Modeling Data Integrity Threats for Scientific Workflows Using OSCRP and MITRE ATT&CK®

Ishan Abhinit, Emily K. Adams, Brian Chase
CACR, Indiana University

Anirban Mandal, Yufeng Xin
RENCI, UNC - Chapel Hill

Karan Vahi, Mats Rynge, Ewa Deelman
ISI, University of Southern California

***Abstract*—Guaranteeing the data integrity of scientific workflows and their associated data products, in the face of non-malicious and malicious threats, is of paramount importance for the validity and credibility of scientific research. In this work, we describe how we can leverage two popular cybersecurity classification frameworks - OSCRP and MITRE ATT&CK®, to systematically model threats to the integrity of scientific workflows and data in a research setting. We enumerate non-malicious and malicious threats to the integrity of scientific workflows, and present the relevant assets, concerns, avenues of attacks and impact of the threats in typical scientific workflow execution scenarios.**

***Index Terms*—cybersecurity, data integrity, scientific workflow, threat model, OSCRP, MITRE ATT&CK**

I. Introduction

With the rise in scale and complexity of scientific workflows, it is extremely important to assure the the integrity of the scientific data, as they are processed and transmitted on distributed infrastructures. When data integrity is not preserved, computational workflows can fail and result in increased computational cost due to reruns, or worse, results can be corrupted compromising the scientific outcomes [1]. Recent works, e.g. [2] are initial efforts to ensure that science workflows and data transfers are guarded against data integrity errors that might arise in complex distributed systems. However, these works have not addressed the diagnosis and pinpointing of the root cause of data integrity errors. Any such analysis will need to incorporate the threat models for data corruption, both non-malicious and malicious threats. In this work, we leverage two existing popular cybersecurity frameworks, Open Science Cyber Risk Profile (OSCRP) [3] and MITRE ATT&CK knowledge base [4] to build threat models for scientific workflows to pinpoint the root cause of unintentional (*i.e.* non-malicious) integrity errors [5], and uncover adversarial tactics and techniques (*i.e.* malicious) [6] that threaten data integrity within scientific workflows.

II. Non-Malicious Data Integrity Threat Model Using OSCRP

Open Science Cyber Risk Profile (OSCRP) is designed to assess cybersecurity risks related to open science projects.Open science research often has unique cybersecurity concerns and OSCRP provides a catalog of typical scientific research assets and the associated risks to research activities. Steps involved in the OSCRP risk profiling process are: 1) Identify the stakeholders; 2) Create an asset inventory; 3) Examine the concerns, consequences and avenues of attack for each mission critical science asset; and 4) for each relevant concern, identify the vectors/methods of attack that could cause the concern to be realized.

A. Categorizing Scientific Workflow Assets using OSCRP

We first translated the OSCRP asset classifications into vernacular suited explicitly for scientific workflows. The resultant asset classification, shown in Figure 1 with their associated OSCRP asset class (in parenthesis) are: I. Transient workflow (”Internal” data); II. Data products (”Public” data); III. Metadata (”Accounting” data), IV. Researcher system (”Desktop”); V. Workflow management system (”Workflow”); VI. Computational systems (”Servers”); VII. Data storage systems (”File storage”); VIII. Network systems (”Networks”). By subsequently overlaying the scientific workflow as enacted by a representative workflow management system, Pegasus [7], with our OSCRP asset classification, indicated with red markers in Figure 1, it enabled us to enumerate possible sources of data integrity errors. Thus, we were able to derive non-malicious threats and conduct an impact analysis to the integrity of the scientific workflow.

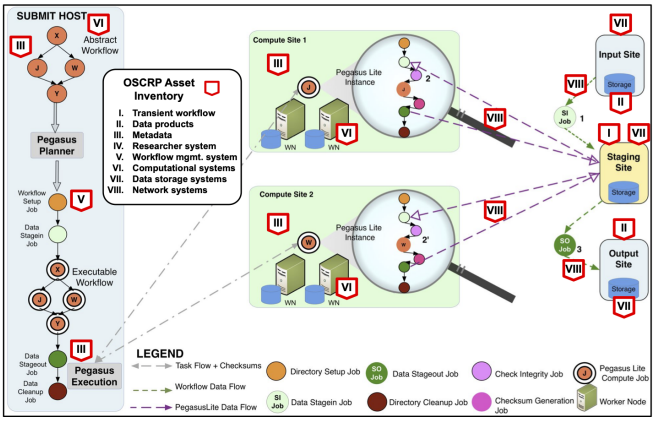

Fig. 1. OSCRP Asset Mapping diagram for Workflows (Adapted).

B. Findings from Non-Malicious Data Integrity Threat Model

We started our non-malicious scientific workflow threat modeling by looking at each asset to determine how and where the data integrity can be degraded. We then identified

concerns (*i.e.* a negative change to an asset that impacts a research activity) for these asset. Finally, we determined the vectors or methods of integrity degradation to complete our non-malicious integrity analysis. Table I exhibits two examples from our first threat model itemizing potential non-malicious impacts to the integrity of scientific workflows.

TABLE I
TWO EXCERPTS FROM "DETAILED ANALYSIS OF ASSETS TO CONCERNS AND AVENUES OF ATTACK" TABLE (ADAPTED)

OSCRP Asset	Concern(s)	Consequence/Impact	Integrity Degradation
I	Corrupted data, incorrect data, or lost data	Workflow producing incorrect/invalid results	Issues with data processing, issue with sensor equipment
V	Lost or incorrect process	Workflow producing incorrect/invalid results	Issues with storage

III. MALICIOUS THREAT MODEL USING MITRE ATT&CK

The MITRE ATT&CK Enterprise knowledge base is used as a foundation for the development of targeted cybersecurity threat models and methodologies. We used the MITRE ATT&CK framework of ATT&CK Tactics and Techniques to conduct a scoped impact analysis enumerating malicious attacks against data integrity within scientific workflows.

A. Defining Data Assets using OSCRP and Attack Type

Prior to leveraging ATT&CK adversarial Tactics and Techniques to develop a threat model for malicious attacks, we scoped our focus to only address the integrity of data within the workflows. Thus, the following subset of OSCRP asset classifications specifically involving data were carried over from the non-malicious threat model for this analysis: I. Transient Data; II. Data Products; and III. Metadata.

We further scoped our malicious threat model by defining attack types that explicitly impact scientific data integrity within scientific workflows: Direct Attacks, defined as malicious activity directly targeting research data with the goal to impact data integrity (*e.g.* delete, alter); Indirect Attacks, defined as malicious activity targeting an asset or process that interfaces with data. *(e.g.* altering scientific software), but did not include General Attacks, defined as malicious activity that impacts the security of assets, functions, and personnel supporting research activities.

B. Findings from Malicious Data Integrity Threat Model

The analysis culminated by determining which MITRE ATT&CK Tactics and Techniques were relevant to malicious attacks against scientific data integrity per OSCRP-derived asset class and attack type. Of the thirteen MITRE ATT&CK Tactics, two ATT&CK Tactics, with a subset of seven ATT&CK Techniques each, were identified to precipitate malicious activity, which *directly* or *indirectly* impacts the integrity of scientific data.

- Impact [TA0040] - The adversary is trying to manipulate, interrupt, or destroy systems and data.
- Execution [TA0002] - The adversary is trying to run malicious code.

Table II contains two examples of ATT&CK Techniques that we determined directly or indirectly impacted data integrity (*i.e.* Data Integrity Concern, and Scientific Workflow Impact).

TABLE II
TWO EXCERPTS FROM ATT&CK TACTIC IMPACT ANALYSIS (ADAPTED)

ATT&CK Tactic & Technique	OSCRP Asset	Attack Type	Data Integrity Concern (An adversary may...)	Scientific Workflow Impact
[Impact Tactic] Data Destruction T1485	I II III	Direct	Destroy data and files. Render stored data irrecoverable by forensic techniques.	Scientific workflow cannot be initiated or executed. Data lost.
[Execution Tactic] Inter-Process Communication T1559	I II III	Indirect	Abuse inter-process communication (IPC) mechanisms for local code or command execution	Scientific workflow cannot be initiated, executed, is degraded, is compromised and/or producing invalid results.

IV. APPLICATION & FUTURE WORK

By systematically identifying tactics and techniques of malicious threats to data within scientific workflows, data stewards will be equipped to proactively design scientific workflows to ensure data integrity, better understand how malicious attacks against data integrity are executed, and proactively build detections and defenses to protect scientific data.

In the future, we will leverage the analysis derived from our two threat models for pinpointing root causes of workflow data integrity errors, by informing Machine Learning (ML) based analysis models to more closely align to threats observed in real systems and workflow execution scenarios. While one can inject errors based on anecdotal evidence during ML model development, injecting errors guided by malicious and non-malicious threat models should be a far superior technique, grounded in well-established cybersecurity frameworks.

ACKNOWLEDGMENT

This work is funded by NSF award OAC-1839900.

REFERENCES

[1] S. Peisert, "Security in high-performance computing environments," https://cacm.acm.org/magazines/2017/9/220422-security-in-high-performance-computing-environments/fulltext.

[2] M. Rynge, K. Vahi, E. Deelman, A. Mandal, I. Baldin, O. Bhide, R. Heiland, V. Welch, R. Hill, W. L. Poehlman, and F. A. Feltus, "Integrity protection for scientific workflow data: Motivation and initial experiences," in *Proceedings of the Practice and Experience in Advanced Research Computing on Rise of the Machines (Learning)*, ser. PEARC '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3332186.3332222

[3] "Open Science Cyber Risk Profile," https://trustedci.github.io/OSCRP.

[4] "MITRE ATT&CK knowledge base," https://attack.mitre.org.

[5] I. Abhinit and V. Welch, "Data Integrity Threat Model using Open Science Cyber Risk Profile," https://hdl.handle.net/2022/27980.

[6] E. K. Adams, "Data Integrity Threat Model using MITRE ATT&CK®," https://hdl.handle.net/2022/28045.

[7] E. Deelman, K. Vahi, G. Juve, M. Rynge, S. Callaghan, P. J. Maechling, R. Mayani, W. Chen, R. Ferreira da Silva, M. Livny, and K. Wenger, "Pegasus: a workflow management system for science automation," *Future Generation Computer Systems*, vol. 46, pp. 17–35, 2015.