

Introducing a Comprehensive, Continuous, and Collaborative Survey of Intrusion Detection Datasets

Philipp Bönninghausen¹

Rafael Uetz¹

Martin Henze^{1,3}



fkie-cad/comidds

Motivation

Intrusion detection

- Highly active research field
- Novel methods require realistic evaluation data
- Thus high demand for public datasets

Intrusion detection datasets and surveys thereof

- Multitude of datasets exist
- Researchers struggle to find appropriate datasets and understand their limitations
- Surveys quickly become outdated and are either incomprehensive or superficial

Addressing survey shortcomings with COMIDDS

- Idea: Survey as website backed by GitHub repo
- Allows for continuous extensions, corrections, and contributions
- Also enables change tracking and automatic processing



Quick Introduction to COMIDDS

A comprehensive, continuous, and collaborative intrusion detection datasets survey

Goal: Aid researchers in finding appropriate datasets and understand their limitations

- Comprehensive: High dataset coverage within scope, thorough information on each
- Continuous: We keep adding new datasets and improving existing entries
- Collaborative: Contributions welcome! – Also check your own dataset descriptions 😊

Scope: Enterprise networks

- Clients and servers running Windows, Linux etc.
- Network appliances (routers, switches, firewalls)
- Common applications and services (e.g., web, mail, directory)

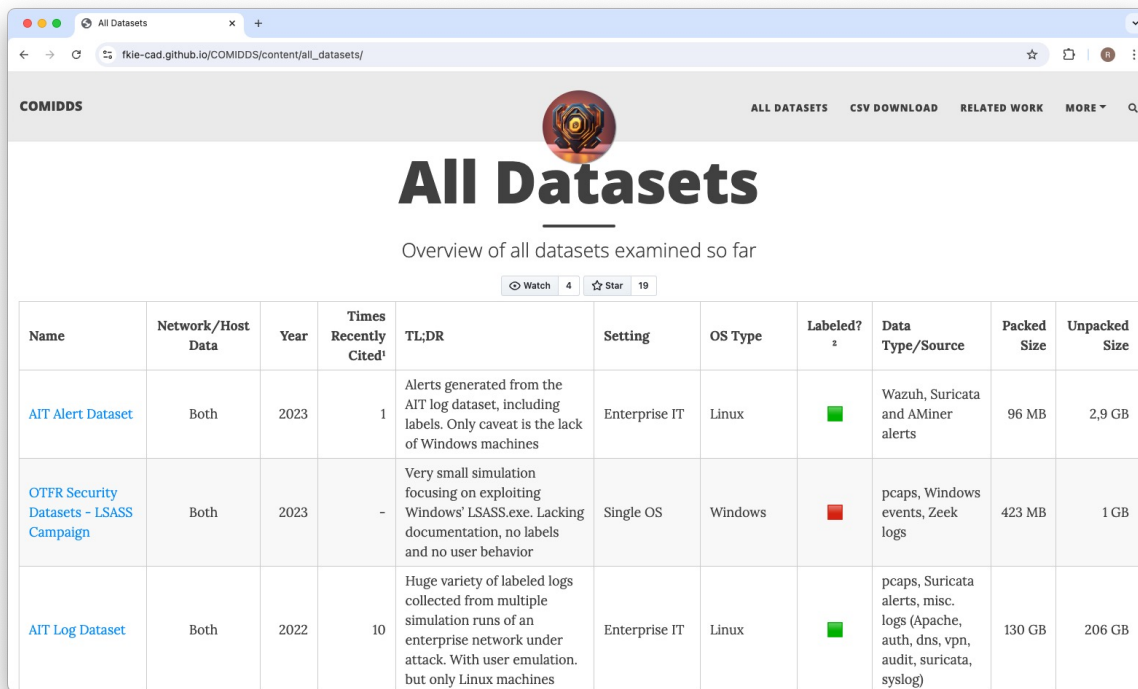
COMIDDS features

- Overview table of all reviewed datasets
- Detailed page for each dataset with info on environment, activity, files, references
- CSV, visualizations, related work, contribution guide






Live Demo

<https://fkie-cad.github.io/COMIDDS/>



The screenshot shows a web browser displaying the COMIDDS website. The page title is "All Datasets" and the subtitle is "Overview of all datasets examined so far". There are navigation links for "ALL DATASETS", "CSV DOWNLOAD", "RELATED WORK", and "MORE". A search icon is also present. Below the header, there is a table with 11 columns: Name, Network/Host Data, Year, Times Recently Cited, TL;DR, Setting, OS Type, Labeled?, Data Type/Source, Packed Size, and Unpacked Size. The table contains three rows of dataset information.

Name	Network/Host Data	Year	Times Recently Cited	TL;DR	Setting	OS Type	Labeled?	Data Type/Source	Packed Size	Unpacked Size
AIT Alert Dataset	Both	2023	1	Alerts generated from the AIT log dataset, including labels. Only caveat is the lack of Windows machines	Enterprise IT	Linux		Wazuh, Suricata and AMiner alerts	96 MB	2,9 GB
OTFR Security Datasets - LSASS Campaign	Both	2023	-	Very small simulation focusing on exploiting Windows' LSASS.exe. Lacking documentation, no labels and no user behavior	Single OS	Windows		pcaps, Windows events, Zeek logs	423 MB	1 GB
AIT Log Dataset	Both	2022	10	Huge variety of labeled logs collected from multiple simulation runs of an enterprise network under attack. With user emulation, but only Linux machines	Enterprise IT	Linux		pcaps, Suricata alerts, misc. logs (Apache, auth, dns, vpn, audit, suricata, syslog)	130 GB	206 GB

Findings and Conclusion

Findings

- Many datasets have significant limitations or deficiencies – always question fitness for purpose!
- There is a trend towards host-based logs (as compared to netflows and pcaps)
- Datasets with realistic benign behavior (i.e., performed by humans in a production environment) hardly exist

Conclusion

We hope that COMIDDS gains acceptance as a reference survey for intrusion detection datasets!



[fkie-cad/comidds](https://github.com/fkie-cad/comidds)



rafael.uetz@fkie.fraunhofer.de



[ru37z](https://twitter.com/ru37z)