Provenance Analytics

Zachary G. Ives
University of Pennsylvania

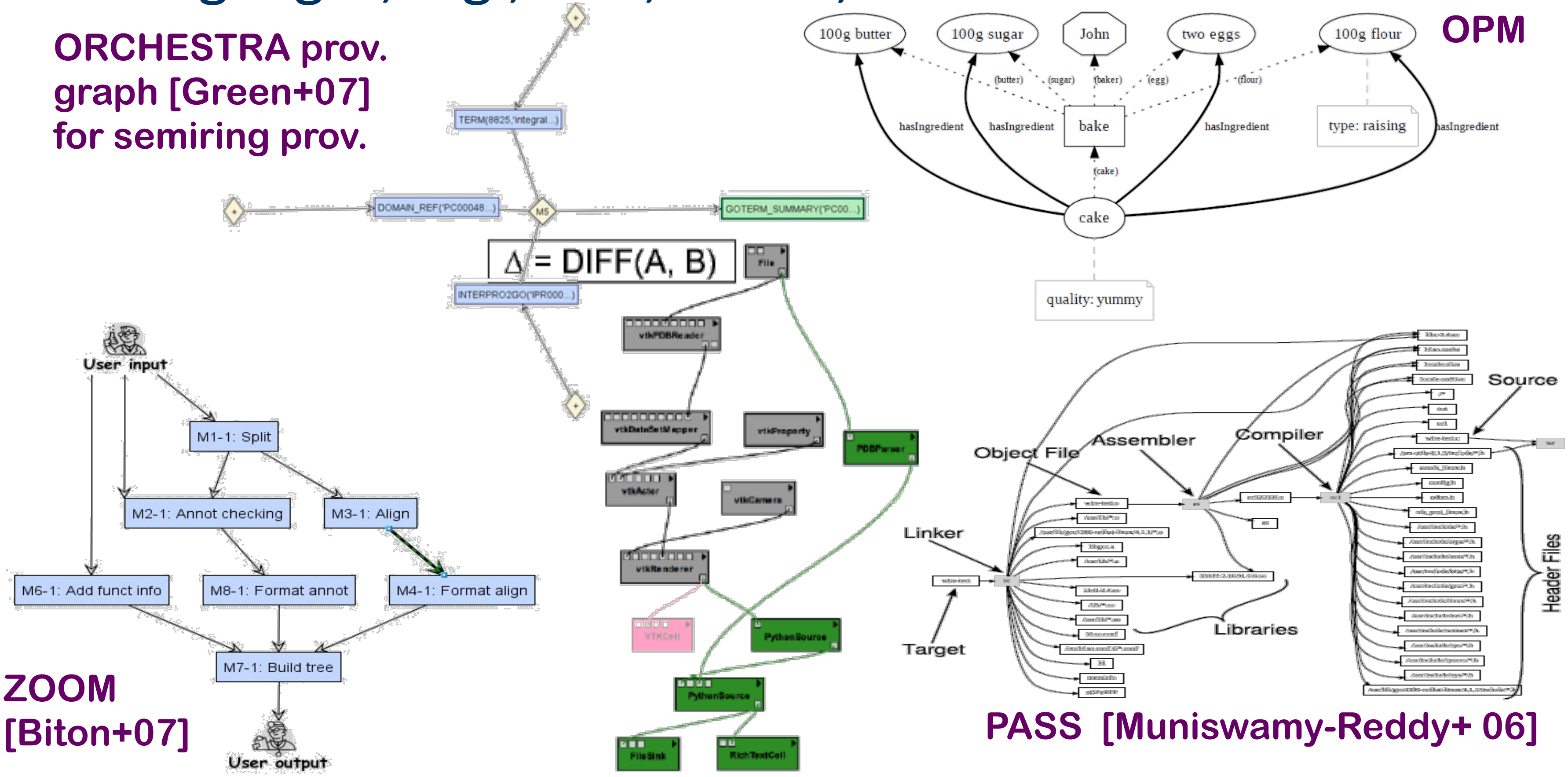
TaPP 2011
June 24, 2011

Provenance: The All-Important Sidekick

- Provenance's role is to enable reasoning about data:
  - Automating a manual procedure
  - Improving a procedure
  - Debugging / diagnosing a problem, or understanding what's going on
  - Assessing data quality / trustworthiness
- "Provenance analytics"
  - Learning from data relationships…
  - Extract→ abstract → visualize, assess, process-mine, …

Enabling "Manual" Analysis: Visualizing Provenance Sub-graphs/trees

- Based on browsing, query by analogy (VisTrails), or languages, e.g., PQL, ProQL, …



ORCHESTRA prov. graph [Green+07] for semiring prov.

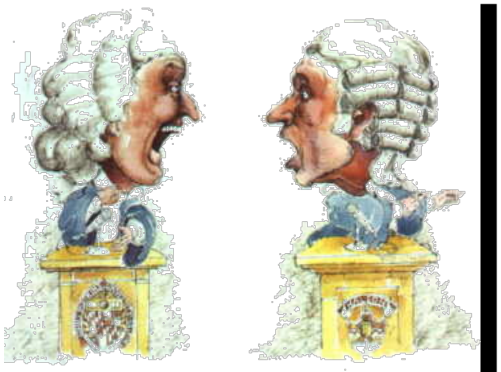
OPM

ZOOM [Biton+07]

VisTrails [Scheidegger+ 08]

PASS [Muniswamy-Reddy+ 06]

Some Still-Debated Questions about Querying Provenance



- Often a human needs to "see what's going on"…
- But what do show as query results?
  - Subtrees (PQL)
  - Subgraphs (SPARQL, ZOOM, …)
  - Data with annotations (semirings)
  - Subgraphs + bindings + annotations (ProQL)

➢ How do we help the user focus on what's important?

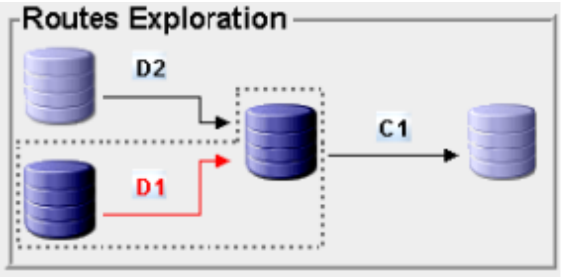
❖Challenges: query, scale, navigation, drill-down interaction

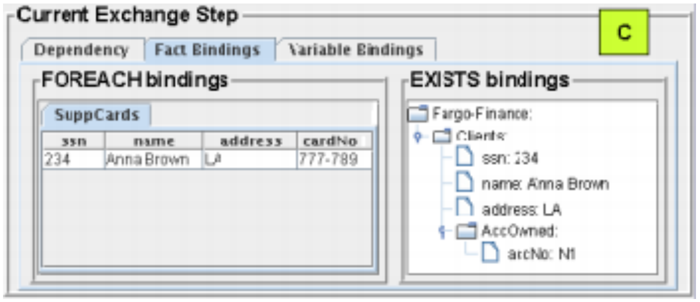
Outline

- Avoiding complexity
  - … Through selective focus
  - … Through scoring the data
  - … Through best-match querying
- Abstracting away complexity
  - … Views and meta-nodes
  - … Generating meta-nodes by clustering
  - … Visualizing meta-nodes
- Generalizing from provenance
- Discussion questions for the session
- Author presentations

Avoiding Complexity 1:

Selecting a Few Items – SPIDER [Chiticariu & Tan 06]





ssn	name	address	cardNo
234	Anna Brown	LA	777-789

```
Fargo-Finance:
  Clients:
    ssn: 234
    name: Anna Brown
    address: LA
    AccOwned:
      accNo: N1
```

- SPIDER debugs “bad” data exchange results
  - User highlights a faulty set of data items (or a schema)
  - System shows which mappings (“routes”) were used in creating it!
    - Can single-step, look at data details in separate windows
  - Not too visually complex: Number of mappings is typically small, and can only see one data item at a time

➢ But debugging can be more complex – as we’ll see in the Chiarini talk in this session

Avoiding Complexity 2: Avoiding the Graph by Ranking Data

- Our goal might be to rate data trustworthiness
  - Define a compositional scoring model, returned data with ranked annotations

  - e.g., data provenance based on semirings [Green+ 07]
    - ❖ Tuples t receive annotations ann_t from a structure called a **commutative semiring**

Relational algebra has two tuple-combining operators, $\cup$, $\bowtie$

A semiring has two operators $\oplus$, $\otimes$

Query operations derive new semiring annotations as follows:

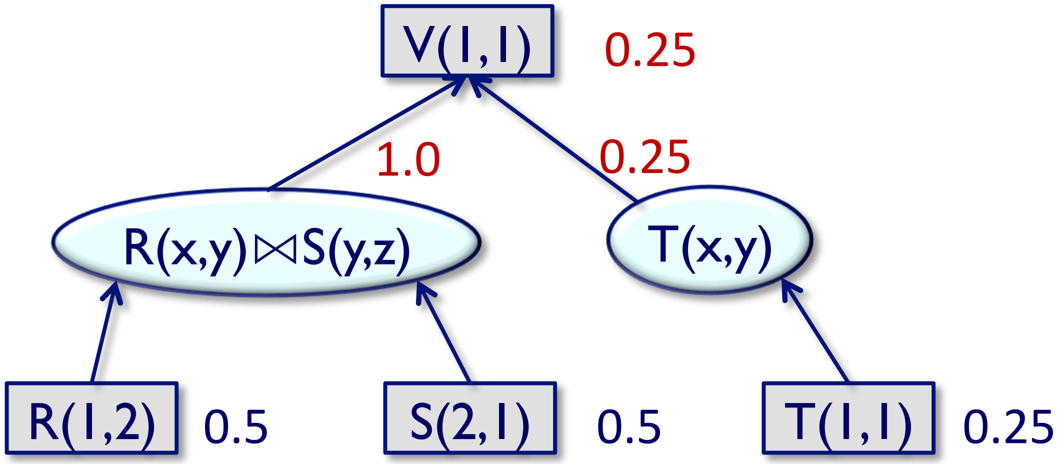
$t_a \cup t_b$: $\text{ann}_{ta} \oplus \text{ann}_{tb}$

$t_a \bowtie t_b$: $\text{ann}_{ta} \otimes \text{ann}_{tb}$

Example: Computing “Distrust” Levels

[Talukdar+08], [Karvounarakis+10]

- Suppose we know the initial quality of a source
  - Annotate with the negative log likelihood of correctness
  - Use the semiring with operators ⊕: min and ⊗:+

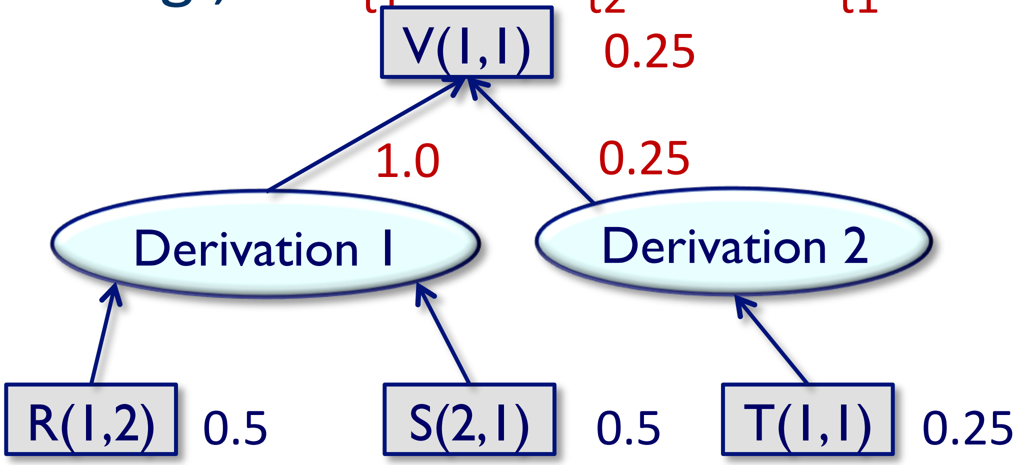


- Can even *learn* a correct ranking, given feedback over answers…

Example: Learning Rankings – & Distrust Scores

[Talukdar+08], [Karvounarakis+10]

- Suppose we get feedback in terms of a *constraint* on its score: e.g., $\text{ann}_{t1} < \text{ann}_{t2}$ or $\text{ann}_{t1} < c$



- View gets a feature vector: $\text{ann}_{R(1,2)}$ $\text{ann}_{S(2,1)}$ $\text{ann}_{T(1,1)}$
  - Derivation 1: (0.5, 0.5, 0) Derivation 2: (0, 0, 0.25)
- Find adjustment to scores using MIRA algorithm [Krammer+06]

➢ But how to generalize to alternate provenance models?

Avoiding Complexity 3:
Avoiding the Graph and Matching by Similarity

- Sometimes we want to find based on similarity:
  - the provenance of data that best matches our example
  - or the data whose provenance best matches our example

- Can query by example:
  - by finding similar graphs (Freire, this session)
  - by distance between feature vectors (Missier, this session)

➢ Are there useful application-agnostic metrics?

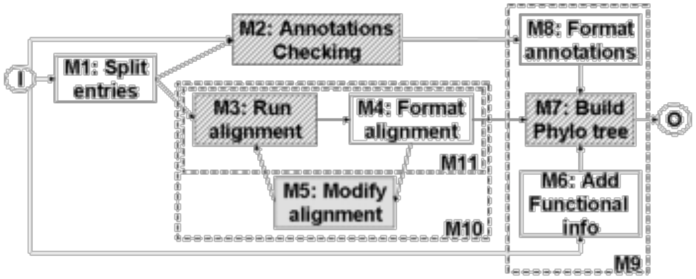
➢ Or general procedures for identifying the features or graph-matching algorithms?

Outline

- Avoiding complexity
- Abstracting away complexity
  - … Views and meta-nodes
  - … Generating meta-nodes by clustering
  - … Visualizing meta-nodes
- Generalizing from provenance
- Discussion questions for the session
- Author presentations

Approach 3: Views & Abstraction

- Many provenance models have a notion of meta-nodes or "collapsed nodes"
  - For privacy but also for visual simplicity!

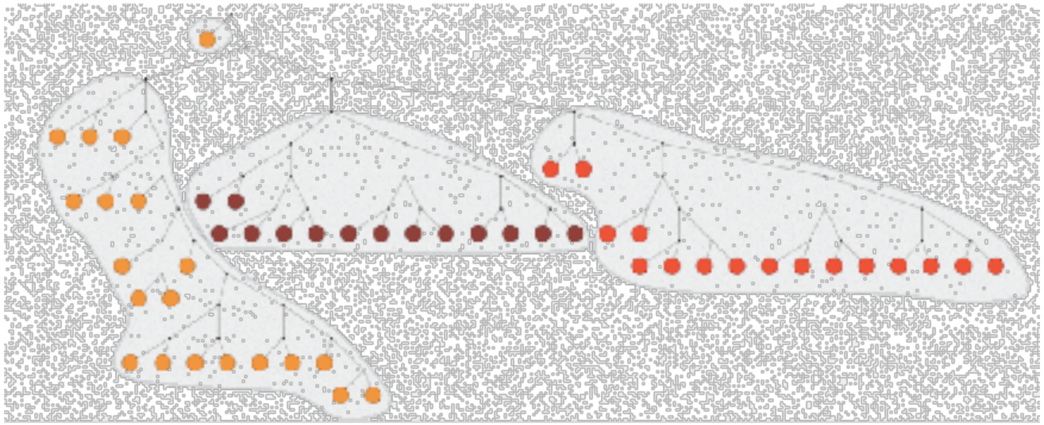


- We "collapse" sets of specified nodes
  - Various rules have been developed for preserving correctness of the abstracted provenance graph
  - ZOOM, PASS, etc.

➢ Can we automate the **specification** of what to abstract?

One Approach: Clustering & Classification

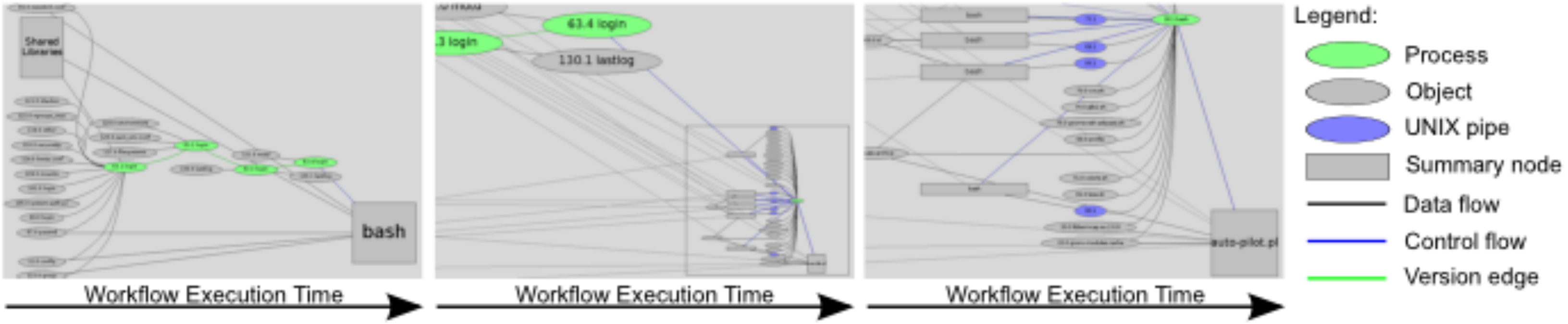
- Idea: define a metric for graph similarity, then run a clustering algorithm



- … Or, go even further and learn classifications
  - Freire talk

Navigating Abstracted Provenance

- Given provenance with collapsed nodes, how do we visualize it and navigate through it?



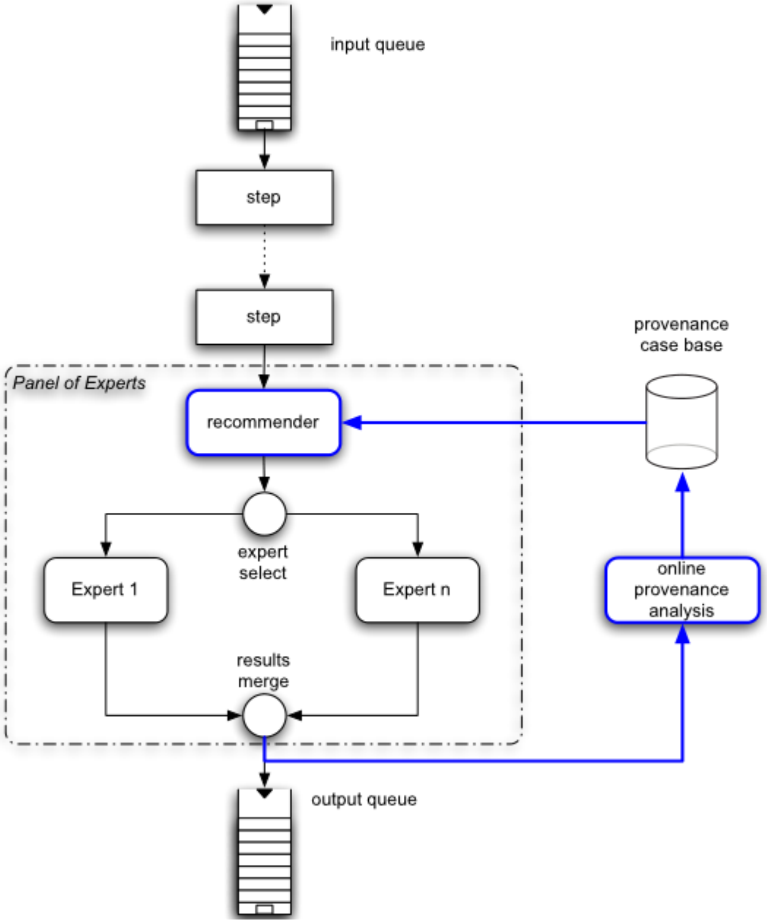
- Two papers on this from the PASS group:
  - Macko & Seltzer – navigational interface
  - Chiarini – requirements for how nodes should be visualized for debugging OS configurations

Outline

- Avoiding complexity
- Abstracting away complexity
- Generalizing from provenance
- Discussion questions for the session
- Author presentations

Beyond Visuals:
Helping Support User Activities

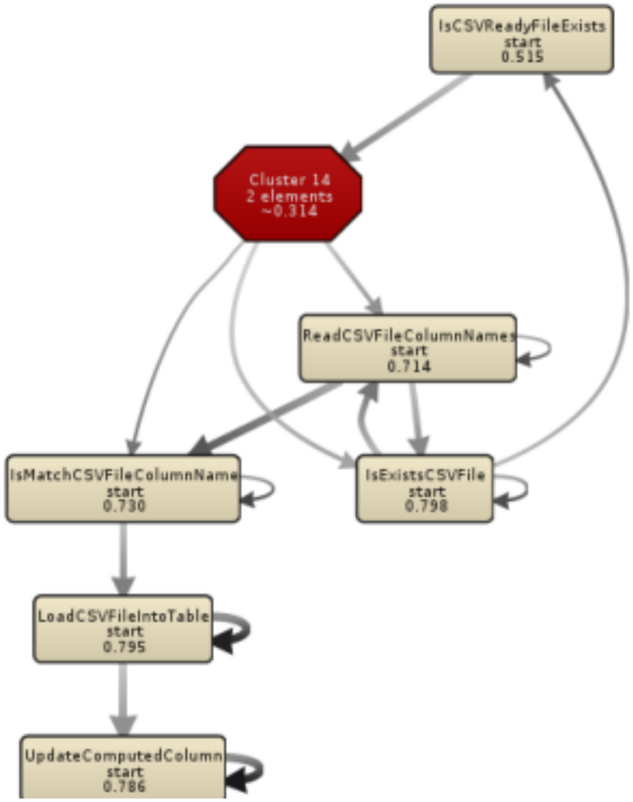
- How do we make recommendations about a workflow, using similarity scores and case-based-reasoning?
  - Missier paper



Beyond Visuals:
Mining Workflows from Provenance

- Many techniques have been developed for learning process models from event logs – “process mining”

- Zeng paper – a study of how useful different techniques are in inferring workflows from provenance



Initial Questions for the Provenance Analytics Session

- For each presenter / theme / paper:
  - How general is your solution? Does it generalize to other provenance models?
  - How far does it go towards solving the main challenge problem in the area?
  - Are there interactions with your task and the need for privacy? (lead-in to tomorrow's session)

Presentations

- Avoiding complexity
- Abstracting away complexity
  - Juliana Freire, *Clustering & Classifying Provenance, Making Recommendations*
  - Marc Chiarini, *Provenance for System Troubleshooting*
  - Peter Macko & Margo Seltzer, *Provenance Map Orbiter: Interactive Exploration of Large Provenance Graphs*
- Generalizing from provenance
  - Reng Zeng et al, *A Method to Build and Analyze Scientific Workflows from Provenance through Process Mining*
  - Paolo Missier, *Incremental Workflow Improvement Through Analysis of Its Data Provenance*