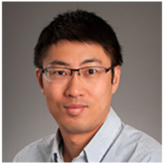
SciLedger: A Scientific Workflow Provenance and Data Sharing Platform

IEEE Conference on Collaboration and Internet Computing '22



Hamilton Hardy   Reagan Hoopes   Min Long   Gaby Dagher

December 14, 2022

UtahStateUniversity

BOISE STATE UNIVERSITY

Table of Contents

1. Introduction
2. Related Work
3. Background
4. Architecture
5. Experimental Evaluation
6. Conclusion

INTRODUCTION

Motivation

- Scientific researchers collaborating from different locations
- Lack of way to ensure research integrity
  - 8.3% committed falsification/fabrication at least once from 2017-2020 [10]
- Increased requirements for data sharing from governmental and private funders [11]
- Flexibility within science
  - 60% of pre-established workflows concluded with null results [7]
    - Invalidation

Challenges

- Balancing contradictory needs of scientific research
  - Integrity limits flexibility
  - Public systems promote accessibility, but limit user privacy
  - Blockchain requires off-chain storage for scientific data which introduces security concerns

The Problem We Address

Scientific researcher’s needs for a system that:

- Is specific to scientific workflow provenance
- Allows for data sharing
- Supports complex processes such as branching and merging
- Provides a sufficient level of user privacy

Contributions

- The SciLedger system
- Public, blockchain-based platform that supports open-access data sharing and complex workflow operations
- Invalidation mechanism
- Implementation and experimental evaluation

RELATED WORK

Scientific Workflow Management Systems

- Kepler [2]
- Taverna [3]
- Galaxy[1]
- KNIME[4]
- Pegasus[5]
- Key Features
  - Locally Maintained Storage
  - Scientific Field Specific

Generic Blockchain Solutions

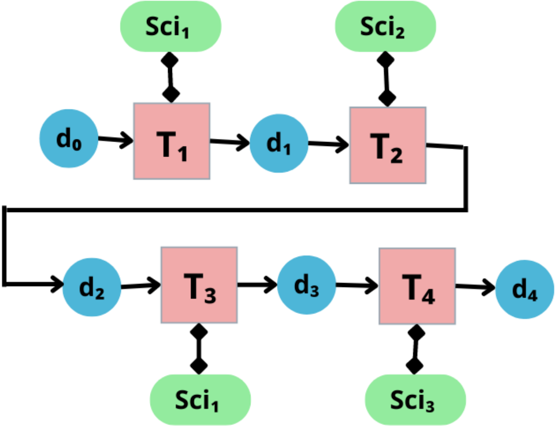
- LineageChain [13]
  - Event Listeners for Data Modification
- BlockCloud [16][15]
  - Network Consesnus by Staking cloud storage
- ProvHL [8]
  - Access Controls for Private Data
- Sifah *et al.* [14]
  - Data Ownership Permissions
- Key Features
  - Private Blockchains
  - Generic Solutions

Scientific Workflow Blockchain Solutions

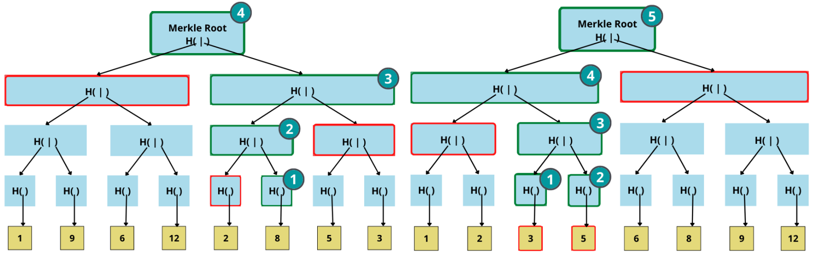
- SmartProvenance [12]
  - Threshold Based Voting Smart Contracts
- Bloxberg [17]
  - Unique Provenance Model
- SciChain [6]
  - Optimized for High Performance Computing
- SciBlock [9]
  - Time Stamp Invalidation Mechanism
- Key Features
  - Private Blockchains
  - Limited in Features

BACKGROUND

Scientific Workflows and Provenance



Merkle Trees



(a) Proving membership of data point 8

(b) Proving non-membership of data point 4

ARCHITECTURE

Overview

- Scientific Provenance Collection
- Complex Multi-Workflow System
- Dependency based Invalidation
- Two Tree Merkle Verification

Scientific Provenance Collection (Cont.)

Provenance Record

Field	**Description**
Task ID	The task's assigned identifier value
Workflow ID	The workflow's assigned identifier value
User ID	Public key belonging to the task performer
Submission Time	Submission time to the quorum
Input Data	Hash pointer to data before modification
Output Data	Hash pointer to data after modification
Valid Merkle Root	Top hash for valid Merkle tree
Invalid Merkle Root	Top hash for invalid Merkle tree
Other	Extra fields custom provenance values

Complex Multi-Workflow System

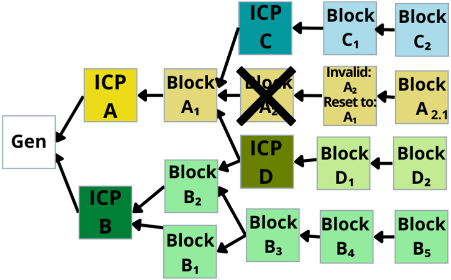


Figure: Sample SciLedger blockchain visualized as Workflows

- Workflow Design
  - Merging
  - Branching
  - Multiple Workflows
- Inception Block
  - Predefined Workflow Design
  - Public Keys of Authorized Users

Dependency Based Invalidation

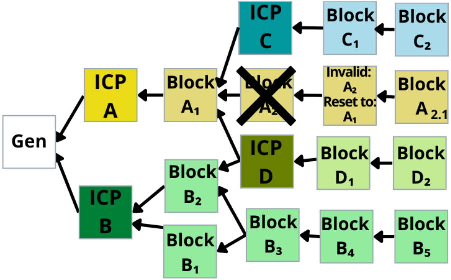


Figure: Sample SciLedger blockchain visualized as Workflows

- Invalidation Block
  - Added to End of Workflow
  - Updates Merkle Trees

EXPERIMENTAL EVALUATION

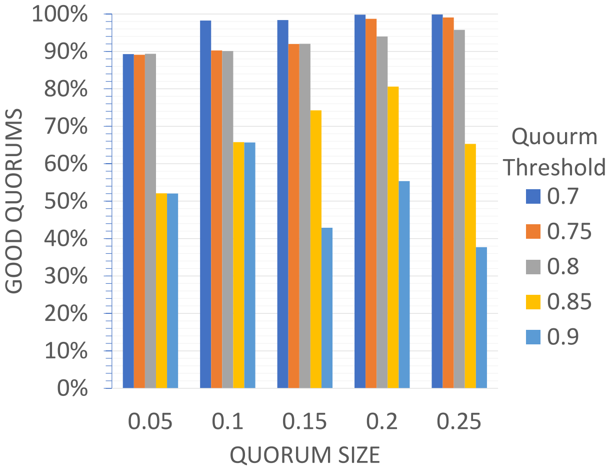
Implementation

- Workflow Generator
    - LoremIpsum data
    - Branching and Merging Complexity
    - Valid and Invalid Merkle Trees
- Block Constructor
    - Provenance Record Construction
    - Transaction Header
- Blockchain
    - Node Consensus

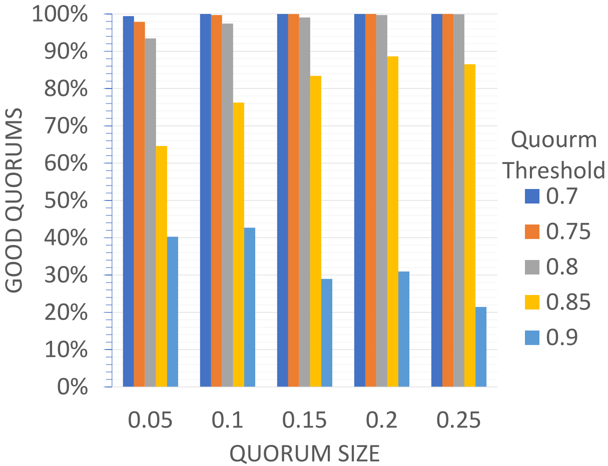
Quorum Experiment Setup

- Malicious Activity in Scientific Research
  - *8.3%* Maliciously Manipulated Data [10]
  - Fix Expected Malicious actors in the Network to be less than 12%.
- Parameters
  - Network Size (Scalability)
  - Quorum Size relative to the Network
  - Quorum Consensus Threshold

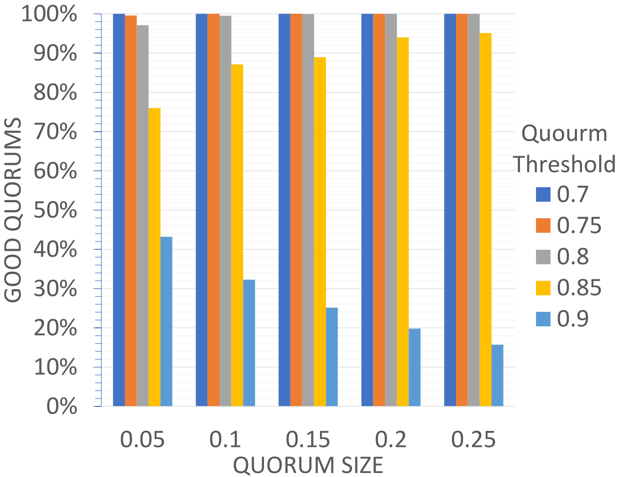
Quorum Parameter Experiment Results



Quorum Parameter Experiment Results



Quorum Parameter Experiment Results



Additional Experiments in the Works

- Block Upload Speed
- Block Verification Transaction Analysis
  - Existence and Validity of Block
    - Valid Merkle Tree of Last Block Added
    - Valid Merkle Tree of the Block in the chain and absent from Invalid Merkle Tree of Last Block
  - Existence of Block
    - Valid Merkle Tree of the Block in the chain
    - Brute Force that recurses over chain until Block found
  - Non Existence of a Block
    - Absence from Valid Invalid Merkle Tree of Last Block
    - Brute Force that recurses over all blockchain until block is not found

CONCLUSION

Summary

- We propose SciLedger: a blockchain-based solution that supports open-access data sharing for scientific workflow provenance and complex workflow operations
- We propose novel invalidation and merkle tree verification methods that allows researchers to modify workflows in a way that minimizes unnecessary repetition.
- SciLedger's implementation shows such a system is possible
- Experimentation proves our system's scalability and efficiency

Future Work

- Differential Data Privacy
- Consensus Mechanisms
- Scientific Data Verification in Blockchain
- Activity Privacy

Conference

The 8th IEEE International Conference on Collaboration and Internet Computing

December 14-16, 2022, Las Vegas, Nevada, USA (tentative)

Questions?

[1] Galaxy community hub.

[2] The kepler project.

[3] Taverna - apache incubator.

[4] Aug 2022.

[5] Pegasus, Apr 2022.

[6] Abdullah Al-Mamun, Feng Yan, and Dongfang Zhao. Scichain: Blockchain-enabled lightweight and efficient data provenance for reproducible scientific computing. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 1853–1858, 2021.

[7] Christopher Allen and David M. Mehler. Open science challenges, benefits and tips in early career and beyond. *PLOS Biology*, 17(5):1–14, May 2019.

[8] Andrey Demichev, Alexander Kryukov, and Nikolai Prikhodko. The approach to managing provenance metadata and data access rights in distributed storage using the hyperledger blockchain platform. In *2018 Ivannikov Ispras Open Conference (ISPRAS)*, pages 131–136, 2018.

[9] Dinuni Fernando, Siddharth Kulshrestha, J. Dinal Herath, Nitin Mahadik, Yanzhe Ma, Changxin Bai, Ping Yang, Guanhua Yan, and Shiyong Lu. Sciblock: A blockchain-based tamper-proof non-repudiable storage for scientific workflow provenance. In *2019 IEEE 5th International Conference on Collaboration and Internet Computing (CIC)*, pages 81–90, 2019.

[10] Gowri Gopalakrishna, Gerben ter Riet, Gerko Vink, Ineke Stoop, Jelte M. Wicherts, and Lex M. Bouter. Prevalence

of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the netherlands. *PLOS ONE*, 17:1–16, 02 2022.

[11] G. Popkin. Setting your data free. *Nature*, 569:445–447, 2019.

[12] Aravind Ramachandran and Murat Kantarcioglu. Smartprovenance: A distributed, blockchain based dataprovenance system. In *Proceedings of the Eighth ACM Conference on Data and Application Security and Privacy*, CODASPY ’18, page 35–42, 2018.

[13] Pingcheng Ruan, Gang Chen, Tien Tuan Anh Dinh, Qian Lin, Beng Chin Ooi, and Meihui Zhang. Fine-grained, secure and efficient data provenance on blockchain systems. *Proc. VLDB Endow.*, 12(9):975–988, May 2019.

[14] Emmanuel Boateng Sifah, Qi Xia, Kwame Opuni-Boachie Obour Agyekum, Hu Xia, Abla Smahi, and Jianbin Gao. A blockchain approach to ensuring provenance to outsourced cloud data in a sharing ecosystem. *IEEE Systems Journal*, 16(1):1673–1684, 2022.

[15] Deepak Tosh, Sachin Shetty, Xueping Liang, Charles Kamhoua, and Laurent L. Njilla. Data provenance in the cloud: A blockchain-based approach. *IEEE Consumer Electronics Magazine*, 8(4):38–44, 2019.

[16] Deepak K. Tosh, Sachin Shetty, Xueping Liang, Charles Kamhoua, and Laurent Njilla. Consensus protocols for blockchain-based data provenance: Challenges and opportunities. In *2017 IEEE 8th Annual Ubiquitous*

Computing, Electronics and Mobile Communication Conference (UEMCON), pages 469–474, 2017.

[17] Kevin Wittek, Neslihan Wittek, James Lawton, Iryna Dohndorf, Alexander Weinert, and Andrei Ionita. A blockchain-based approach to provenance and reproducibility in research workflows. In *2021 IEEE International Conference on Blockchain and Cryptocurrency (ICBC)*, pages 1–6, 2021.