JUMPSEC

Knowledge & Data Engineering Systems

University of Glasgow

OSPtrack: A Labelled Dataset Targeting Simulated Execution of Open-Source Software

Zhuoran Tan, **Christos Anagnostopoulos, Jeremy Singer**
School of Computing Science, University of Glasgow, UK

OSPtrack Dataset

Challenges

- **Lack of Source Code Access:**
  - Current solutions assume plaintext access, **but** in practice, source code of third-party software is often unavailable.
- **No Runtime Datasets:**
  - There is lack of datasets capturing the runtime behaviour of malicious packages or libraries.
- **Limited Labelled Data for Threat Detection:**
  - Existing datasets do not support real-time or runtime threat detection in third-party packages.

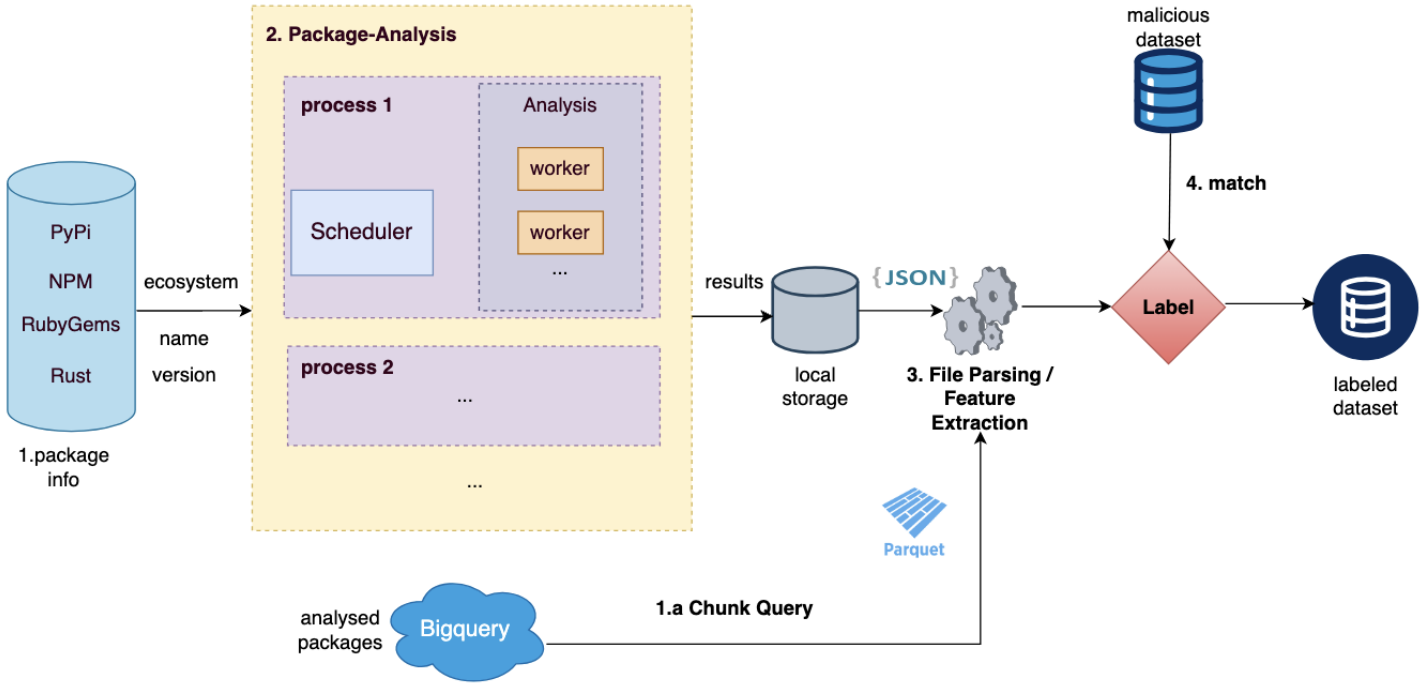
OSPtrack Dataset

Our Approach

- **Sandbox Execution:**
  Run each library/package once in a sandbox environment using *package-analysis* tools.
- **Dynamic Feature Collection:**
  Monitor and extract runtime behaviors and features during execution.
- **Malicious Dataset Sourcing:**
  Collect known malicious packages from **OpenSSF**
- **Feature Engineering:**
  Extract key features and reduce noise to improve threat detection accuracy

Data Simulation & Extraction

Knowledge & Data Engineering Systems

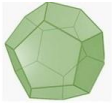


Data Structure and Features

Five Ecosystems:	Metadata:	Features:

• npm
• pypi
• crates.io
• nuget
• packagist
	• ecosystem
	• package_name
	• version
	• features
		• Two sections: import and install
		• Sub features:
		  • File ---> file-related activities
		  • Sockets ---> socket operations
		  • Commands ---> execution of system commands
		  • DNS ---> DNS queries

Dataset Size and Distribution



Knowledge & Data Engineering Systems

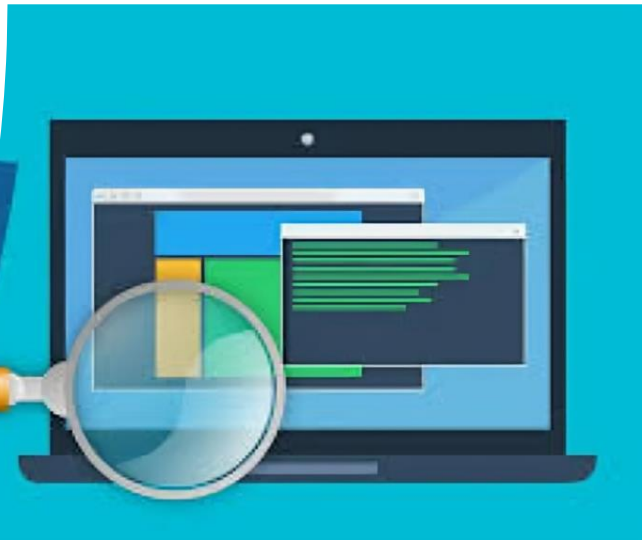
- 9,461 package instances
- 1,962 are malicious

TABLE II
PACKAGE COUNTS BY ECOSYSTEM, PACKAGE COUNT, LABEL, AND SUB-LABEL.

Ecosystem	Count	Label	Sub_Label
crates.io	1205	0	na
	1	1	na
packagist	265	0	na
	-	-	-
rubygems	61	0	na
	269	1	na
	8	1	C2
pypi	1323	0	na
	812	1	na
	38	1	C2
	2	1	command exec
npm	4645	0	na
	800	1	na
	18	1	C2
	11	1	root shell
	2	1	command exec

Research Opportunities

- Running vulnerability detection
- malicious software classification
- Threat hunting
- Differential analysis of vulnerabilities in diverse ecosystem



Resources

Knowledge & Data Engineering Systems

Takeway:

- Dynamic features
- Multiple ecosystems
- Creditable labels
- Software supply chain security

→ **Data:**
https://zenodo.org/records/14680781

→ **Code:**
https://github.com/Wapiti08/OSPTrack