A METHOD TO BUILD AND ANALYZE SCIENTIFIC WORKFLOWS FROM PROVENANCE THROUGH PROCESS MINING

Reng Zeng, Xudong He

Florida International University
Miami, Florida, USA

Jiafei Li

JiLin University
China

Zheng Liu, W.M.P. van der Aalst

Eindhoven University of Technology
The Netherlands

MOTIVATION

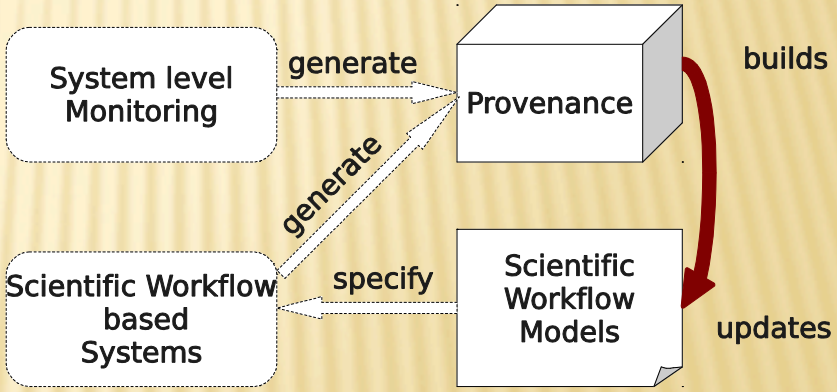
- Before creating scientific workflows, the provenance can only be captured from provenance enabled applications.
- It is often very hard to manually create and maintain scientific workflows.
- Can we leverage existing provenance to build scientific workflows automatically?

MOTIVATION (CONT)

- Process mining has become an active research area in the past decade,
- Process mining synthesizes a process model from event logs,
- We aim to automatically generate a scientific workflow model from provenance using established process mining techniques
  - Offers an effective approach for creating an initial scientific workflow model,
  - Facilitates analysis techniques such as simulation and verification for detecting potential scientific workflow design problems,
  - Helps to discover hidden dependencies among different scientific workflows,
  - Supports automated synthesis of existing scientific workflows

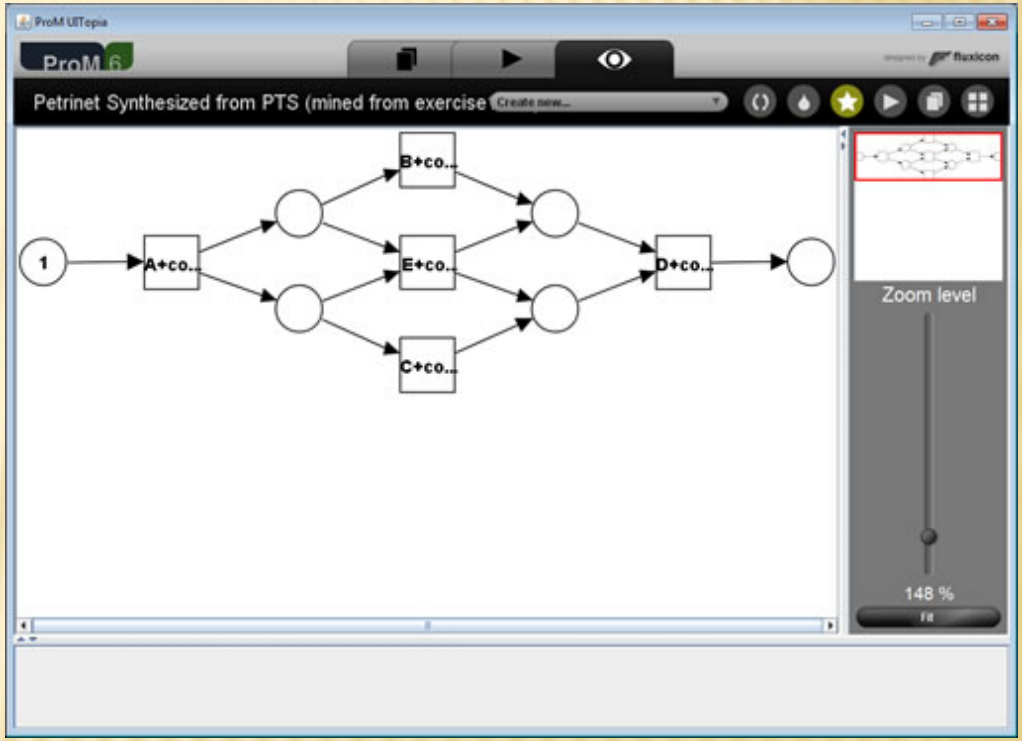
A METHOD TO BUILD AND ANALYZE SCIENTIFIC WORKFLOWS FROM PROVENANCE

- Issues when applying process mining in the context of scientific workflow
  - Control flow mining
    - In this paper we focus on control flow mining
  - Data dependency
  - Incremental mining



TOOLS

- ProM is a generic open-source framework for implementing process mining tools in a standard environment.
- The ProM framework accepts input logs in the XES or MXML format.
- The ProM framework has plug-ins for process *mining*, *analysis*, *monitoring* and *conversion*.
  - Conversion from event logs in relational databases to XES or MXML.
  - We have converted provenance in Taverna and Kepler to XES / MXML.



http://prom.sourceforge.net/

RESULTS OF PROCESS DISCOVERY ALGORITHMS

	Description	Result
Fuzzy Miner	Provides a **zoomable view** of scientific workflows by controlling significance cutoff to show tasks at different importance levels.	Under certain significance cutoff, the fuzzy miner successfully gives the changed part and unchanged part. The fuzzy miner gets most dependency correctly in the original scientific workflows, but includes some non-existent dependency.
Alpha Miner	Provides a view of **direct succession** between tasks in provenance.	Assuming the completeness of direct succession, the alpha miner fails to give a view close to the original scientific workflow.
Genetic Miner	Provides a view of **frequency** for both tasks and succession between tasks, and discovers all common control-flow structures assuming the existence of noises.	The genetic miner gets a good view of structures and frequencies, yet gives some wrong dependencies which do not exist in both the original scientific workflows and the results of the fuzzy miner.
Heuristic Miner	Provides a view of scientific workflows by considering **long distance dependency**.	The heuristic miner gives long distance dependency successfully, but gives too much dependency for some tasks such as ReadCSVFileColumnNames.

CONCLUSION

- A method using process mining to build and analyze scientific workflows
- The method can be applied to provenance data in many different forms
  - it is quite straight forward to translate the provenance to XES format acceptable to process mining tools