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

Philipp Bönninghausen
Rafael Uetz
philipp.boenninghausen@fkie.fraunhofer.de
rafael.uetz@fkie.fraunhofer.de
Fraunhofer FKIE
Wachtberg, Germany

Martin Henze
henze@spice.rwth-aachen.de
RWTH Aachen University
Aachen, Germany
Fraunhofer FKIE
Wachtberg, Germany

ABSTRACT

Researchers in the highly active field of intrusion detection largely rely on public datasets for their experimental evaluations. However, the large number of existing datasets, the discovery of previously unknown flaws therein, and the frequent publication of new datasets make it hard to select suitable options and sufficiently understand their respective limitations. Hence, there is a great risk of drawing invalid conclusions from experimental results with respect to detection performance of novel methods in the real world. While there exist various surveys on intrusion detection datasets, they have deficiencies in providing researchers with a profound decision basis since they lack comprehensiveness, actionable details, and up-to-dateness. In this paper, we present Comidds, an ongoing effort to comprehensively survey intrusion detection datasets with an unprecedented level of detail, implemented as a website backed by a public GitHub repository. Comidds allows researchers to quickly identify suitable datasets depending on their requirements and provides structured and critical information on each dataset, including actual data samples and links to relevant publications. Comidds is freely accessible, regularly updated, and open to contributions.

CCS CONCEPTS

• **Security and privacy** → **Intrusion/anomaly detection and malware mitigation**; • **Applied computing** → Enterprise computing; • **Computing methodologies** → Modeling and simulation.

KEYWORDS

Intrusion Detection, Dataset, Log Data, Netflow Data, Cyberattack, Enterprise Network, Testbed, Cyber Range, Simulation, Survey

ACM Reference Format:
Philipp Bönninghausen, Rafael Uetz, and Martin Henze. 2024. Introducing a Comprehensive, Continuous, and Collaborative Survey of Intrusion Detection Datasets. In *Workshop on Cyber Security Experimentation and Test (CSET 2024), August 13, 2024, Philadelphia, PA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3675741.3675754

CSET 2024, August 13, 2024, Philadelphia, PA, USA

ACM ISBN 979-8-4007-0957-9/24/08
https://doi.org/10.1145/3675741.3675754

1 INTRODUCTION

Intrusions of enterprise networks continue to affect thousands of organizations each year, often resulting in data theft, sabotage, and extortion [59]. Detecting such intrusions in a timely manner is difficult [1, 53], yet crucial to stop adversaries from reaching their final goals [38]. It is thus hardly surprising that intrusion detection is a highly active area of research for more than three decades now, with thousands of papers being published each year [31].

A large number of these works propose novel intrusion detection methods [3, 30, 60] and consequently require realistic data (resembling both benign and adversarial activity) to evaluate them against. Since many researchers lack access to enterprise networks or permission to run representative attacks against them, there is a high demand for appropriate public datasets [28]. In addition, public datasets (in contrast to private ones) allow for quantitative comparisons of works by different authors as well as independent analyses of the dataset itself to discover potential flaws [58].

Reacting to this high demand, researchers have created a multitude of datasets, which vary greatly in objective, age, and effort put into them [28]. Their contents cover a wide range of environments (e.g., office, cloud, or industrial context), activity (e.g., real or simulated benign activity as well as various attacks), and data formats (e.g., network flows, host log files, or system call traces) [3, 45].

Since there is no central registry for such datasets and relevant publications are spread over a large number of media and years, researchers may struggle to find datasets fitting their requirements and to fully understand their limitations and potential deficiencies [28]. In particular, some of the most popular and widely used datasets [31] show significant weaknesses [12, 28, 36, 40, 56]. Consequently, researchers using datasets should have an adequate knowledge of available datasets and their characteristics to avoid drawing invalid conclusions from experimental results.

To spare researchers from having to read hundreds of papers before using a dataset, various surveys give an overview of available datasets as well as independent analyses thereof (cf. Section 5). However, they suffer from three fundamental shortcomings: (1) They are *static* in the sense that they cannot be updated or corrected once published, (2) their descriptions of datasets are mostly superficial due to limited space, and (3) contained data (such as tables and plots) cannot be sorted, filtered, or otherwise processed automatically, e.g., to narrow down choices or create own statistics.

Addressing these shortcomings, we present Comidds – a **com**prehensive, continuous, and collaborative **i**ntrusion **d**etection **d**atasets **s**urvey. Comidds is freely accessible as a website backed by a public GitHub repository [5], thus allowing for ongoing extensions,

corrections, and change tracking. It provides an overview of key characteristics of all surveyed datasets (currently 48) and dedicated pages for each dataset containing detailed, structured, and critical information on their environment, activity, data format, related publications, and exemplary data snippets. Thus, Comidds assists researchers in finding and selecting appropriate datasets for their experiments and furthermore raises awareness of known limitations, eventually fostering advances in real-world intrusion detection. Overall, we make the following contributions:

- We introduce Comidds [5], a novel effort to survey intrusion detection datasets based on a GitHub repository (Section 2).
- We describe our methodology for finding relevant datasets, reviewing them, and adding them to Comidds (Section 3).
- We visualize key characteristics of the datasets surveyed so far to showcase our machine-readable survey data (Section 4).
- We compare Comidds to existing surveys, showing that it overcomes all shortcomings that we identified (Section 5).

2 COMIDDS: A REPOSITORY-BASED SURVEY OF INTRUSION DETECTION DATASETS

We begin with giving an overview of Comidds [5], including its goals, scope, and current features in the following. Based on this, we describe our methodology for finding, reviewing, and adding datasets to Comidds in Section 3.

Comidds' purpose is to aid researchers in finding and selecting suitable datasets to work with and to understand their potential limitations and deficiencies. It is *comprehensive* in the sense that it provides a structured and critical description for each contained dataset with a level of detail not seen in other surveys before. It is *continuous* in the sense that we will continue adding further datasets in the future, extend existing entries, and potentially correct discovered errors. Due to regular versioned releases with changelogs, users can directly track changes and reference fixed snapshots if desired. Comidds is *collaborative*, i.e., we strongly welcome contributions, both in the form of adding new dataset entries and improving existing ones. At the moment, Comidds contains information on 48 datasets as well as various short paragraphs on related work (13 survey papers and nine websites).

Goals. Motivated by the shortcomings of related surveys (cf. Sections 1 and 5), we set the following goals for Comidds:

- High coverage of datasets within our scope (see below): While the broadest survey that we found covers 52 datasets, we are striving to significantly exceed this number soon (cf. Section 3).
- Actionable description: Each dataset should be represented in a way that researchers can profoundly decide which dataset(s) to use and how to interpret experimental results based on them.
- Practical format: The survey should be easily accessible, extensible, maintainable, logically structured, and machine-readable.

Scope. We currently focus on datasets suited for developing and evaluating methods for intrusion detection in *enterprise networks*, i.e., environments usually involving client and server computers with common operating systems (particularly Windows and Linux), network hardware (e.g., routers, switches, firewalls), and typical applications and services (e.g. web, mail, directory). Adding datasets stemming from fundamentally different environments such as industrial control systems, Internet of Things, or otherwise specialized hardware or software is currently not planned by us, but might be considered if contributed by respective domain experts.

Features. To begin with, all included datasets are summarized in an overview table, which comprises the following columns:

- the **name** of the dataset as introduced by its author(s),
- a very **brief description** of the dataset,
- the fundamental **data type(s)** contained: *network* (e.g., network flows), *host* (e.g., operating system log files), or *both*,
- the **year(s)** of creation or, if unknown, of publication,
- the basic **environment**, e.g., *single system* or *enterprise IT*,
- the **operating system(s)**, e.g., *Windows* or *Linux*,
- the **labeling**: *direct* if data records are directly labeled as attack (class) or benign, *indirect* if only indirect labeling such as periods of attack are given, and *none* if no labels are present,
- the **data format(s)**, e.g., *NetFlow*, *syslog*, or *Suricata alerts*,
- the packed and unpacked **size** of the dataset in MB or GB.

In addition, for each dataset, there is a dedicated page containing an in-depth description, divided into the following sections:

- a **detailed table** showing concrete information beyond the summary table, such as *attack categories* and *benign activity*,
- an **overview** summarizing the origin, purpose, and contents of the dataset in a few sentences,
- information on the **environment** in which the dataset was recorded, e.g., the involved systems and network architecture,
- what **activity** was performed while recording the dataset (either by humans or synthetically), both benign and adversarial, and
- which **files** are actually contained in the dataset (with respect to data sources, formats, and labeling);
- moreover a list of **relevant papers**, including the original publication of the dataset as well as independent analyses thereof,
- links to **relevant websites** (especially the download location),
- a list of **related datasets**, and finally
- **sample records** for each data format contained in the dataset (excluding binary formats such as pcap).

Appendix A exemplarily shows the description of the popular CSE-CIC-IDS2018 dataset [49] as contained in the current version of Comidds. Last but not least, all key characteristics can be downloaded as a CSV (comma-separated values) file to facilitate custom sorting, filtering, or plotting, as we will showcase in Section 4.

3 SURVEY METHODOLOGY

In the following, we describe our methodology for identifying relevant datasets in the literature and analyzing them, respectively.

Literature Review. Prior to searching for original publications that contribute new intrusion detection datasets, we searched for already existing surveys of such datasets. For this purpose, we leveraged Google Scholar combined with domain knowledge from personally known researchers working in this field. We did not aim for a full coverage of such surveys since there is a large number of arguably redundant publications covering the same few datasets such as KDD Cup 1999 [22] or CSE-CIC-IDS2018 [49], often discussing

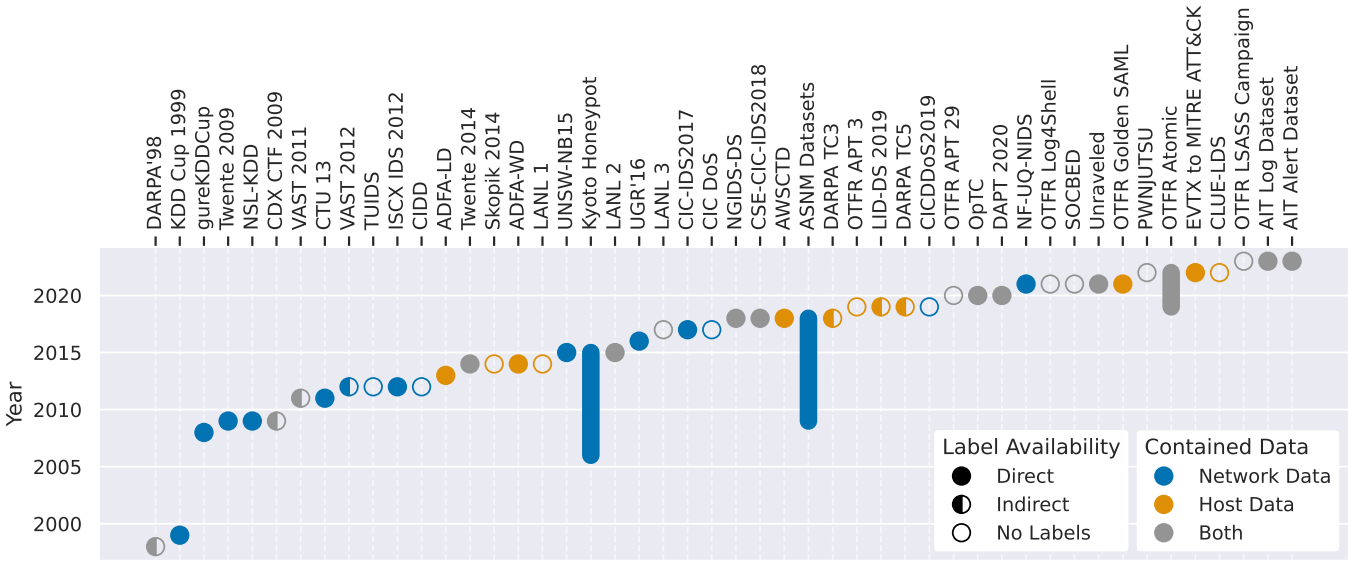

Figure 1: Age, labeling, and data types of all intrusion detection datasets surveyed until now (see Appendix B for references)

them with regards to some specific flaw or research question. Instead, we focused on a selection of recent surveys offering the most comprehensive overview (cf. Section 5).

Using these surveys as a starting point, we found a total of 90 datasets that fit our scope. To find further relevant datasets (especially those published after the latest surveys), we again utilized Google Scholar, using the search term *"intrusion detection dataset"*, but limiting our search to works published in the year 2023 or later to keep the number of results manageable. As this search resulted in 2030 works, we defined the following exclusion criteria:

- The publication does not contribute its own novel dataset,
- the contributed dataset is not publicly available,
- the contributed dataset does not contain adversarial activity,
- the publication is not available in English, or
- the publication is not available in electronic form.

After applying these criteria, we were left with 30 publications that contribute their own dataset, of which only ten match our scope (i.e., a focus on enterprise networks). Consequently, the total number of relevant datasets grew to exactly 100.

Lastly, we leveraged two more sources to find further datasets: (1) references within the selected publications (usually in the related work section) and (2) the domain knowledge of researchers in this field, in both cases following the exclusion criteria as defined above. This resulted in a grand total of 126 datasets. While this number might not be definitive, it excludes only those datasets that are not referenced by any major survey, not cited in any of these 126 works, and are unknown to several domain experts. At the time of writing, Comidds already covers 48 of these 126 datasets, focusing on the most popular ones, with more being added continuously. To include datasets into Comidds, we analyze them as follows.

Dataset Analysis. There are a number of dataset characteristics that various researchers regard as desirable, e.g., *documentation of labeling methodology* [32, 49, 58]. In an ideal world, every dataset would fulfill all of these characteristics, while also describing the process leading to their fulfillment. In reality, few publications document such issues, making it difficult to determine whether or not characteristics are present/fulfilled. For example, many works describe the simulation of benign activity in just a few sentences, making it close to impossible to determine if or to which extent the requirement of *realistic benign activity* is fulfilled.

Consequently, we do not aim to check each dataset against all requirements proposed in the literature, both because it is not feasible and requirements are often vague, making classification difficult or sometimes impossible, even with a lot of effort. Instead, we resort to an approach in part similar to that of Ring et al. [45], defining key characteristics (cf. Section 2) and reviewing all datasets with respect to them. We believe that this information serves as a sufficient representation for a given dataset, providing researchers with the means to quickly obtain detailed information and decide if this dataset could be suitable for their current undertaking. At the same time, this level of detail allows us to spend a feasible amount of time per dataset (usually a few hours).

During our analysis, we found that 23 of the 126 identified datasets are not backed by an academic publication or otherwise sufficient documentation. We decided that these datasets do not undergo the full analysis process as described above, but are instead listed separately on the Comidds website and each described in a single paragraph since at least some of the key characteristics cannot be determined from the documentation. However, we found cases where well-defined parts of such works were documented in a dedicated paper, thus being eligible for the previously described analysis process. For example, the *Malware Capture Facility Project* [15] provides a large number of pcap files collected from real networks, with little to no explanation for each of them – except for a select subset, known as CTU-13, which has its own paper [14] and is thus included in Comidds. We continue our discussion with presenting statistics of the datasets analyzed so far.

4 DATASETS STATISTICS

Since CoMIDDS provides key characteristics of all surveyed datasets in machine-readable CSV format, generating statistics and plotting them is straightforward. To illustrate this, we present two exemplary visualizations created from the data in the CSV file. They are also available in the repository (including source code) and updated automatically whenever dataset entries are added or changed.

Figure 1 depicts all datasets surveyed so far, where the y-axis shows the year of data creation or, if unknown, of publication. Datasets comprising more than one year are visualized accordingly. In addition, data types and label availability are shown (cf. Section 2). Note that while this figure provides a broad overview of the current datasets landscape, it also simplifies some aspects. For example, while the DARPA'98 and CSE-CIC-IDS2018 datasets contain both network and host data and are visualized as such, only their network data is labeled and thus typically used by other researchers.

Figure 2 plots multiple characteristics of the surveyed datasets, grouped into five categories: Network data formats, host data formats, type of benign activity, involved operating systems, and number of systems (in the sense of data-generating operating systems). Except for the last category, these classifications are not mutually exclusive, so the sum of a category does not necessarily match the total number of datasets surveyed. Note that we deliberately omit some characteristics, namely, dataset size, runtime, and number of machines, since we generally find them ineligible for a qualitative comparison of datasets with one another. For example, datasets containing packet captures can be orders of magnitude larger than NetFlow datasets despite resembling less activity. Similarly, runtime and number of machines do not necessarily correlate with the quantity and quality of benign or adversarial activity.

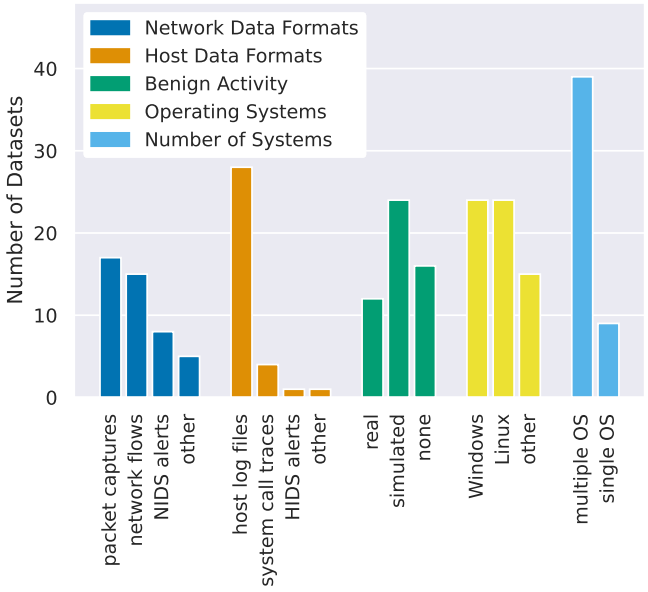

Figure 2: Characteristics of the surveyed datasets

5 RELATED WORK

While there exists a multitude of publications touching upon the topic of intrusion detection datasets, our discussion of related work focuses on works that share our principal goal of providing a broad yet actionable overview to help researchers choose appropriate datasets and understand their respective limitations as well as potential deficiencies.

Gümüşbaş et al. [19] discuss various intrusion detection methods based on deep learning, alongside which they present a list of datasets commonly used to benchmark these approaches. Twenty network-based datasets are described in a short manner, with the six most frequently cited undergoing further analysis regarding properties such as number of features and attack types. Bridges et al. [3] provide a survey focused on methods and datasets leveraging host data. They offer an overview of 22 datasets in the form of a brief description for each, listing information such as origin or data types, though not in a consistent manner. Yang et al. [60] compiled the broadest survey listed here, covering a large variety of publications and topics related to anomaly-based network intrusion detection, ranging from data preprocessing over evaluation metrics to datasets. They cover 52 datasets, although very little detail is provided for each of them.

Other surveys place their emphasis solely on datasets themselves. Ring et al. [45] provide the most in-depth overview of all studied papers, doing so by first defining 15 different dataset properties to describe, such as year of creation, format, duration, or type of network, and then applying this methodology to 34 network datasets, along with a description. Kenyon et al. [28] follow a similar approach, supplying a short paragraph per dataset but defining fewer features (origin, anonymization, data types, attack types). Furthermore, they define characteristics a dataset should fulfill in order to be suitable for intrusion detection research, and discuss a selection of the datasets with respect to them.

Lastly, in addition to these surveys, there are several works featuring a substantially smaller number of datasets, with the goal of answering specific research questions. For example, Landauer et al. [34] and Engelen et al. [12] analyze six and five popular datasets, respectively, discussing flaws affecting anomaly-based approaches and their consequences on state-of-the-art research. However, as the objective of our work is to offer a comprehensive survey that helps researchers to narrow down dataset choices, we consider their work to be complementary to ours.

A unifying property of the discussed broad surveys [3, 19, 28, 45, 60], as well as a driving motivator for the creation of CoMIDDS, is their lack of detail required to choose an appropriate dataset and become aware of its potential limitations and deficiencies. While certainly helpful in providing a general overview of a portion of currently existing intrusion detection datasets, researchers looking to process a dataset for their specific use case will have to invest substantial amounts of time into manual analysis, or resort to one of the most popular datasets (e.g., CSE-CIC-IDS2018 [49]) without questioning its suitability. As an example, only Ring et al. [45] provide basic information such as the data format (and even then only differentiate between "packet", "flow", and "other") and none provide actionable information on the simulation environment, ongoing activity within that environment, or samples of contained data – all of which are highly relevant, if not crucial, for performing and evaluating experiments based on these datasets.

6 CONCLUSION

This work addresses the challenges of selecting suitable datasets for intrusion detection research while taking into account their characteristics and potential deficiencies. We found that existing dataset surveys have significant shortcomings in the sense that they are static and either incomprehensive or superficial. With COMIDDS, we strive to resolve these issues by providing a repository-based survey that is comprehensive, continuous, and collaborative. COMIDDS currently covers 48 datasets and allows for sorting, filtering, and plotting of key characteristics to facilitate dataset selection. We will regularly add new datasets in the future and welcome contributions from other researchers. In addition, we intend to add further automatically-updated statistics and plots to the website. Ultimately, we hope that COMIDDS gains acceptance as a reference survey for intrusion detection datasets within its scope and thereby facilitates sound research in this practically relevant field.

ACKNOWLEDGMENTS

We would like to thank Frédéric Majorczyk and Maxime Lanvin for providing feedback on COMIDDS and suggesting additional datasets and papers. We also thank the anonymous reviewers for their time and valuable comments on the paper.

REFERENCES

[1] Stefan Axelsson. 2000. The Base-Rate Fallacy and the Difficulty of Intrusion Detection. *ACM Transactions on Information and System Security* (2000). https://doi.org/10.1145/357830.357849

[2] Aimad Berady, Mathieu Jaume, Valérie Viet Triem Tong, and Gilles Guette. 2022. PWNJUTSU: A Dataset and a Semantics-Driven Approach to Retrace Attack Campaigns. *IEEE Transactions on Network and Service Management* (2022). https://doi.org/10.1109/TNSM.2022.3183476

[3] Robert A. Bridges, Tarrah R. Glass-Vanderlan, Michael D. Iannacone, Maria S. Vincent, and Qian (Guenevere) Chen. 2019. A Survey of Intrusion Detection Systems Leveraging Host Data. *Comput. Surveys* (2019). https://doi.org/10.1145/3344382

[4] Dainius Čeponis and Nikolaj Goranin. 2018. Towards a robust method of dataset generation of malicious activity for anomaly-based HIDS training and presentation of AWSCTD dataset. *Baltic Journal of Modern Computing* (2018).

[5] COMIDDS contributors. 2024. COMIDDS: A comprehensive survey of datasets for research in host-based and/or network-based intrusion detection, with a focus on enterprise networks – GitHub. https://github.com/fkie-cad/COMIDDS

[6] Kristin Cook, Georges Grinstein, Mark Whiting, Michael Cooper, Paul Havig, Kristen Liggett, Bohdan Nebesh, and Celeste Lyn Paul. 2012. VAST Challenge 2012: Visual analytics for big data. In *IEEE Conference on Visual Analytics Science and Technology (VAST)*. https://doi.org/10.1109/VAST.2012.6400529

[7] Gideon Creech. 2014. *Developing a high-accuracy cross platform Host-Based Intrusion Detection System capable of reliably detecting zero-day attacks.* Ph. D. Dissertation. School of Engineering and Information Technology, University of New South Wales, Australia. https://doi.org/10.26190/unsworks/16615

[8] Gideon Creech and Jiankun Hu. 2013. Generation of a new IDS test dataset: Time to retire the KDD collection. In *IEEE Wireless Communications and Networking Conference (WCNC)*. https://doi.org/10.1109/WCNC.2013.6555301

[9] Gideon Creech and Jiankun Hu. 2014. A Semantic Approach to Host-Based Intrusion Detection Systems Using Contiguous and Discontiguous System Call Patterns. *IEEE Trans. Comput.* (2014). https://doi.org/10.1109/TC.2013.13

[10] DARPA. 2020. Operationally Transparent Cyber (OpTC) Data Release. https://github.com/FiveDirections/OpTC-data

[11] DARPA. 2020. Transparent Computing Engagement 3 and 5 Data Release. https://github.com/darpa-i2o/Transparent-Computing

[12] Gints Engelen, Robert Flood, Lisa Liu-Thorrold, Vera Rimmer, Henry Clausen, David Aspinall, and Wouter Joosen. 2022. Poster: Pillars of Sand: The current state of Datasets in the field of Network Intrusion Detection. In *European Symposium on Security and Privacy (EuroS&P)*. https://doi.org/10.5281/zenodo.7068716

[13] EVTX-to-MITRE-Attack contributors. 2024. mdecrevoisier/EVTX-to-MITRE-Attack – GitHub. https://github.com/mdecrevoisier/EVTX-to-MITRE-Attack

[14] Sebastián García, Martin Grill, Jan Stiborek, and Alejandro Zunino. 2014. An empirical comparison of botnet detection methods. *Computers & Security* (2014). https://doi.org/10.1016/j.cose.2014.05.011

[15] Sebastián García and Vojtech Uhlir. 2014. Malware Capture Facility Project. https://mcfp.weebly.com/

[16] Prasanta Gogoi, Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita. 2012. Packet and Flow Based Network Intrusion Dataset. In *Contemporary Computing*. https://doi.org/10.1007/978-3-642-32129-0_34

[17] Martin Grimmer, Martin Max Röhling, Dennis Kreußel, and Simon Ganz. 2019. A Modern and Sophisticated Host Based Intrusion Detection Data Set. In *Deutscher IT-Sicherheitskongress des BSI*. https://dbs.uni-leipzig.de/files/research/publications/2019-5/pdf/BSI-LID-DS.pdf

[18] Georges Grinstein, Kristin Cook, Paul Havig, Kristen Liggett, Bohdan Nebesh, Mark Whiting, Kirsten Whitley, and Shawn Konecni. 2011. VAST Challenge 2011: Mini Challenge 2 – Computer Networking Operations. https://visualdata.wustl.edu/varepository/benchmarks.php#VAST2011

[19] Dilara Gümüşbaş, Tulay Yıldırım, Angelo Genovese, and Fabio Scotti. 2021. A Comprehensive Survey of Databases and Deep Learning Methods for Cybersecurity and Intrusion Detection Systems. *IEEE Systems Journal* (2021). https://doi.org/10.1109/JSYST.2020.2992966

[20] Aric Hagberg, Nathan Lemons, Alex Kent, and Joshua Neil. 2014. Connected Components and Credential Hopping in Authentication Graphs. In *International Conference on Signal-Image Technology and Internet-Based Systems (SITIS)*. https://doi.org/10.1109/SITIS.2014.95

[21] Waqas Haider. 2018. *Developing reliable anomaly detection system for critical hosts: a proactive defense paradigm.* Ph. D. Dissertation. School of Engineering and Information Technology, University of New South Wales, Australia. https://doi.org/10.26190/unsworks/20924

[22] S. Hettich and S. D. Bay. 1999. The UCI KDD Archive. http://kdd.ics.uci.edu

[23] Rick Hofstede, Luuk Hendriks, Anna Sperotto, and Aiko Pras. 2014. SSH Compromise Detection using NetFlow/IPFIX. *ACM SIGCOMM Computer Communication Review* (2014). https://doi.org/10.1145/2677046.2677050

[24] Ivan Homoliak, Kamil Malinka, and Petr Hanacek. 2020. ASNM Datasets: A Collection of Network Attacks for Testing of Adversarial Classifiers and Intrusion Detectors. *IEEE Access* (2020). https://doi.org/10.1109/ACCESS.2020.3001768

[25] Hossein Hadian Jazi, Hugo Gonzalez, Natalia Stakhanova, and Ali A. Ghorbani. 2017. Detecting HTTP-based application layer DoS attacks on web servers in the presence of sampling. *Computer Networks* (2017). https://doi.org/10.1016/j.comnet.2017.03.018

[26] Alexander D. Kent. 2014. User-Computer Authentication Associations in Time. Los Alamos National Laboratory. https://doi.org/10.11578/1160076

[27] Alexander D. Kent. 2016. *Cyber security data sources for dynamic network research.* https://doi.org/10.1142/9781786340757_0002

[28] Tony Kenyon, Lipika Deka, and David Elizondo. 2020. Are public intrusion datasets fit for purpose characterising the state of the art in intrusion event datasets. *Computers & Security* (2020). https://doi.org/10.1016/j.cose.2020.102022

[29] Hisham A. Kholidy and Fabrizio Baiardi. 2012. CIDD: A Cloud Intrusion Detection Dataset for Cloud Computing and Masquerade Attacks. In *Information Technology - New Generations (ITNG)*. https://doi.org/10.1109/ITNG.2012.97

[30] Ansam Khraisat, Iqbal Gondal, Peter Vamplew, and Joarder Kamruzzaman. 2019. Survey of intrusion detection systems: techniques, datasets and challenges. *Cybersecurity* (2019). https://doi.org/10.1186/s42400-019-0038-7

[31] Satish Kumar, Sunanda Gupta, and Sakshi Arora. 2021. Research Trends in Network-Based Intrusion Detection Systems: A Review. *IEEE Access* (2021). https://doi.org/10.1109/ACCESS.2021.3129775

[32] Max Landauer, Florian Skopik, Maximilian Frank, Wolfgang Hotwagner, Markus Wurzenberger, and Andreas Rauber. 2023. Maintainable Log Datasets for Evaluation of Intrusion Detection Systems. *IEEE Transactions on Dependable and Secure Computing* (2023). https://doi.org/10.1109/TDSC.2022.3201582

[33] Max Landauer, Florian Skopik, Georg Höld, and Markus Wurzenberger. 2022. A User and Entity Behavior Analytics Log Data Set for Anomaly Detection in Cloud Computing. In *IEEE International Conference on Big Data (Big Data)*. https://doi.org/10.1109/BigData55660.2022.10020672

[34] Max Landauer, Florian Skopik, and Markus Wurzenberger. 2023. A Critical Review of Common Log Data Sets Used for Evaluation of Sequence-based Anomaly Detection Techniques. arXiv:2309.02854 [cs.LG]

[35] Max Landauer, Florian Skopik, and Markus Wurzenberger. 2023. Introducing a New Alert Data Set for Multi-Step Attack Analysis. arXiv:2308.12627 [cs.CR]

[36] Maxime Lanvin, Pierre-François Gimenez, Yufei Han, Frédéric Majorczyk, Ludovic Mé, and Éric Totel. 2023. Errors in the CICIDS2017 Dataset and the Significant Differences in Detection Performances It Makes. In *Risks and Security of Internet and Systems (CRiSIS)*. https://doi.org/10.1007/978-3-031-31108-6_2

[37] Richard P. Lippmann, David J. Fried, Isaac Graf, Joshua W. Haines, Kristopher R. Kendall, David McClung, Dan Weber, Seth E. Webster, Dan Wyschogrod, Robert K. Cunningham, and Marc A. Zissman. 2000. Evaluating intrusion detection systems: The 1998 DARPA off-line intrusion detection evaluation. In *DARPA Information Survivability Conference and Exposition (DISCEX)*. https://doi.org/10.1109/DISCEX.2000.821506

[38] Nate Lord. 2020. Cyber Security Investments: Experts Discuss Detection vs. Prevention. https://digitalguardian.com/blog/cyber-security-investments

[39] Gabriel Maciá-Fernández, José Camacho, Roberto Magán-Carrión, Pedro García-Teodoro, and Roberto Therón. 2018. UGR'16: A new dataset for the evaluation of cyclostationarity-based network IDSs. *Computers & Security* (2018). https://doi.org/10.1016/j.cose.2017.11.004
[40] John McHugh. 2000. Testing Intrusion Detection Systems: A Critique of the 1998 and 1999 DARPA Intrusion Detection System Evaluations as Performed by Lincoln Laboratory. *ACM Transactions on Information and System Security (TISSEC)* (2000). https://doi.org/10.1145/382912.382923
[41] Nour Moustafa and Jill Slay. 2015. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set). In *Military Communications and Information Systems Conference (MilCIS)*. https://doi.org/10.1109/MilCIS.2015.7348942
[42] Sowmya Myneni, Ankur Chowdhary, Abdulhakim Sabur, Sailik Sengupta, Garima Agrawal, Dijiang Huang, and Myong Kang. 2020. DAPT 2020 - Constructing a Benchmark Dataset for Advanced Persistent Threats. In *Deployable Machine Learning for Security Defense*. https://doi.org/10.1007/978-3-030-59621-7_8
[43] Sowmya Myneni, Kritshekhar Jha, Abdulhakim Sabur, Garima Agrawal, Yuli Deng, Ankur Chowdhary, and Dijiang Huang. 2023. Unraveled — A semi-synthetic dataset for Advanced Persistent Threats. *Computer Networks* (2023). https://doi.org/10.1016/j.comnet.2023.109688
[44] Iñigo Perona Balda, Olatz Arbelaiz Gallego, Ibai Gurrutxaga Goikoetxea, José Ignacio Martín, Javier Francisco Muguerza Rivero, and Jesús María Pérez de la Fuente. 2017. *Generation of the database gurekddcup*. Technical Report EHU-KAT-IK-02-16. University of the Basque Country. https://addi.ehu.es/handle/10810/20608
[45] Markus Ring, Sarah Wunderlich, Deniz Scheuring, Dieter Landes, and Andreas Hotho. 2019. A survey of network-based intrusion detection data sets. *Computers & Security* (2019). https://doi.org/10.1016/j.cose.2019.06.005
[46] Roberto Rodriguez and Jose Luis Rodriguez. 2022. Security Datasets. https://securitydatasets.com/
[47] Benjamin Sangster, T. J. O'Connor, Thomas Cook, Robert Fanelli, Erik Dean, William J. Adams, Chris Morrell, and Gregory Conti. 2009. Toward instrumenting network warfare competitions to generate labeled datasets. In *Conference on Cyber Security Experimentation and Test (CSET)*. https://www.usenix.org/legacy/events/cset09/tech/full_papers/sangster.pdf
[48] Mohanad Sarhan, Siamak Layeghy, and Marius Portmann. 2022. Towards a Standard Feature Set for Network Intrusion Detection System Datasets. *Mobile Networks and Applications* (2022). https://doi.org/10.1007/s11036-021-01843-0
[49] Iman Sharafaldin, Arash Habibi Lashkari, and Ali A. Ghorbani. 2018. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization. In *International Conference on Information Systems Security and Privacy (ICISSP)*. https://doi.org/10.5220/0006639801080116
[50] Iman Sharafaldin, Arash Habibi Lashkari, Saqib Hakak, and Ali A. Ghorbani. 2019. Developing Realistic Distributed Denial of Service (DDoS) Attack Dataset and Taxonomy. In *International Carnahan Conference on Security Technology (ICCST)*. https://doi.org/10.1109/CCST.2019.8888419
[51] Ali Shiravi, Hadi Shiravi, Mahbod Tavallaee, and Ali A. Ghorbani. 2012. Toward developing a systematic approach to generate benchmark datasets for intrusion detection. *Computers & Security* (2012). https://doi.org/10.1016/j.cose.2011.12.012
[52] Florian Skopik, Giuseppe Settanni, Roman Fiedler, and Ivo Friedberg. 2014. Semi-synthetic data set generation for security software evaluation. In *Annual International Conference on Privacy, Security and Trust (PST)*. https://doi.org/10.1109/PST.2014.6890935
[53] Robin Sommer and Vern Paxson. 2010. Outside the Closed World: On Using Machine Learning for Network Intrusion Detection. In *IEEE Symposium on Security & Privacy (IEEE S&P)*. https://doi.org/10.1109/SP.2010.25
[54] Jungsuk Song, Hiroki Takakura, Yasuo Okabe, Masashi Eto, Daisuke Inoue, and Koji Nakao. 2011. Statistical analysis of honeypot data and building of Kyoto 2006+ dataset for NIDS evaluation. In *Workshop on Building Analysis Datasets and Gathering Experience Returns for Security (BADGERS)*. https://doi.org/10.1145/1978672.1978676
[55] Anna Sperotto, Ramin Sadre, Frank van Vliet, and Aiko Pras. 2009. A Labeled Data Set for Flow-Based Intrusion Detection. In *IP Operations and Management (IPOM)*. https://doi.org/10.1007/978-3-642-04968-2_4
[56] Mahbod Tavallaee, Ebrahim Bagheri, Wei Lu, and Ali A. Ghorbani. 2009. A detailed analysis of the KDD CUP 99 data set. In *IEEE Symposium on Computational Intelligence for Security and Defense Applications (IEEE CISDA)*. https://doi.org/10.1109/CISDA.2009.5356528
[57] Melissa J. M. Turcotte, Alexander D. Kent, and Curtis Hash. 2018. *Unified Host and Network Data Set*. https://doi.org/10.1142/9781786345646_001
[58] Rafael Uetz, Christian Hemminghaus, Louis Hackländer, Philipp Schlipper, and Martin Henze. 2021. Reproducible and Adaptable Log Data Generation for Sound Cybersecurity Experiments. In *Annual Computer Security Applications Conference (ACSAC)*. https://doi.org/10.1145/3485832.3488020
[59] Verizon. 2023. 2023 Data Breach Investigations Report.
[60] Zhen Yang, Xiaodong Liu, Tong Li, Di Wu, Jinjiang Wang, Yunwei Zhao, and Han Han. 2022. A systematic literature review of methods and datasets for anomaly-based network intrusion detection. *Computers & Security* (2022). https://doi.org/10.1016/j.cose.2022.102675

Table 1: Detailed table for the exemplary CSE-CIC-IDS2018 dataset as contained in the current version of Comidds

Network Data Source(s)	pcaps, NetFlows
Network Data Labeled	Yes, NetFlows are labeled
Host Data Source(s)	Ubuntu & Windows event logs
Host Data Labeled	No
Overall Setting	Enterprise IT
OS Types	Windows 7/8/10/Vista/Server 2016, Ubuntu 14.04/16.04, MacOS; Kali & Windows 8.1 (Attacker)
Number of Machines	450
Total Runtime	~5 days
Year of Collection	2018
Attack Categories	Bruteforce, Heartbleed, Botnet, DoS/DDoS, Web-Based, Infiltration from Inside Network
Benign Activity	Synthetic, models complex behavior
Packed Size	220 GB
Unpacked Size	n/a
Download Link	Instructions at bottom of page

A EXEMPLARY DATASET ENTRY

This section shows information on the popular CSE-CIC-IDS2018 dataset as contained in Comidds as a concrete example for one of the currently 48 covered datasets. Note that this entry might be extended, corrected, or otherwise improved in future releases.

Overview

A collaboration between the Communications Security Establishment (CSE) and the Canadian Institute for Cybersecurity (CIC), this dataset uses the notion of profiles to generate cybersecurity datasets in a systematic manner, including various attack types and a large and diverse infrastructure. It is a continuation of previous efforts (CIC IDS2017), featuring similar attacks and benign behavior, but being significantly larger in scale (14 vs. 450 victim machines, 1 vs. 6 victim networks). While being one of the primary benchmark datasets in the current field of NIDS research, researchers have discovered errors within this dataset, affecting aspects like attack orchestration, feature generation, or labeling. Essential details of this dataset are summarized in Table 1.

Environment

The attacking infrastructure contains 50 machines, the victim infrastructure consists of 5 departments with a total of 420 PCs and 30 servers. An overview is provided by the diagram below [note: omitted in this paper to save space]. Presumably, vulnerable software versions have been installed to facilitate certain exploits, but this is more suggested than specified in their description.

Activity

Simulated behavior is defined in the form of profiles, divided into benign (B) and malicious (M) profiles. B-profiles are derived from observing human behavior, from which some features are learned/

extracted. M-profiles consist of seven different attack scenarios, each based on a certain attack type: Bruteforce, Heartbleed, Botnet, DoS, DDoS, Web-Based, and Infiltration from Inside Network. The total capturing period lasted ~5 days, with attacks being performed on every day except the first. Details for each attack as well as the timing are available on the linked homepage.

Contained Data

The dataset includes the network traffic and log files of each victim machine, combined with 80 network features extracted from captured traffic using CICFlowMeter. Data is divided into two main directories, `Network Traffic and Log Data` as well as `Processed Traffic Data for ML Algorithms`, with data being organized per day, respectively. The former contains raw data in the form of unlabeled network traffic (pcap) and event logs (Windows/Ubuntu). The latter consists of labeled features derived from the aforementioned network traffic (although the labeling logic is not transparently documented); these features are what is most commonly leveraged when using this dataset. Each feature is explained in detail on the homepage linked below. The aforementioned flaws of this dataset, such as some simulation artifacts making detection artificially easy, are for example laid out in Paper 2 referenced below.

Example Data

Note: We only show a small excerpt of the example data in this paper to give an idea of the structure on the Comidds website.

Labeled features from "Processed Traffic Data for ML Algorithms/ Thursday-01-03-2018_TrafficForML_CICFlowMeter.csv":

```
Dst Port,Protocol,Timestamp,Flow Duration,Tot Fwd...
3389,6,01/03/2018 09:56:59,4046191,14,7,1386,392,...
58655,6,01/03/2018 09:56:59,86620951,2,0,0,0,0,0,...
50657,6,01/03/2018 09:56:59,0,2,0,0,0,0,0,0,0,0,0,...
```

Ubuntu event logs taken from "Network Traffic and Log data/ Friday-16-02-2018/logs/U172.31.69.25":

```
Feb 16 07:39:01 ip-172-31-69-25 CRON[11625]: (root)...
Feb 16 07:48:09 ip-172-31-69-25 dhclient[922]: ...
Feb 16 07:48:09 ip-172-31-69-25 dhclient[922]: ...
```

Papers

1 Toward Generating a New Intrusion Detection Dataset and Intrusion Detection Traffic Characterization (2017)
2 Error Prevalence in NIDS datasets: A Case Study on CIC-IDS-2017 and CSE-CIC-IDS-2018 (2022)

Links

- Homepage. For download, install AWS CLI and run `aws s3 sync –no-sign-request –region <your-region> "s3://cse-cic-ids2018/" dest-dir`, where your-region is your AWS region and destination-dir is the target directory. If you only need the labeled features, use `s3://cse-cic-ids2018/Processed Traffic Data for ML Algorithms` as your URL.
- Secondary Source

Related Entries

- CIC IDS2017
- NF-UQ-NIDS

B LIST OF COVERED DATASETS

The following intrusion detection datasets are currently described in Comidds, ordered by year of creation/publication (newest first):

(1) AIT Alert Dataset [35]
(2) AIT Log Dataset [32]
(3) OTFR Security Datasets - LSASS Campaign [46]
(4) CLUE-LDS [33]
(5) EVTX to MITRE ATT&CK [13]
(6) OTFR Security Datasets - Atomic [46]
(7) PWNJUTSU [2]
(8) NF-UQ-NIDS [48]
(9) OTFR Security Datasets - Log4Shell [46]
(10) OTFR Security Datasets - SimuLand Golden SAML [46]
(11) SOCBED Example Dataset [58]
(12) Unraveled [43]
(13) DAPT 2020 [42]
(14) OpTC [10]
(15) OTFR Security Datasets - APT 29 [46]
(16) CICDDoS2019 [50]
(17) DARPA TC5 [11]
(18) LID-DS 2019 [17]
(19) OTFR Security Datasets - APT 3 [46]
(20) ASNM Datasets [24]
(21) AWSCTD [4]
(22) CSE-CIC-IDS2018 [49]
(23) DARPA TC3 [11]
(24) NGIDS-DS [21]
(25) CIC DoS [25]
(26) CIC-IDS2017 [49]
(27) Unified Host and Network Data Set [57]
(28) UGR'16 [39]
(29) Comprehensive, Multi-Source Cyber-Security Events [27]
(30) Kyoto Honeypot [54]
(31) UNSW-NB15 [41]
(32) ADFA-WD [7]
(33) Skopik 2014 [52]
(34) Twente 2014 [23]
(35) User-Computer Associations in Time [20, 26]
(36) ADFA-LD [7–9]
(37) CIDD [29]
(38) ISCX IDS 2012 [51]
(39) TUIDS [16]
(40) VAST Challenge 2012 [6]
(41) CTU 13 [14]
(42) VAST Challenge 2011 [18]
(43) CDX CTF 2009 [47]
(44) NSL-KDD [56]
(45) Twente 2009 [55]
(46) gureKDDCup [44]
(47) KDD Cup 1999 [22]
(48) DARPA'98 Intrusion Detection Program [37]

We will continue adding and improving dataset entries in the future.