Temporal Representation of Scientific Data Provenance

Peng Chen, Beth Plale, Mehmet Aktas

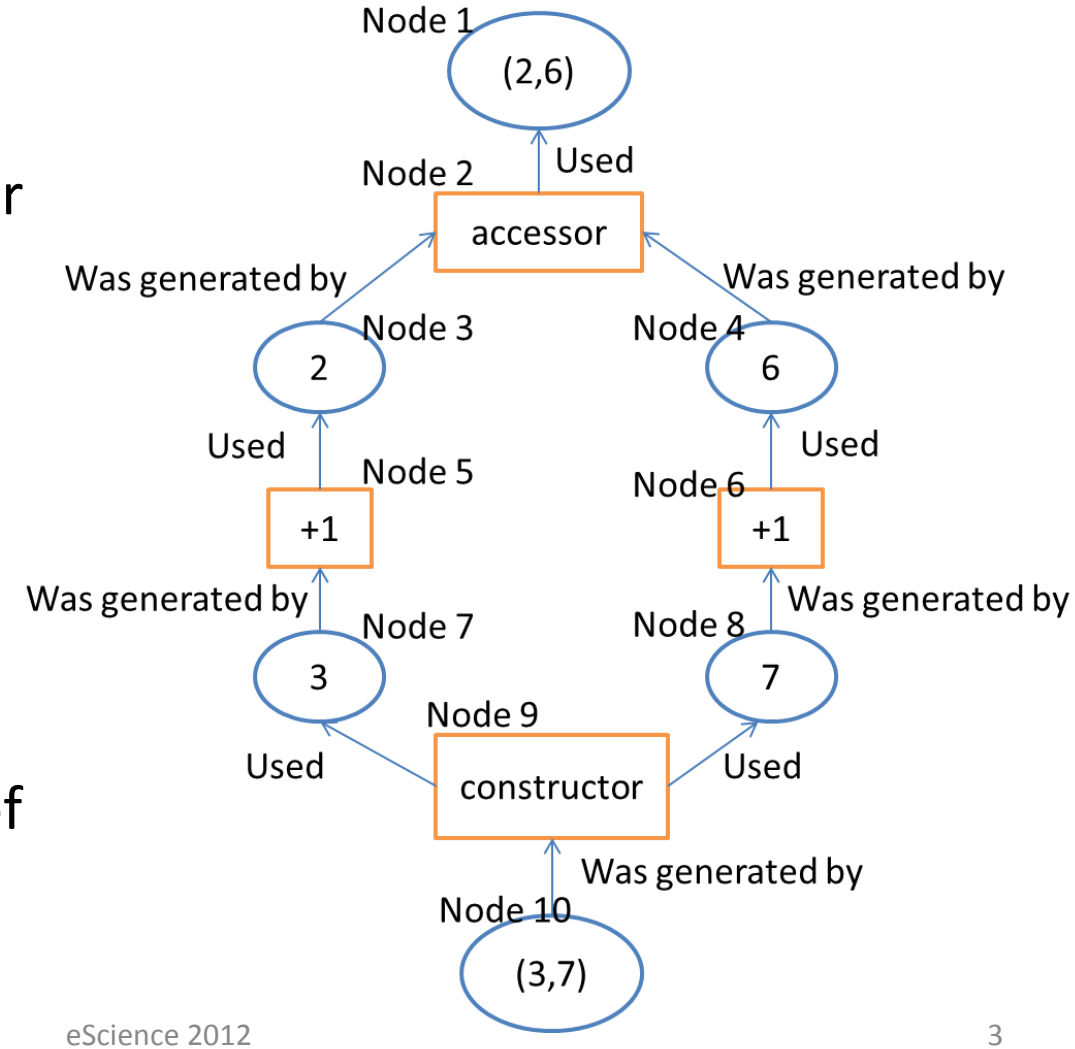
INDIANA UNIVERSITY
SCHOOL OF INFORMATICS AND COMPUTING
Bloomington

Provenance of scientific data

- The provenance of a scientific data product or collection is its lineage; a record of the factors contributing to the product as it exists today.
- The provenance of a data product can be used to gain a deeper understanding of the data, particularly in cases where the data are shared more broadly.

Provenance as a causal graph

- Open Provenance Model (OPM) is community consensus data model for representation of provenance for interoperability
- An OPM graph is a historical representation
- Provenance is often enriched beyond causal relationships by means of annotation.



Outline

- Introduction
- Motivation
- Temporal representation
- Feature selection
- Experiment
- Future work

Motivation

- Provenance information can be highly voluminous
- Existing approaches to graph representation lose either *structural information* or *attribute information* in the graph
  - Labeled workflow graph (Santos et al 2008)
  - Weighted complete dependency graph (Jung and Bae 2006)
  - Multidimensional vector (Salton et all 1975)

Limitations to existing approaches

- Labeled Workflow Graph
  - Capture labeled nodes (such as name, parameters) and the unlabeled edges between nodes.
  - Similarity is calculated by edit distance, subgraph isomorphism, and Maximum Common Induced Subgraph (MCIS)
  - Most attribute information is lost
- Weighted Complete Dependency Graph
  - A weighted graph supplemented with inferential transitions to consider delicate structural independencies of their activities.
  - Most attribute information get lost

Limitations to existing approaches, cont.

- Multidimensional Vector
  - The dimensions in the vector space are defined by the union of all the possible node attributes the workflows in the input set may contain.
  - < Node Name 1, Node Name 2, Node Name 3 …>
  - Structure information is lost

Other related work

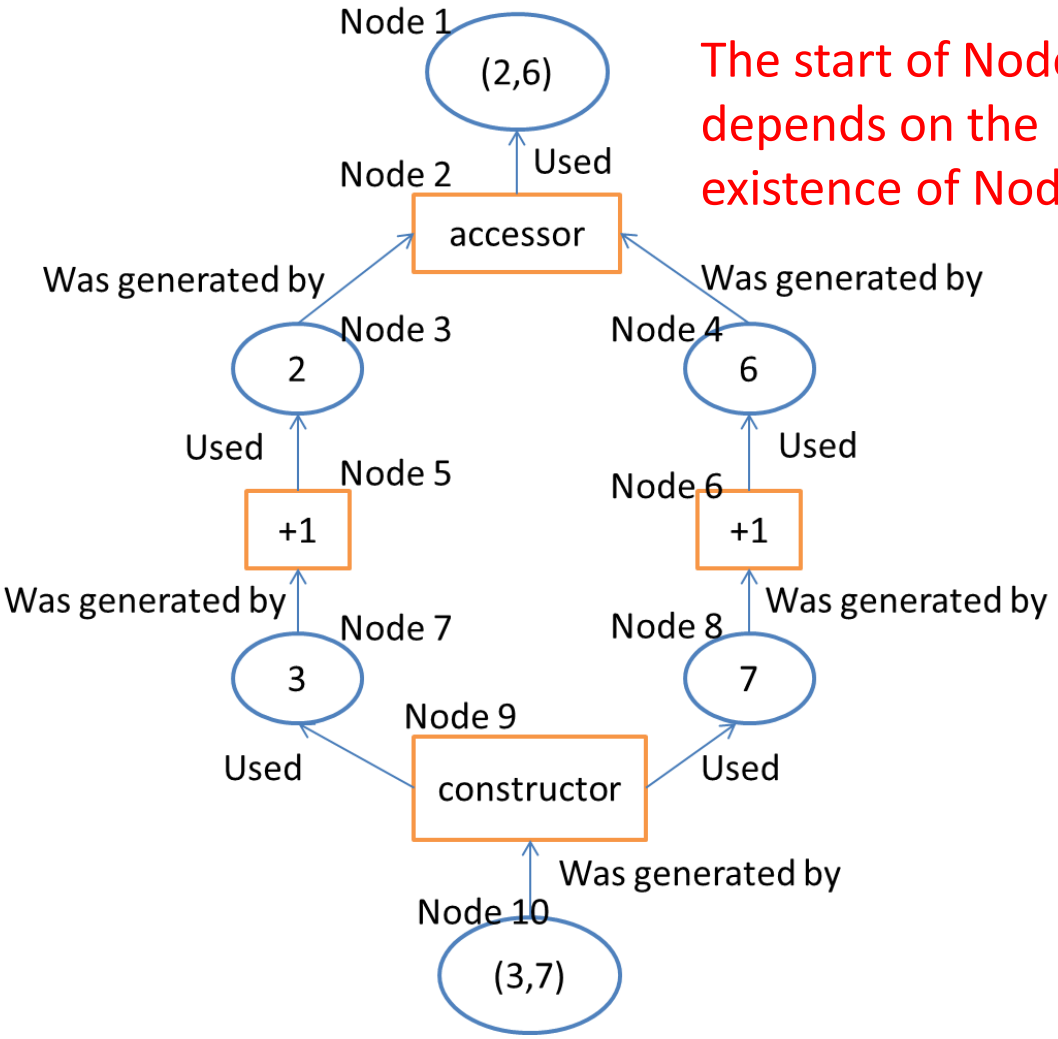
- Data mining inside provenance graph
  - Margo et al. (USENIX 2010) extract semantic information from file system provenance through application of data mining and machine learning techniques to file metadata.
  - Simmhan, Y.L. et al. (SciFlow 2006) uses decision tree inductive machine learning technique to classify discrete and continuous valued attributes into quality scores, thus to automatically determine the quality of provenance.
- To our best knowledge, there is no prior research on mining a collection of provenance graphs

Outline

- Introduction
- Motivation
- Temporal representation
  - Logical-P
  - Graph Partitioning
- Feature selection
- Experiment
- Future work

Representation of Provenance

- OPM compliant graphs have a logical temporal ordering



Proposed Approach

- We propose a logical time representation of provenance graphs that:
  - Substantially reduces the amount of provenance information needed

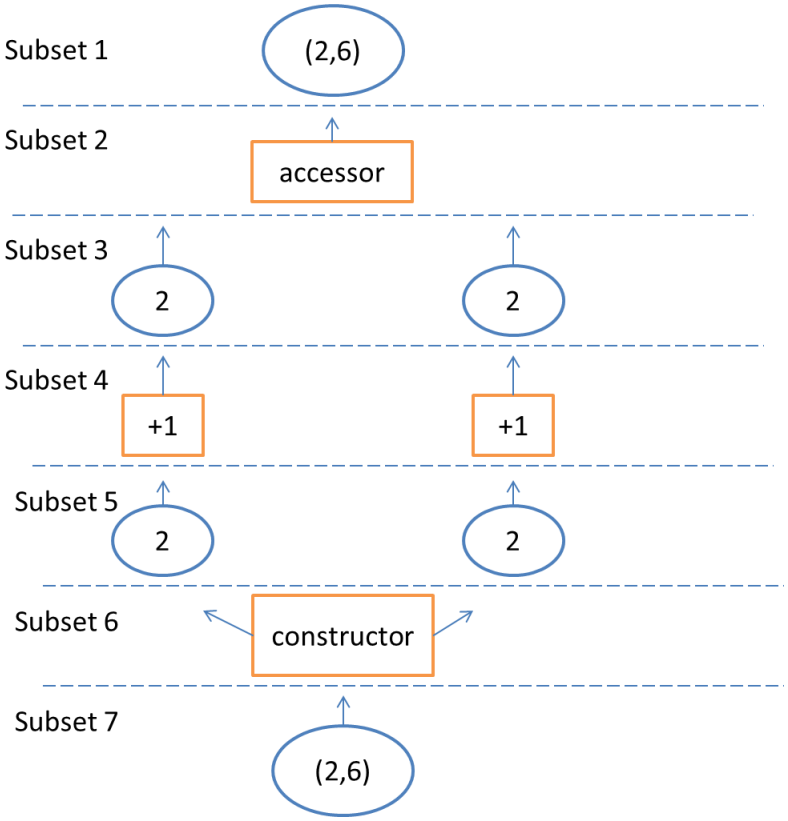
  While still …
  - Preserving key functionality over the graphs as evidenced by application of data mining algorithms.

Goals of Proposed Representation

- Able to support patterns that describe and distinguish general properties of datasets in provenance repositories
  - Through training classifier and mining association rule set
- Able to support detection of faulty provenance data
  - Through clustering, by checking cluster centroids in case where correct and faulty provenance are naturally separated into different clusters
- Able to find more descriptive knowledge of provenance clusters
  - Through mining association rules that reflect workflow variants

Solution Outline

- Partition provenance graph into subsets with temporal order, then extract information from each subset to generate a representation sequence

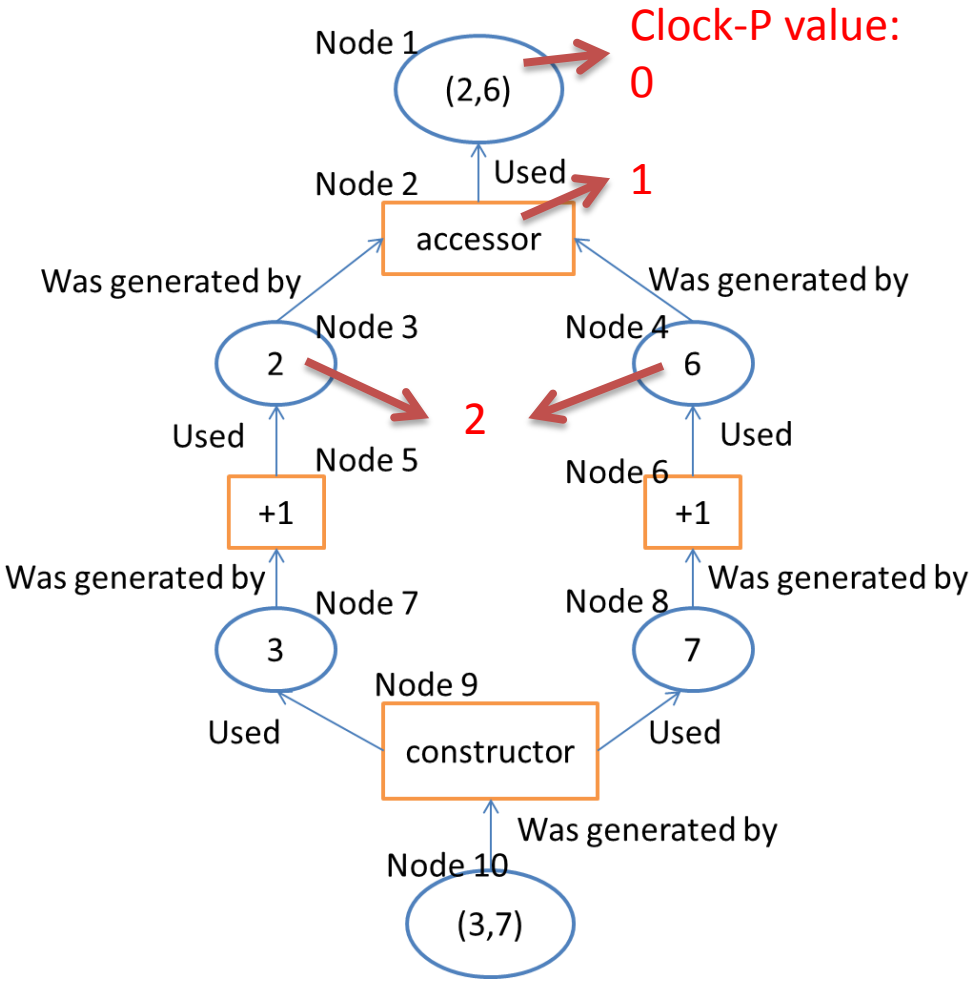


Logical Clock-P

- Logical Clock-P is a function C that takes a node as input and produces an non-negative integer as output. This function maps an integer to each node of a given provenance graph.
- Correct logical clocks must satisfy:

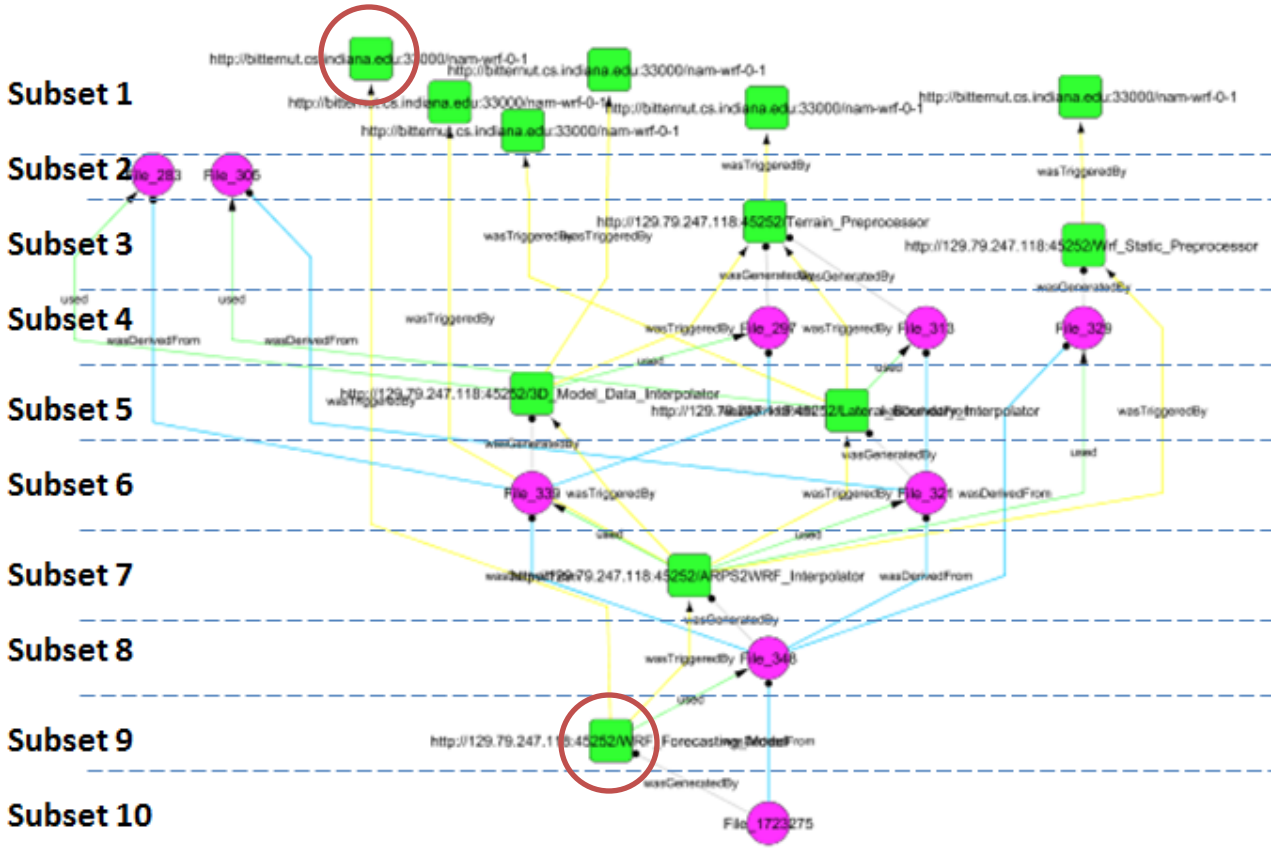
Clock Condition: For any node a, b:

If a -> b, then C(a) < C(b)



Example of partitioning

Partitioning of provenance graph is set of non-overlapping and non-empty subsets of nodes based on logical clock-P and node type.

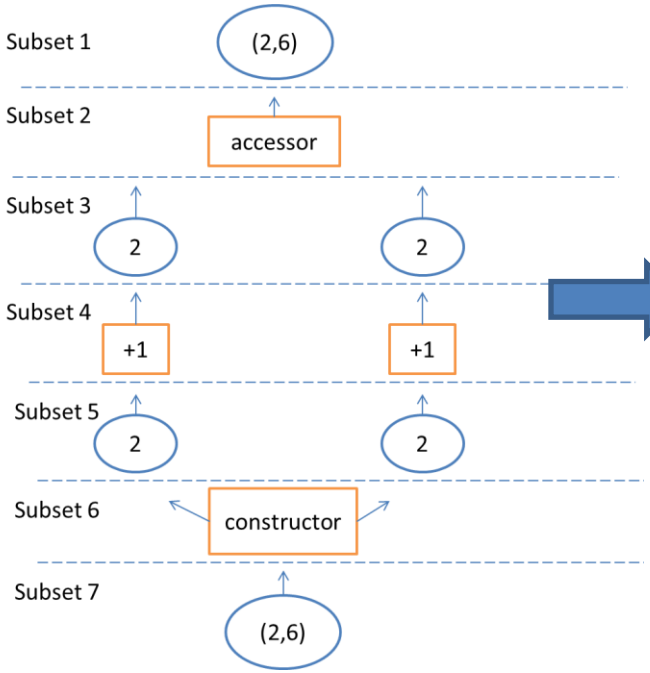


Outline

- Introduction
- Motivation
- Problem Outline
- Temporal Representation
- Feature selection
- Experiment
- Future work

Feature Selection

- Statistical feature space: extract statistical features from each subset by:
  - Application of statistical functions: avg, max, min, dev
  - Identify node features: node type, node number, node in/out-degree, node label length, …

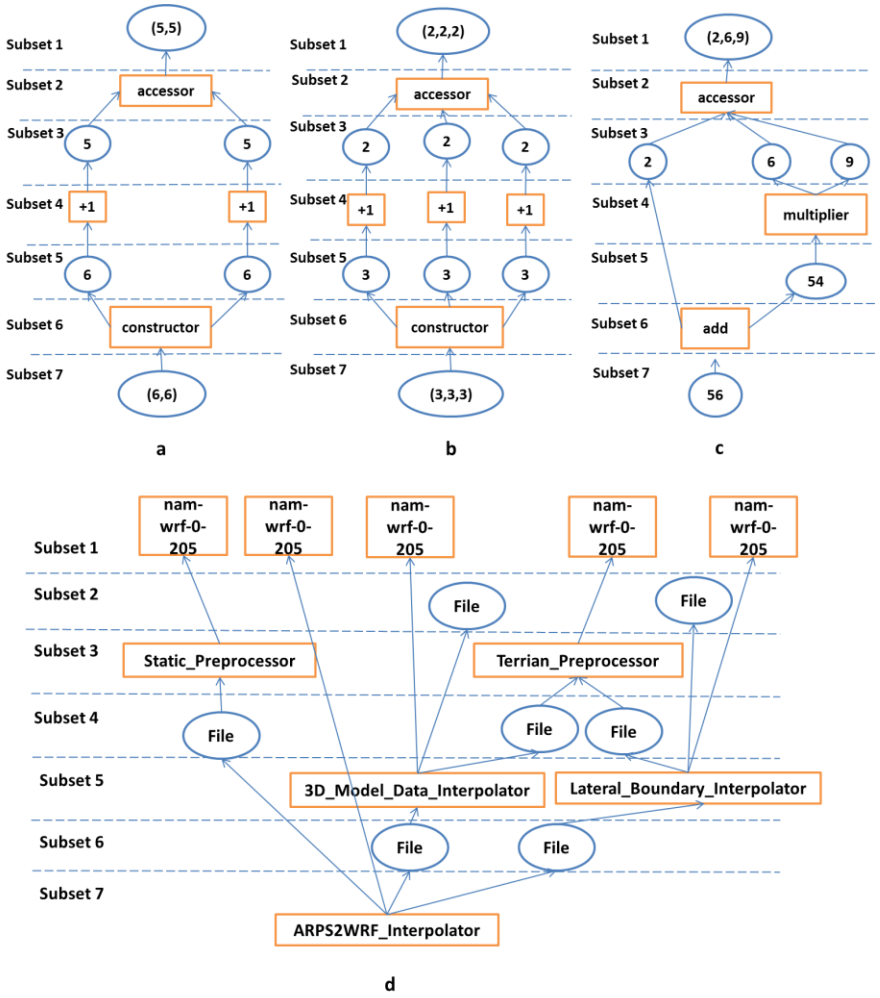


<Type of nodes in subset, Number of nodes in subset, Average number of in-degree of nodes in subset, Average number of out-degree of nodes in subset >

$(<Artifact, 1, 1, 0>, <Process, 1, 2, 1>, <Artifact, 2, 1, 1>, < Process, 2, 1, 1 >, < Artifact, 2, 1, 1>, < Process, 1, 1, 2>, < Artifact, 1, 0, 1>)$

Feature selection

- In current research, we select feature set by studying provenance data and mining objectives, and then evaluate it by testing the performance of clustering
  - We do not utilize feature selection algorithms that can lead to optimal feature set (future work)



Similarity Distance Evaluation

- To distinguish between two graphs based on their *attribute difference*, we capture:

  <Type of nodes in subset, Number of nodes in subset, Average number of characters in name of nodes>

Distance between	Euclidean distance in time domain	Euclidean distance in frequency domain
Fig (a) – Fig (b)	2.6458	0.1879
Fig (a) – Fig (c)	12.0416	0.6078
Fig (b) – Fig (c)	12.083	0.5821

- We similarly choose features to distinguish between two graphs based on their *structure difference*. See paper for details.

Outline

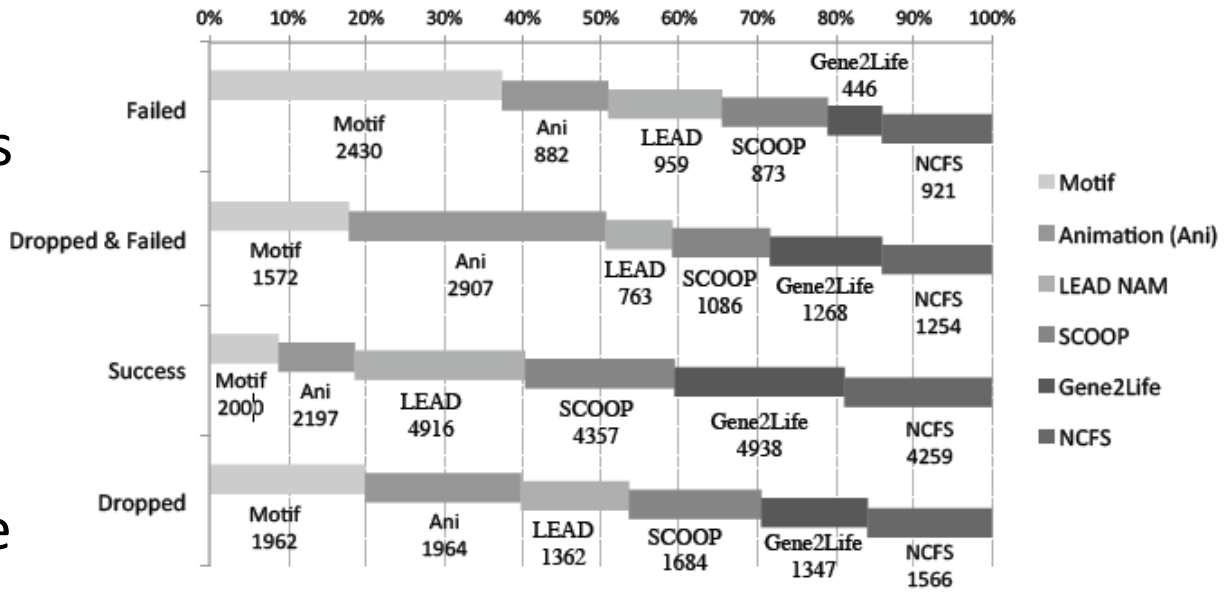
- Introduction
- Motivation
- Temporal representation
- Feature selection
- Experiment
- Future work

Evaluation

- Logical Clock-P: apply to generate temporal representations in a) time domain and b) frequency domain.
- Apply data mining:
  - Task 1.) unsupervised clustering, time domain
  - Task 2.) classification, freq domain
  - Task 3.) association rule mining, time domain
  - Task 4.) unsupervised clustering, freq domain (not included, see paper)
- Data: 10GB database of synthetic provenance
- Environment: Dual-Core Intel M540, 4GB, Windows Vista, Weka Version: 3.6.4

10 GB Provenance Database

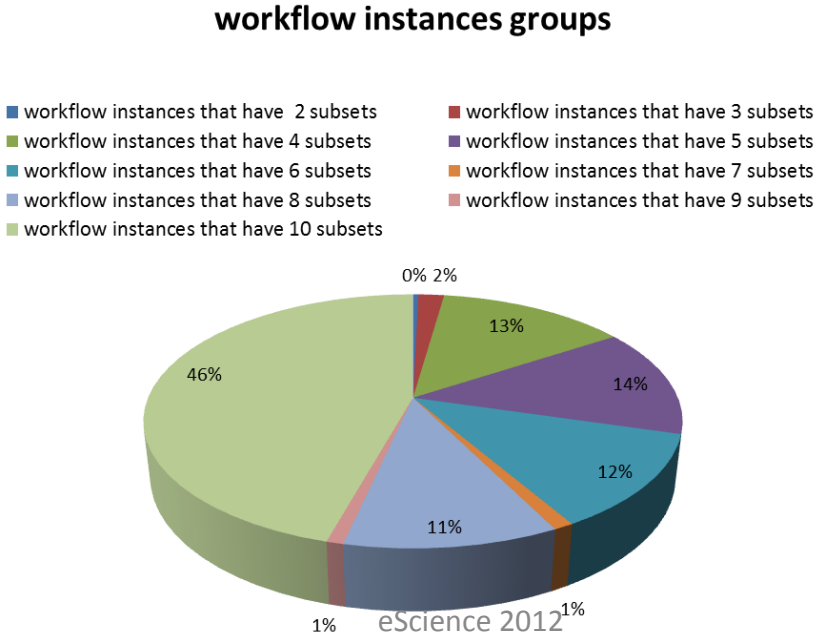
- Provenance from 48,000 workflows modeled on six real e-science workflows in weather, coastal, bioinformatics, CS, and biomedical
- Failure model used in creating database produced workflows with known failure properties



*Cheah, Y. and Plale, B. and Kendall-Morwick, J. and Leake, D. and Ramakrishnan, L.: A Noisy 10GB Provenance Database. 2nd Int' l Workshop on Traceability and Compliance of Semi-Structured Processes (TC4SP2011), 2011

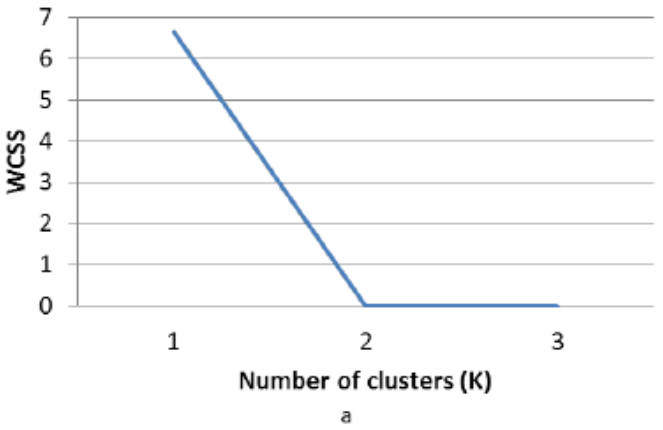
Task 1: unsupervised clustering time domain

- Use Euclidean distance as similarity measurement *limits application of k-means clustering to representation sequences of same length*
- Hence, group together provenance representations with same number of temporal subsets, then apply k-means clustering algorithm within each group.

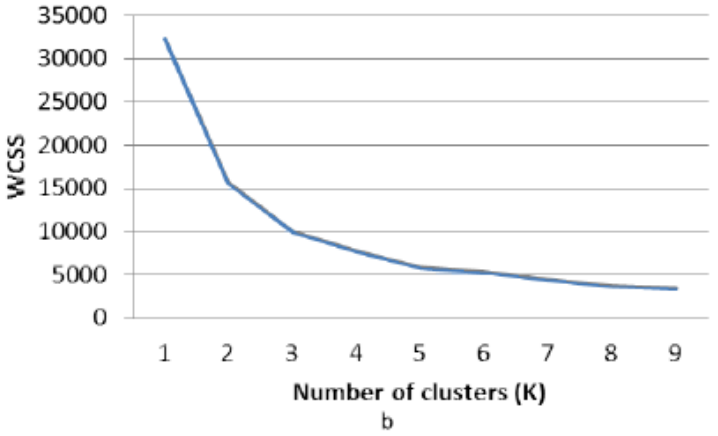


Selecting “K” for K-means

- To determine “k”, we plot within-cluster sum of squares (WCSS) and look for “elbow point”
- e.g., when subset is of length 9, k=3



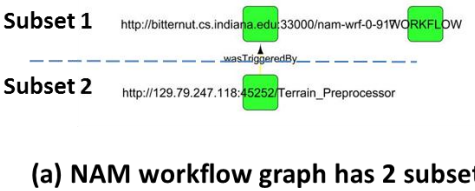
a



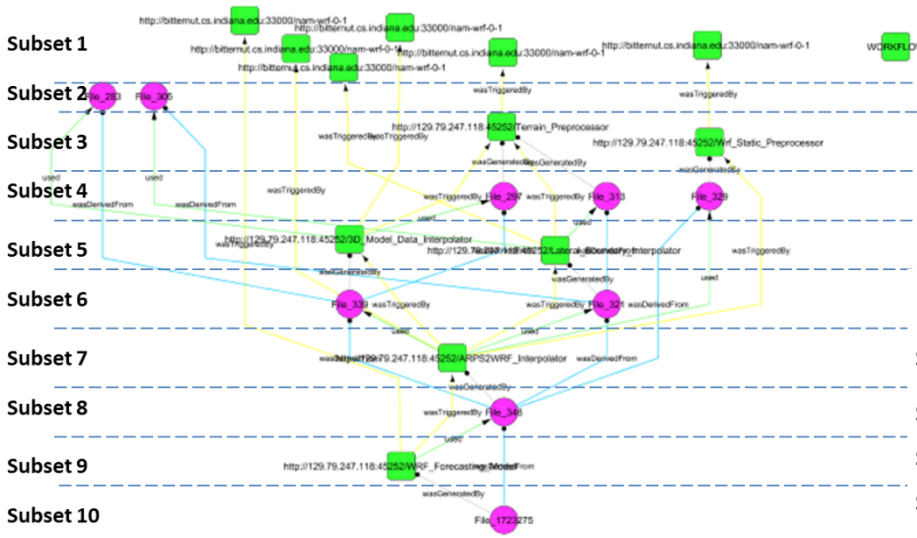
b

Task 1: Outcome

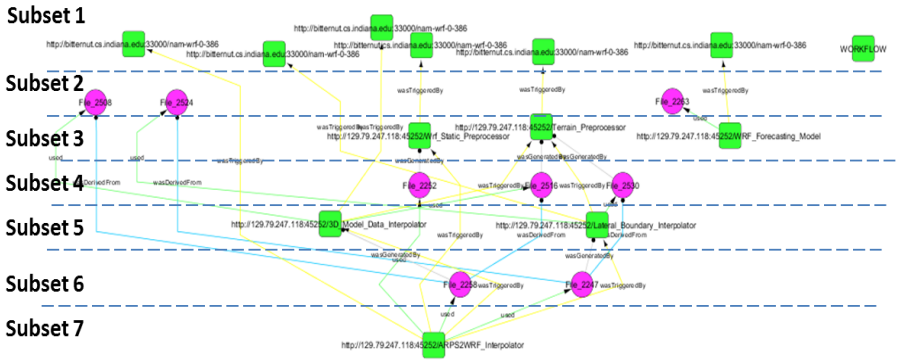
- Correct and faulty provenance graphs are naturally separated into different clusters, so we can tell faulty provenance by comparing centroid provenance graphs



(a) NAM workflow graph has 2 subsets



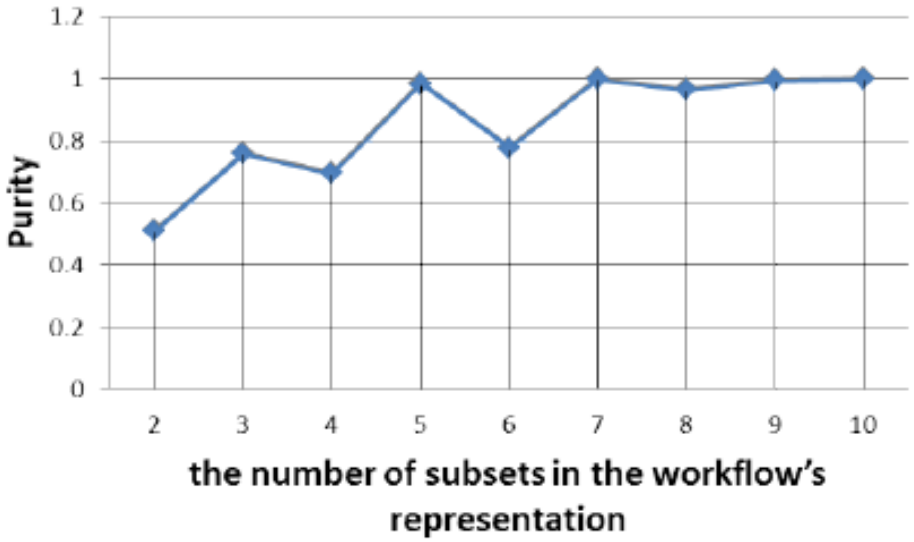
(c) NAM workflow graph that has 10 subsets



(b) NAM workflow graph that has 7 subsets

Task 1: Quality Assessment

- Purity is a measure of quality of clustering process. Purity is shown to be high.



* Y. Zhao and G. Karypis. Criterion functions for document clustering: Experiments and analysis. 2002

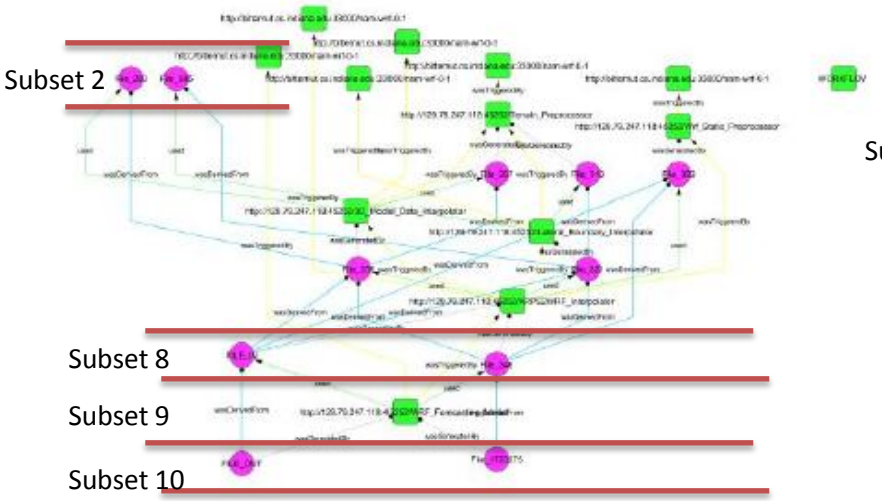
Task 2: Workflow type classification, freq domain

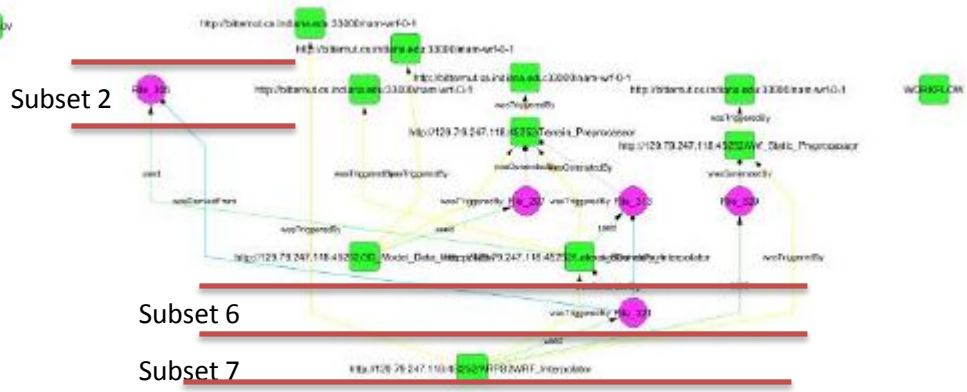
- Desired outcome: Given a new provenance graph, can we tell which workflow type it belongs to?
- Evaluation: utilize Bayes Network Classifier (Weka). Summary of 10-fold-cross validation given below

Weka Scheme	10-fold-cross validation summary
weka. classiers. bayes. BayesNet –D -Q weka. classi ers. bayes.net. search.local.K2 - -P 1 -S BAYES -E weka. classi ers. bayes.net. estimate.SimpleEstimator-A 0.5	Correctly Classified Instances 46307 96.6461 % Incorrectly Classified Instances 1607 3.3539 % Kappa statistic 0.9598 Mean absolute error 0.0113 Root mean squared error 0.089 Relative absolute error 4.053 % Root relative squared error 23.872% Coverage of cases (0.95 level) 98.7811 %

Task 3: Association rules mining, time domain

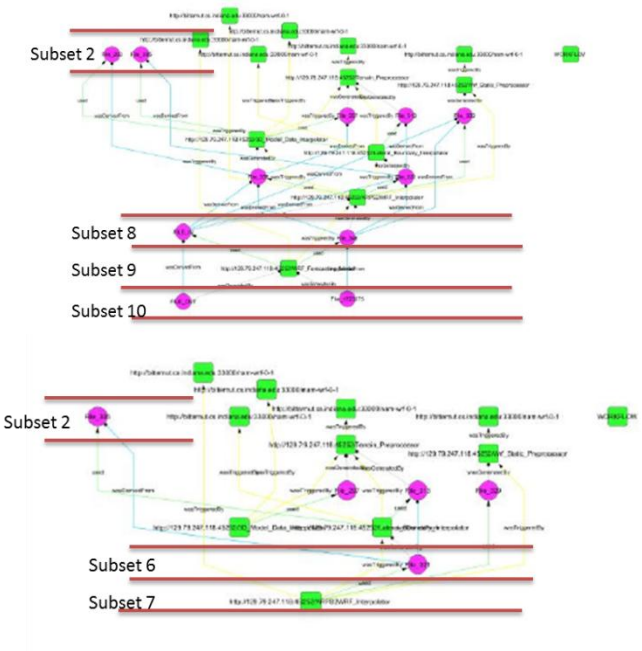
- We manually introduce two additional error cases to weather forecast workflow to explore value of association rules mining





Task 3: association rules mining

- Apriori algorithm is less efficient when dealing with long sequences, so select attribute “number of nodes in subset“ from each subset, forming new representation sequence of length 10.



Weka Scheme	Sample of association rules found
Weka.associatio n.Aprior –N 10 – T 0 –C 0.9 –D 0.05 –U 0.4 –M 0.1 –S -1.0 –c -1	1) numberOfNodes_8=‘(0.8 – 1]’ ➔ numberOfNodes_10 = ‘(0.8 – 1]’ 2) numberOfNodes_8=‘(1.8– inf]’ ➔ numberOfNodes_10 = ‘(1.8 – inf]’ 3) numberOfNodes_2=‘(-inf – 1]’ ➔ numberOfNodes_8 = ‘(-inf – 0]’

Key outcomes

Approach	Key outcomes
Unsupervised clustering / time-domain representation	Can be used to detect failed workflow instances
Unsupervised clustering / frequency domain representation	Representations do not maintain meaningful information for mining association rules
Classification / frequency domain representation	Can predict workflow type for new workflow instances
Association rules mining/ time domain representation	Can capture causal relationships between subsets Some association rule sets can distinguish different clusters Association rules on time-domain representation reflects patterns only on statistical features

Future Work

- Feature selection: utilize feature ranking and subset selection algorithms in feature selection
- Application: extend work to a less “well behaved” provenance data set
- Compare: other temporal data mining techniques
- Extend this approach to other provenance-specific questions
- Performance: improve scalability of representation process using MapReduce

Peng Chen <chenpeng@umail.iu.edu>