Revealing the Detailed History of Script Outputs with Hybrid Provenance Queries

Yang Cao[1], Duc Vu[2], Qiwen Wang[1], Qian Zhang[1], Priyaa Ramesh[3],Timothy McPhillips[1], Paolo Missier[3], Bertram Ludäscher[1]

[1]University of Illinois, Urbana-Champaign, [2]Department of Electrical and Computer Engineering, Univeristy of Illinois at Chicago, [3]School of Computing Science, Newcastle University, UK

Motivation

- Data- and Workflow-Provenance are crucial for **transparency** and **reproducibility** in computational and data-driven science.
- Scientific workflow systems (Kepler, Taverna, …) provide both **prospective provenance** (the workflow graph) and **retrospective provenance** (runtime observables).

Challenges

- Most computational analyses and workflows are conducted using **scripts** (Python, R, MATLAB, bash, …) rather than workflow systems.
- **Retrospective Provenance Observables**, e.g., from DataONE RunManagers (file-level), ReproZip (OS-level), or noWorkflow (Python code-level) only yield **isolated fragments** of the overall data lineage and processing history.
- **Prospective Provenance** could be used to link and contextualize fragments into a meaningful and comprehensible workflow, but **scripts alone do not reveal the underlying workflow graph**.
- Provenance (like other metadata) appears to be **rarely actionable or immediately useful** for those who are expected to provide it (provenance is "*for others*").

Approach

Simple **YesWorkflow (YW) annotations** allow users to **reveal workflow** (prospective provenance graph) **implicit in scripts**.

- **Prospective provenance queries to expose and test data dependencies** at the workflow level.
- **Hybrid provenance queries** that **situate runtime observables** (retrospective provenance) in the overall **workflow**, yielding meaningful knowledge artifacts.
- Easily **share comprehensible workflow graphs and customizable provenance reports for script runs**, along with data, code in scientific studies ("*provenance for self*").

Demo Queries

- Q1 (**prospective query**): Render prospective upstream subgraph of the YW model of the script for a given output data product D.
- Q5 (**hybrid query**): Render retrospective graph with with concrete filename for a given output data product D.

Fine-Grained Prospective Provenance

Fig 1.1 (a) YesWorkflow model of a MATLAB script of C3C4

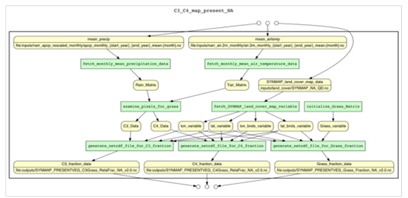

Fig 1.1 (b) Upstream figure of a MATLAB script of C3C4

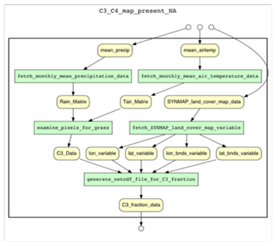

Fig 1.1 (c) Prospective queries Q4-pro of a MATLAB script of C3C4

Q2_Pro : **List the script inputs that are *upstream* of a given data product D.**
q2_pro('C3_fraction_data',mean_precip).
q2_pro('C3_fraction_data',mean_airtemp).
q2_pro('C3_fraction_data','SYNMAP_land_cover_map_data').
q2_pro('Grass_fraction_data','SYNMAP_land_cover_map_data').

Q4_Pro : List the outputs that depend on a particular script Input (*downstream*).
q4_pro(mean_airtemp,'C4_fraction_data').
q4_pro(mean_airtemp,'C3_fraction_data').

Fig 1.2 (a) YesWorkflow model of a Python script of LIGO

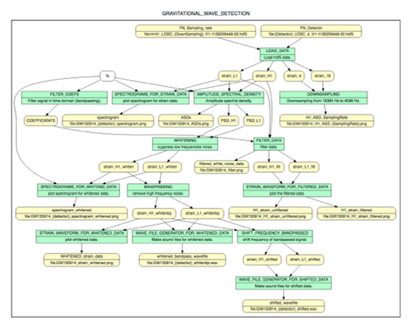

Fig 1.2 (c) Upstream figure of a Python script of LIGO

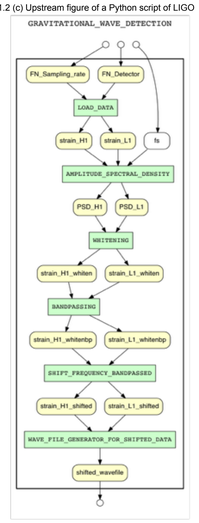

Fig 1.2 (c) Prospective queries of a Python script of LIGO

Q2_Pro : **List the script inputs that are *upstream* of a given data product D.**
q2_pro(shifted_wavefile,fs).
q2_pro(shifted_wavefile,'FN_Sampling_rate').
q2_pro(shifted_wavefile,'FN_Detector').

Q4_Pro : List the outputs that depend on a particular script Input (*downstream*).
q4_pro(fs,'ASDs').
q4_pro(fs,spectrogram).
q4_pro(fs,filtered_white_noise_data).
q4_pro(fs,'WHITENED_strain_data').
q4_pro(fs,shifted_wavefile).
q4_pro(fs,'H1_strain_filtered').
q4_pro(fs,whitened_bandpass_wavefile).
q4_pro(fs,'H1_strain_unfiltered').
q4_pro(fs,spectrogram_whitened).

Run Reconstruction

- **YW recon**
- **YW recon facts**

Coarse and Fine-Grained Observations of Runs

- Matlab Run Manager -> list of files input or output
- YesWorkflow -> list of files matching @URI annotations
- noWorkflow -> values assigned to variables

Hybrid Queries for Fine-Grained Retrospective Provenance

- **YW recon**
- **YW recon facts**
- **YW noworkflow facts**

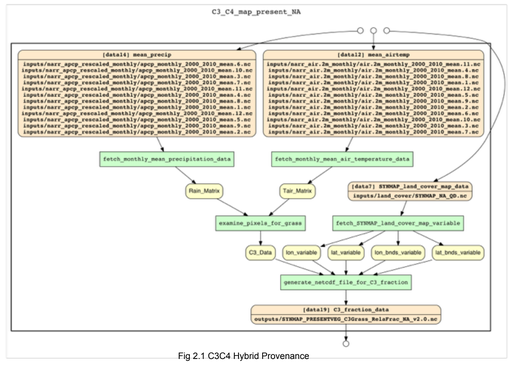

Fig 2.1 C3C4 Hybrid Provenance

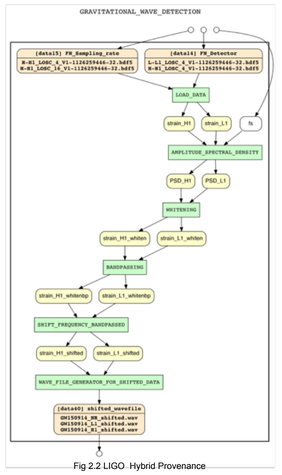

Fig 2.2 LIGO Hybrid Provenance

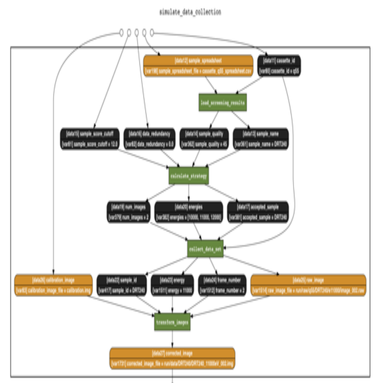

Fig 2.3 C3C4 RunManager

Conclusions and Future Work

- Provenance from script runs can be revealed graphically and made actionable (e.g., to yield customizable data lineage reports) via (1) simple YW user annotations, (2) linking runtime observables (e.g. DataONE RunManager, ReproZip, noWorkflow), and (3) sharing provenance artifacts and executable queries.
- Extend YW toolkit to support other (optional) workflow modeling constructs (e.g., simple control-flow to complement dataflow); to support graph pattern queries; to support project-level provenance.
- Evolve ProvONE to support project-level provenance and graph queries.

References

- Y Cao, D Vu, Q Wang, Q Zhang, P Ramesh, T McPhillips, P Missier, B Ludäscher (2016). DataONE AHM Provenance Demonstration: https://github.com/idaks/dataone-ahm-2016-poster
- YesWorkflow Project and Tools, https://github.com/yesworkflow-org
- T. McPhillips, T. Song, et al.(2015). YesWorkflow: A User-Oriented, Language-Independent Tool for Recovering Workflow Information from Scripts. Intl. Journal of Digital Curation 10, 298-313.
- T. McPhillips, S. Bowers, K. Belhajjame, B. Ludäscher (2015). Retrospective Provenance Without a Runtime Provenance Recorder. Workshop on the Theory and Practice of Provenance (TaPP).
- Cao, Y., Jones, C., Cuevas-Vicenttín, V., Jones, M.B., Ludäscher, B., McPhillips, T., Missier, et al., 2016, June. DataONE: A Data Federation with Provenance Support. Intl. Provenance and Annotation Workshop (IPAW). Springer.
- Pimentel, J.F., Dey, S., McPhillips, T., Belhajjame, K., Koop, D., Murta, L., Braganholo, V. and Ludäscher, B., 2016, June. Yin & Yang: demonstrating complementary provenance from noWorkflow & YesWorkflow. Intl. Provenance and Annotation Workshop (IPAW). Springer.

Acknowledgments. Supported by NSF awards ACI-1430508 and NSF ABI-1262458.

School of Information Sciences
The iSchool at Illinois
ILLINOIS
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN