UC Santa Barbara

Computer Science Department

On Similarity of Object-Aware Workflows

Mohammad Javad Amiri, Mahnaz Koupaee, Divyakant Agrawal

The 13th IEEE International Conference on Service-Oriented System Engineering

Workflows and Workflow Similarity

- A *workflow* consists of a set of activities performed in coordination in an organizational Environment to accomplish a business goal
- Many large enterprises require hundreds of workflows to fulfill their duties
    - More than **8000** workflows in the Office Automation (OA) systems of China Mobile Communications Corporation (CMCC)
    - SAP reference model covers over **1000** workflows
- Finding similar workflows in workflow repositories helps enterprises to reduce their cost and increase their performance.

Problem: Given a pair of *workflows*, determine whether those two workflows exhibit similar behaviors

Why is Similarity Measurement Important?

Prevent the duplication of activities by merging similar workflows being executed in different parts of an organization

Identify branch workflows that no longer comply with the enterprise reference model

Reduce the cost of expanding businesses by identifying similar workflows when small businesses unite with each other and form a single business

Workflow Similarity using Activity Labels

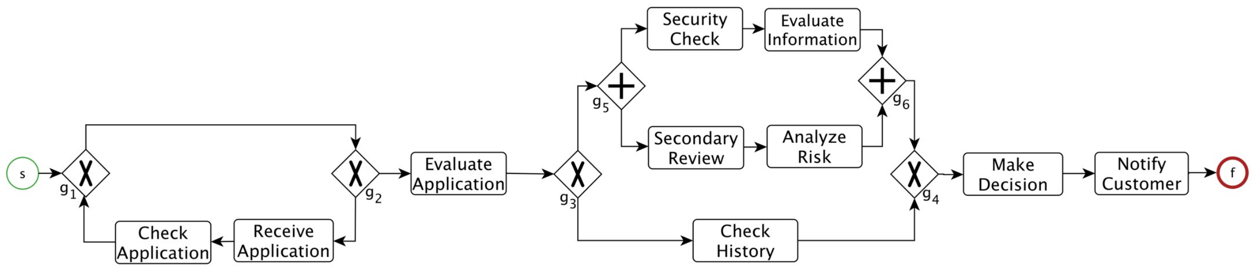
Step 1: Find similar activities using activity labels

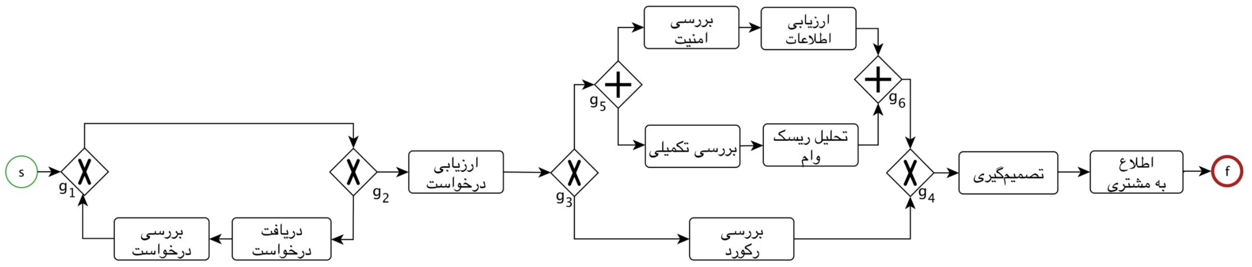
Labels can be compared either *syntactically* or *semantically*

- Syntactically: String edit distance
- Semantically: Natural language processing techniques

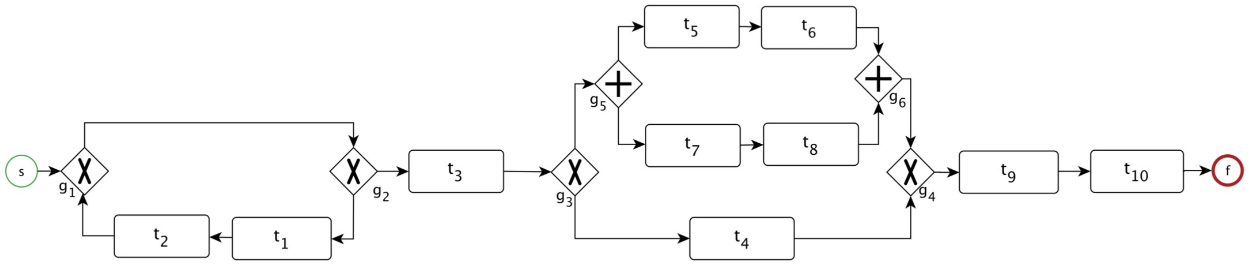
Step 2: measure the similarity of a pair of workflows using the similarity of activities (*structural similarity*)

Activity Labels issues

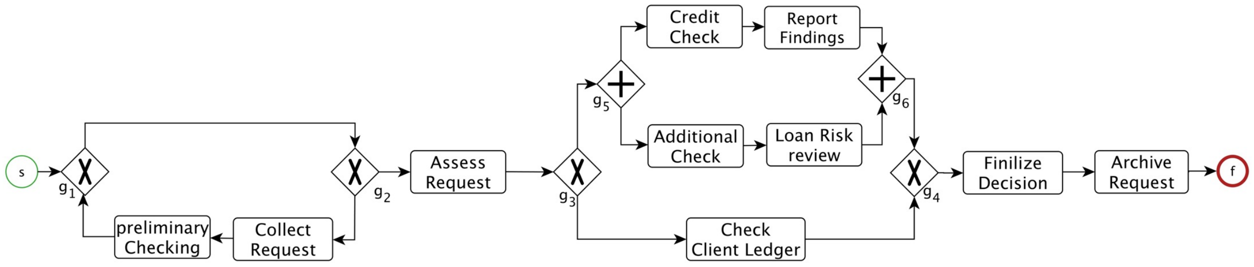




Incomplete or multilingual labels



Meaningless labels



Different words or synonyms

Measure workflow similarity using

Data Objects

Many paradigms model data

- **Decision-aware**
- **Data-aware**
- **Artifact-centric**

Workflow Similarity using Data Object

Use data access patterns (Reads and Writes)

Step 1: Find similar activities using data access patterns

Step 2: Measure the similarity using the similarity of activities

- Activities might have different granularities
  - fine-grained activities: perform a single read/write operation
  - coarse-grained activities: fulfill a service

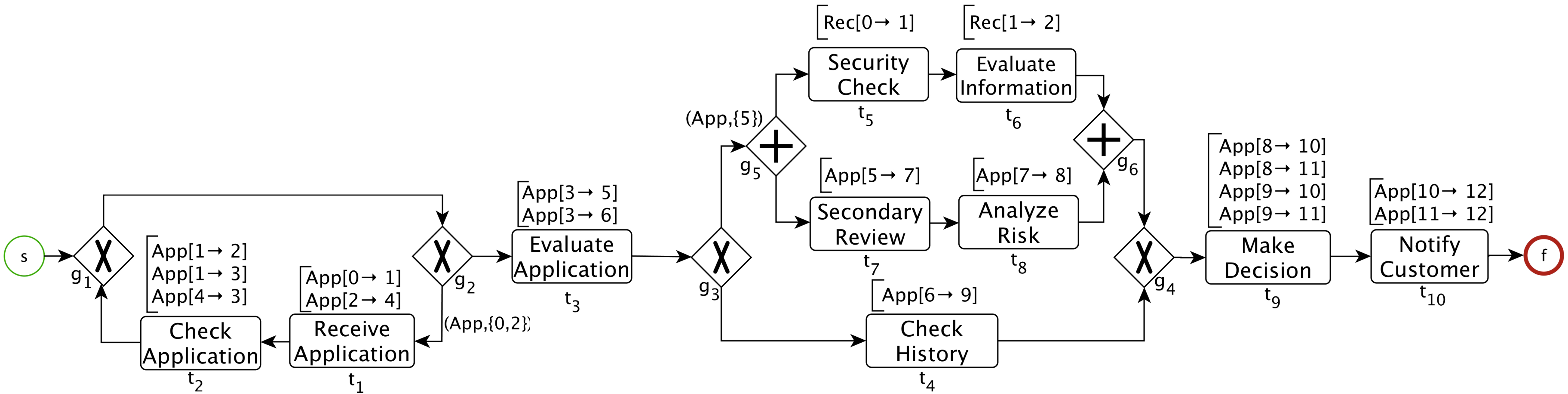
Object life cycles

The state of each object evolves during the execution of a workflow

object life cycle: the behavior of a data object in terms of state changes

(Object-aware) Workflow Schema

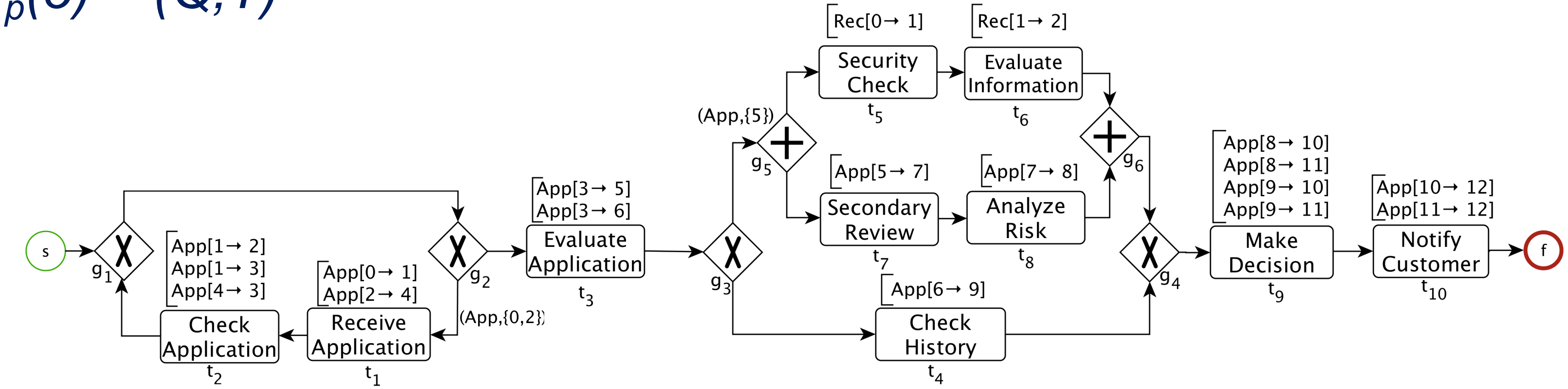
- P = (N, s, f, L, E, O)
  - N = {g_1, …, g_6, t_1, …, t_{10}}
  - O = {App, Rec}
  - Activity: a set of objects and transitions: (α,O,τ)
  - Schema size |P|: number of activity nodes within the workflow



App: Application, Rec: LoanRecord. States of Rec: 0(RecCreated), 1(SecurityChecked), 2(InfoEvaluated). States of App: 0(Initiated), 1(Received), 2(Incomplete), 3(Complete), 4(Resubmitted), 5(MoreInfoNeeded), 6(Evaluated), 7(Reviewed), 8(RiskAnalyzed), 9(HistoryChecked), 10(LoanApproved), 11(LoaanDenied), 12(Archived).

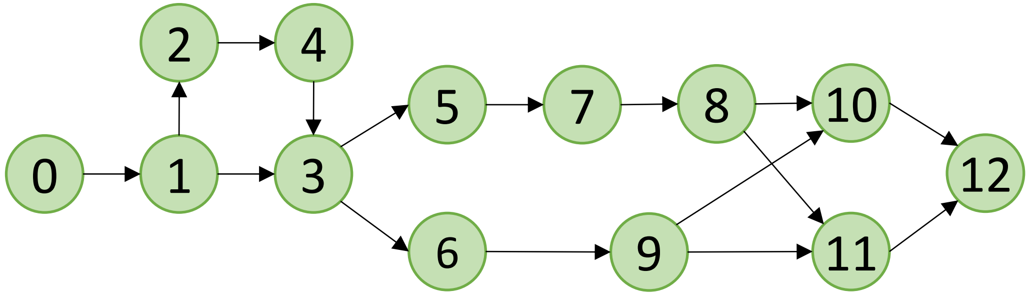
Object Life Cycles

$G_p(o) = (Q, T)$



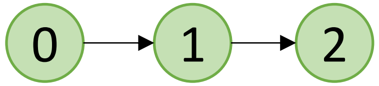
App Object

13



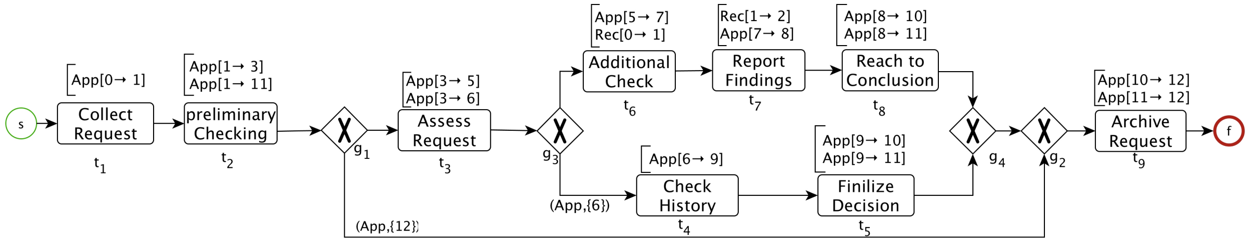
Rec Object

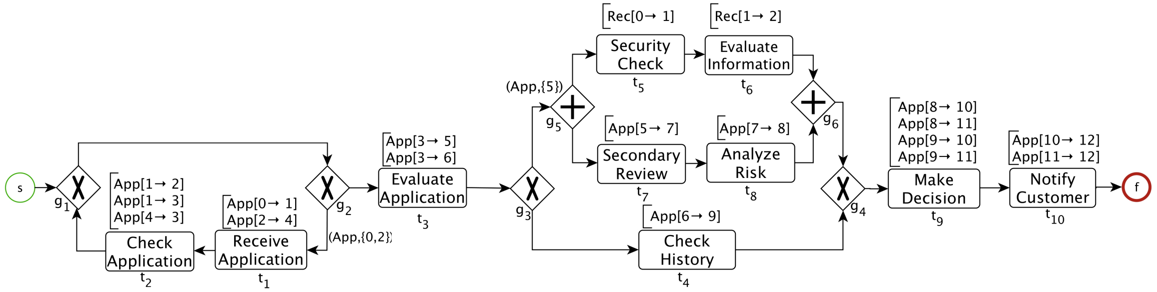
3



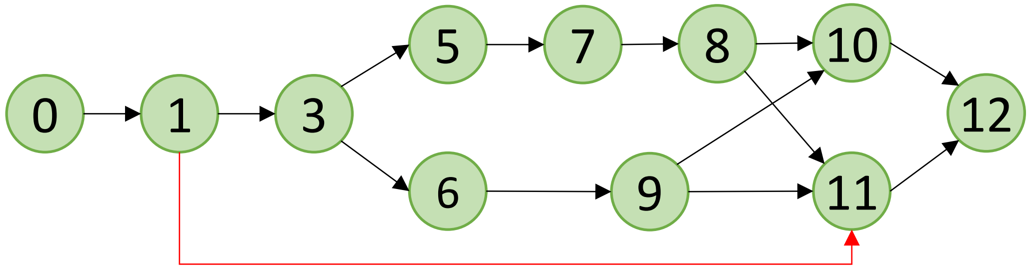
Object life cycle size |*G*|: number of states within the life cycle graph

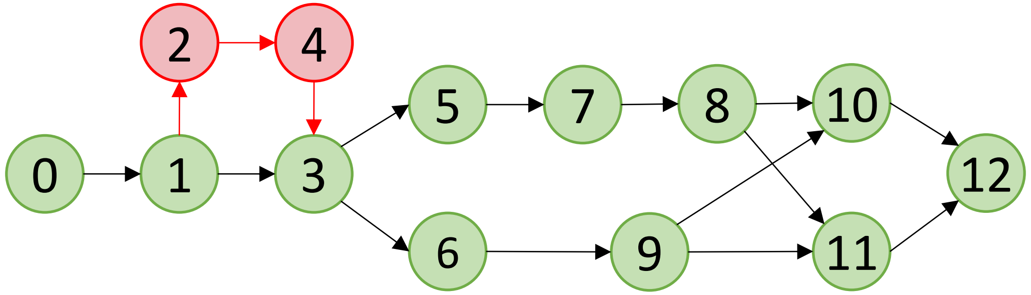
Object Life Cycles Similarity



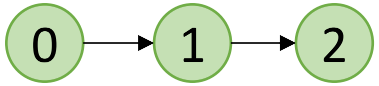


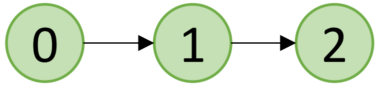
App Object





Rec Object

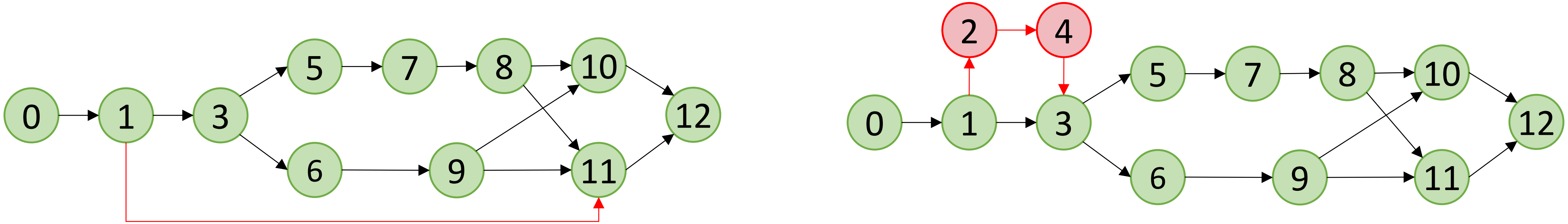




State Similarity

- For each state *q*: two sets of *successor* (σ_q) and *predecessor* (π_q) states
- State similarity of a state *q* in two workflows *P* and *P’*

$$Sim^{s}_{(P,P')}(q) = 0.5 * \left(\frac{\sigma_P(q) \cap \sigma_{P'}(q)}{\sigma_P(q) \cup \sigma_{P'}(q)} + \frac{\pi_P(q) \cap \pi_{P'}(q)}{\pi_P(q) \cup \pi_{P'}(q)}\right)$$

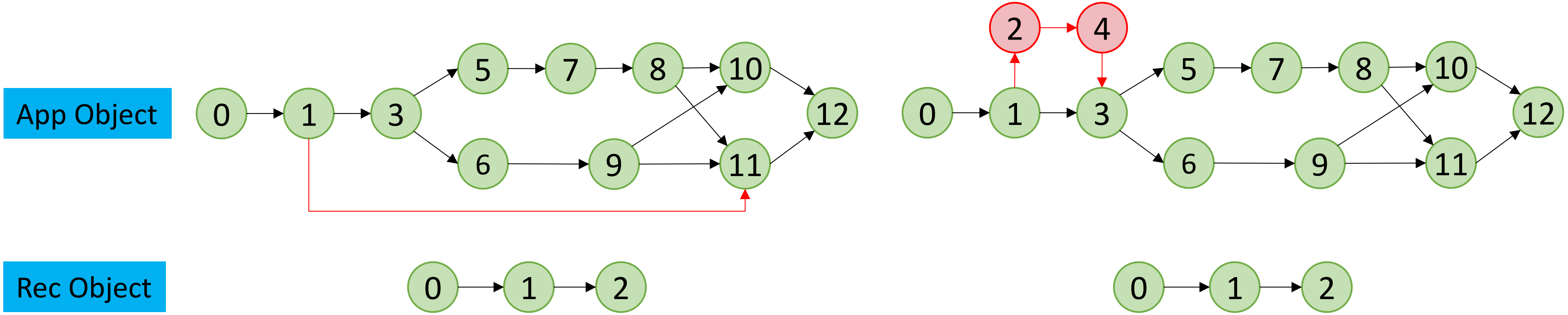


$$Sim^{s}_{(p1,p2)}(1) = 0.5 * \left(\frac{\{3,11\} \cap \{2,3\}}{\{3,11\} \cup \{2,3\}} + \frac{\{0\} \cap \{0\}}{\{0\} \cup \{0\}}\right) = 0.66$$

Object similarity

- Let $G_P(o)=(Q,T)$ and $G_{P'}(o)=(Q',T')$ be the life cycles of object *o* in *P* and *P'*
- The object similarity of *o* is

$$\mathrm{Sim}^o(G_P(o),G_{P'}(o)) = \frac{\sum_{q\,\in(Q\cap Q')} \mathrm{sim}^s_{(P\,P')}(q)}{|Q\cup Q'|}$$



$$\mathrm{Sim}^o(G_{P1}(\mathrm{App}),G_{P2}\,(\mathrm{App})) = 0.79$$

$$\mathrm{Sim}^o(G_{P1}(\mathrm{Rec}),G_{P2}\,(\mathrm{Rec})) = 1$$

Workflow Similarity

- Similarity of two workflows is computed using the similarity of their objects

Object life cycles have different sizes

$$Sim^D(P,P') = \frac{\sum_{o \in (O \cap O')} (sim^o(G_p(o), G_{p'}(o)) * (|\{Q_o \cup Q'_o\}|))}{\sum_{o \in (O \cap O')} |\{Q_o \cup Q'_o\}|}$$

$$Sim^D(P_1,P_2) = \frac{(0.79 * 13)+(1 * 3)}{13+3} = 82.75$$

Evaluation Setup

- Algorithms: Object-aware similarity and Activity-based similarity
- Dataset: 10 workflows, each with five instances (a reference workflow and four variants)
- For each workflow , three experts are asked to rank variants separately [1,2,3,4]
- The conformance of each of the two approaches with experts' opinions is quantified using the resulting ranks.
    - Ranking score
    - Number of similar workflow

Evaluation Results

	P_{i1}			P_{i2}			P_{i3}			P_{i4}			Score (10)	
	Sim^A	Sim^D		Sim^A	Sim^D		Sim^A	Sim^D		Sim^A	Sim^D		Sim^A	Sim^D
P_0	0.93	0.94	1	0.84	0.75	2	0.65	0.77	4	0.61	0.64	3	9	8
P_1	0.88	0.76	2	0.84	0.88	3	0.82	0.81	1	0.45	0.70	4	1	10
P_2	1	1	1	0.81	0.92	2	0.68	0.77	3	0.62	0.81	4	10	9
P_3	0.94	0.78	2	0.91	0.82	1	0.72	0.75	3	0.64	0.72	4	7	10
P_4	0.73	0.85	1	0.54	0.73	2	0.48	0.61	4	0.44	0.79	3	9	8
P_5	0.91	0.88	1	0.84	0.78	2	0.70	0.83	3	0.54	0.67	4	10	8
P_6	0.79	0.85	1	0.75	0.69	3	0.72	0.76	2	0.57	0.71	4	8	9
P_7	0.96	0.96	1	0.93	0.91	2	0.87	0.82	4	0.87	0.84	3	9	10
P_8	0.98	0.96	1	0.91	0.89	2	0.66	0.71	4	0.63	0.75	3	9	10
P_9	0.66	0.74	1	0.58	0.71	2	0.47	0.63	3	0.44	0.66	4	10	10

Rank 1: 4 points
Rank 2: 3 points
Rank 3: 2 points
Rank 4: 1 points
1⇆ 2: 4 points (from 7)
2 ⇆ 3: 3 points (from 5)
3 ⇆ 4: 2 points (from 3)

activity-based: 82/100
object-aware: 92/100

Evaluation Results

	Activity-based	Object-aware	Experts
P_0	P_{01}, P_{02}	P_{01}	P_{01}
P_1	P_{11}, P_{12}, P_{13}	P_{11}, P_{12}	P_{11}, P_{12}
P_2	P_{21}, P_{22}	P_{21}	P_{21}
P_3	P_{31}, P_{32}	P_{31}	P_{31}, P_{32}
P_4	∅	P_{41}	P_{41}
P_5	P_{51}, P_{52}	P_{51}, P_{53}	P_{51}
P_6	∅	P_{61}	P_{61}
P_7	P_{71}, P_{72}, P_{73}, P_{74}	P_{71}, P_{72}, P_{73}, P_{74}	P_{71}, P_{72}, P_{73}, P_{74}
P_8	P_{81}, P_{82}	P_{81}, P_{82}	P_{81}, P_{82}
P_9	P_{91}	P_{91}, P_{92}	P_{91}

Metrics	# of variants for different approach	
	Activity-based	Object-aware
True Positive (TP)	14	15
False Negative (FN)	2	1
False Positive (FP)	4	2
True Negative (TN)	20	22

	Activity-based	Object-aware
Precision	0.78	0.88
Recall	0.87	0.94
Accuracy	0.85	0.93

$$\text{Precision} = \frac{TP}{TP+FP} \qquad \text{Recall} = \frac{TP}{TP+FN} \qquad \text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

Future Work

Extend the approach to other attributes of data objects

Extend the approach to relations between data objects

Find the most similar pair without comparing all the existing workflows

Consider the dependencies between object life cycles

THANK YOU!

Questions?!