and end(P) for each process P
- use(P, r, A) for each $P \xrightarrow{r} A$

A **temporal interpretation** is a mapping

$$\tau : Vars \to \mathbb{N}$$

assigning timepoints to the temporal variables

Temporal theory of the OPM graph

Ax.1: begin(P) $\leq$ end(P) for each P

Ax.2: begin(P) $\leq$ create(A) $\leq$ end(P) for each $A \xrightarrow{r} P$

Ax.3: begin(P) $\leq$ use(P, r, A) $\leq$ end(P) and create(A) $\leq$ use(P, r, A) for each $P \xrightarrow{r} A$

Ax.4: create(B) $\leq$ create(A) for each $A \rightarrow B$

Ax.5: begin(P) $\leq$ create(A) for each $A \rightarrow P$

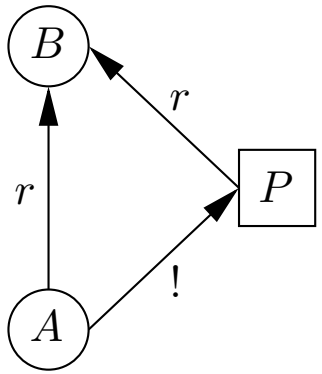
Ax.6: create(A) $\leq$ end(P) for each $P \rightarrow A$

Ax.7: begin(Q) $\leq$ end(P) for each $P \rightarrow Q$

Axiom 8

Ax.8: use$(P, r, B) \leq$ create(A) for each $\underline{\triangle(A, B, P, r)}$

"Generate–use–derive triangle"



$A \xrightarrow{!} P$ is an abbreviation for $\exists s : A \xrightarrow{s} P$

Temporal models

Any temporal interpretation that satisfies Axioms 1–8 is called a **temporal model** of the OPM graph

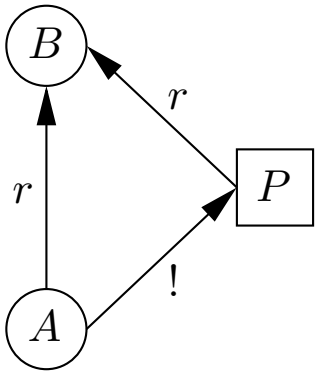
E.g.: $A \rightarrow P \xrightarrow{r} B$

interpretation	τ_1	τ_2	τ_3	τ_4
create(B)	1	1	3	2
begin(P)	2	2	1	3
use(P, r, B)	3	4	4	4
create(A)	4	3	2	1
end(P)	5	5	5	5
model?	yes	yes	yes	no

Temporal inference

Given: An OPM graph G

Find: All inequalities that logically follow from G

E.g.: 

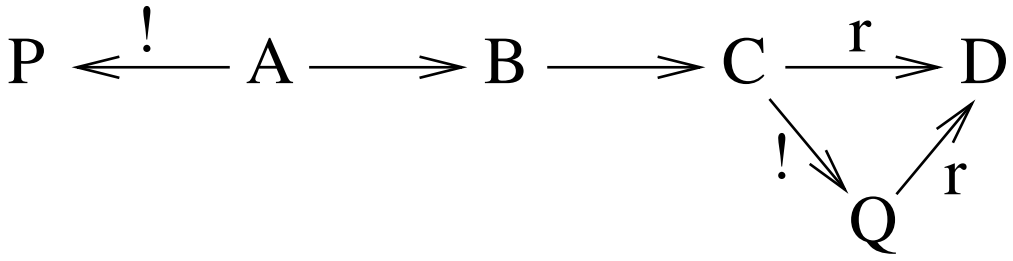
logically implies create(B) $\leq$ create(A)

E.g.: $A \rightarrow P \rightarrow B$ does **not** imply create(B) $\leq$ create(A)

Reasoning with inequalities

Example:

the following OPM graph implies use$(Q, r, D) \leq$ end(P)



use$(Q, r, D) \overset{\text{Ax.8}}{\leq}$ create$(C) \overset{\text{Ax.4}}{\leq}$ create$(B) \overset{\text{Ax.4}}{\leq}$ create$(A) \overset{\text{Ax.2}}{\leq}$ end(P)

- Would be better to do inference in the graph itself

Revenge of the OPM edge inference rules

0. **if** $X \rightarrow Y$ or $X \stackrel{!}{\rightarrow} Y$
   then $X \dashrightarrow Y$

1. **if** $A \dashrightarrow B \dashrightarrow C$
   then $A \dashrightarrow C$

2. **if** $A \dashrightarrow B \dashrightarrow P$
   then $A \dashrightarrow P$

3. **if** $P \dashrightarrow A \dashrightarrow B$ or $P \stackrel{!}{\leftarrow} A \dashrightarrow B$
   then $P \dashrightarrow B$

4. **if** $P \dashrightarrow A \dashrightarrow Q$ or $P \stackrel{!}{\leftarrow} A \dashrightarrow Q$
   then $P \dashrightarrow Q$

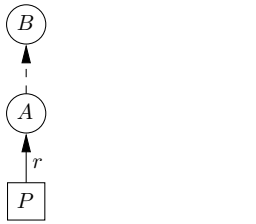
Completeness of the edge inference rules

Theorem:

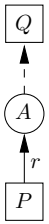
1. create(B) $\leq$ create(A) logically follows iff A --↠ B
2. begin(P) $\leq$ create(A) logically follows iff A --↠ P
3. create(A) $\leq$ end(P) logically follows iff P --↠ A
4. begin(Q) $\leq$ end(P) logically follows iff P --↠ Q

Inequalities involving use-variables

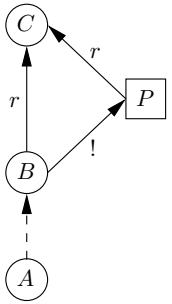
Theorem: An inequality involving use-variables logically follows from the OPM graph if and only if it already belongs to the axioms, or it matches one of six cases:



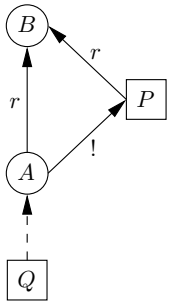
$$\overline{\text{create}(B) \preceq \text{use}(P, r, A)}$$



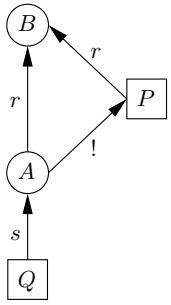
$$\overline{\text{begin}(Q) \preceq \text{use}(P, r, A)}$$



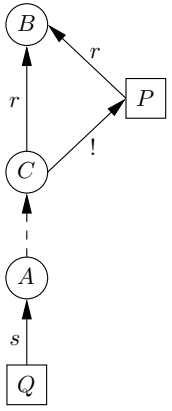
$$\overline{\text{use}(P, r, C) \preceq \text{create}(A)}$$



$$\overline{\text{use}(P, r, B) \preceq \text{end}(Q)}$$



$$\overline{\text{use}(P, r, B) \preceq \text{use}(Q, s, A)}$$



$$\overline{\text{use}(P, r, B) \preceq \text{use}(Q, s, A)}$$

Refinement of OPM graphs

Method of Stepwise Refinement in Software Engineering

Definition: OPM graph H is a **refinement** of OPM graph G if every inequality, involving only variables common to G and H, that logically follows from G, also logically follows from H.

Trivial example: if G is a subgraph of H

Refinement by renaming/merging operations

Let ρ be an arbitrary mapping on artifact ids, process ids, and role ids.

- ids may be mapped to existing ids $\Rightarrow$ merging
- ids may be mapped to new ids $\Rightarrow$ renaming

Call ρ **proper** if $x \neq \rho(x)$ and $\rho(x) \in G$ implies $\rho(\rho(x)) = \rho(x)$.

Theorem: The OPM graph obtained by performing a *proper* merge/renaming is always a refinement.

Further foundational research on OPM

Define a complete set of graph transformation operations that generates all and only refinements

Explore other than temporal semantics for causality (e.g., probabilistic reasoning)

Reference

L. Moreau, N. Kwasnikowska, J. Van den Bussche
A Formal Account of the Open Provenance Model
University of Southampton ECS EPrint 21819, 2010.