CICI:IPAAI:CANIS: Curated AI-ready Network telescope datasets for Internet Security

Overview

The UCSD network telescope (UCSD-NT) has been a long-standing NSF-funded scientific cyberinfrastructure (CI), supporting the collection of unsolicited Internet (IPv4) traffic (Internet background radiation or IBR). Researchers use IBR data to detect a variety of malicious activities. But applying AI to UCSD-NT data is a double-edged sword. Advanced ML/AI models excel at identifying threats in IBR, but their success hinges on the *integrity, provenance, and authenticity* of the underlying data. Emerging risks to the integrity of the UCSD-NT data have coincided with exploding use of AI tools in cybersecurity research, creating an urgent challenge. This project directly tackles this challenge, with the ultimate goal of delivering high-quality, large-scale labeled datasets to safely train, validate, and benchmark AI models. But it requires infrastructure innovation.

The UCSD-NT faces mounting operational hurdles as usage of the underlying address space grows in magnitude and scope. The data's integrity also faces two external risks. First, IBR traffic growth increasingly strains UCSD-NT's packet-capture capacity, leading to data loss. Second, Internet routing disruptions, like misconfigurations and hijacks, impair connectivity to the UCSD-NT monitored address space, undermining the completeness of the data. Without constant monitoring, UCSD-NT is at risk of capturing legitimate (non-IBR) traffic and/or of reduced visibility.

Increasing use of AI tools and models that rely on this data amplifies the urgency of re-architecting its collection and curation infrastructure. Conventional cybersecurity datasets are used to train models to isolate rare malicious traffic from mostly legitimate flows, a process ill-suited for IBR's unique anomaly detection needs. The scarcity of large-scale, labeled IBR reference datasets stifles accurate AI model training and evaluation. UCSD-NT generates the largest such data set in the world, and as its use with AI tools expands to those with less expertise in the underlying network traffic characteristics, the integrity challenges become crucial to overcome.

To tackle these challenges, we propose CANIS, a suite of modules to transform the applicability of UCSD-NT in AI contexts through three complementary tasks. First, we will develop and deploy a new monitoring framework to safeguard cybersecurity research workflows. We will leverage CAIDA's active measurement infrastructure and public BGP data to monitor its connectivity, and use IBR generated by known scanning campaigns to *continuously verify the data integrity of UCSD-NT*. We will develop a new data format to disseminate information on status of the platform, lowering the risk of research use of inaccurate data in AI applications. Second, we'll enhance metadata by tagging IPs on blocklists or associated with network abuse or malware probes. We will also log system and network data, letting researchers trace whether their findings rest on flawed inputs. Third, we'll curate a library of curated, labeled reference datasets–snapshots of real-world events like malware outbreaks and scans–paired with AI-generated analyses, empowering researchers to efficiently benchmark and validate models.

Keywords: network telescope, AI-ready datasets, benchmarking, security, cyberinfrastructure.

Intellectual Merit

CANIS aligns with CICI's mission to strengthen the integrity and provenance of data generated by the UCSD-NT CI, delivering a transformative framework that enables trust and reproducibility in AI-driven cybersecurity and network research.

Broader Impacts

This project will yield curated datasets that facilitate anomaly detection, threat intelligence, and attack mitigation, ultimately strengthening global cybersecurity and workforce training efforts. Moreover, by bridging AI infrastructure and cybersecurity, CANIS sets a precedent for data-intensive fields, aligning research pipelines with transparency and accountability in AI innovation.

1 Introduction

Threat intelligence is vital to securing cyberinfrastructure (CI), pinpointing emerging attacks that exploit system weaknesses. For two decades, the NSF-funded UCSD Network Telescope (UCSD-NT) [1] has served as a cornerstone of such activities, capturing *Internet Background Radiation (IBR)*, unsolicited traffic to unused IPv4 addresses. Researchers rely on this scientific CI and the data it generates to expose malicious activities, from widespread vulnerability scans and malware outbreaks to backscatter from randomly spoofed denial-of-service (RSDoS) attacks.

Yet, applying AI to UCSD-NT data is a double-edged sword. Advanced ML/AI models–Autoencoders [2], Word2Vec [3, 4], LSTMs [5, 6], and Large Language Models (LLMs)–excel at spotting anomalies and identifying threats in IBR, but their success hinges on the *integrity, provenance, and authenticity* of the underlying data. However, emerging threats to the integrity of this data have coincided with exploding use of AI tools in cybersecurity research, creating an urgent need to mitigate these threats. This project directly tackles this challenge, with the ultimate goal of delivering high-quality, large-scale labeled datasets to safely train, validate, and benchmark AI models. Achieving this ambitious goal requires a transformative overhaul of the long-standing UCSD-NT infrastructure.

Originally built on a /8 IPv4 block allocated to the amateur radio community the UCSD-NT faces mounting operational hurdles as usage of the underlying address space grows. In 2019, a quarter of the address block (a /10 subnet) was sold to a cloud provider. The remaining address space has seen a variety of increased experimental usage, including temporary leases of subnets, by members of the amateur radio community. These shifting dynamics complicate efforts to track unused addresses. Compounding this challenge are two external threats to data integrity. First, IBR traffic growth increasingly strains UCSD-NT's packet-capture capacity, leading to data loss. Second, Internet routing disruptions, like BGP misconfigurations and hijacks, impair connectivity to the UCSD-NT monitored address space, undermining the completeness of the data. Without constant monitoring, UCSD-NT is at risk of capturing legitimate (non-IBR) traffic and/or of reduced visibility.

The increasing use of AI tools and models that rely on this data amplifies the urgency of re-architecting its collection and curation infrastructure. Currently, UCSD-NT provides only basic aggregated flow statistics [7], annotated with source IP geolocation and Autonomous System Number (ASN). The scientific community deems this inadequate for ML/AI tasks like anomaly detection and malware identification, forcing researchers to wrestle with re-processing terabyte-scale raw IBR traffic data.

This inconvenience is symptomatic of a larger more critical problem. The scarcity of large-scale, labeled IBR reference datasets stifles accurate AI model training and evaluation. Conventional cybersecurity datasets, such as CIC [8] and ISCX [9], are designed to train ML/AI models to detect rare malicious traffic from mostly legitimate flows, a process ill-suited for IBR's unique anomaly detection needs. To bridge these gaps, we propose CANIS, a suite of modules to revolutionize UCSD-NT through three complementary tasks.

First, we will develop and deploy a new monitoring framework to safeguard cybersecurity research workflows. We will leverage CAIDA's active measurement infrastructure to send beacon packets from globally distributed vantage points, and use the IBR generated by known Internet scanning campaigns to *continuously verify the data integrity of UCSD-NT*. We will also integrate public BGP data streams to detect changes in IP prefixes used by UCSD-NT. Additionally, we will develop a data format to disseminate information to inform researchers about the status of the platform, lowering the risk of using inaccurate data in their AI applications.

Our second task is to enrich the annotations of IBR, leveraging several available Internet mea-

surement datasets and labeling techniques. We will tag blocklisted source IPs and those used by cloud providers, whose services are often abused by spammers and attackers. We will apply advanced fingerprinting techniques to identify and tag probe packets generated by malware or malicious scanning campaigns. We will format the data in AI-friendly formats to ensure seamless compatibility with popular programming languages used in data science.

Our third task will fill the gap in AI-ready reference cybersecurity datasets. We will create a library of manually curated reference datasets for ML/AI training and benchmarking. By utilizing our recent research, we will select representative real-world events, such as the onset of malware spreading and scanning events targeted at vulnerabilities, and provide snapshots of both raw traffic data and flow data.

This work directly advances CICI's goals for Integrity, Provenance, and Authenticity for Artificial Intelligence Ready Data (IPAAI), by enhancing UCSD-NT's trustworthiness as a scientific CI. Our rich expansion of metadata, which we will document to facilitate adoption by other cybersecurity data providers, will increase the transparency and accountability of ML/AI applications, promising a transformative impact on AI-driven cybersecurity research.

2 Data Gaps in the UCSD Network Telescope

2.1 A brief history of UCSD-NT

UCSD-NT is a cyberinfrastructure for cybersecurity research. It leverages a largely unused IPv4 address space (darknet) to capture unsolicited Internet traffic. Over the last two decades, researchers at CAIDA have obtained an intermittent series of short-term research grants, including multiple NSF grants, that have supported the UCSD telescope infrastructure as scientific community CI [10–16], while also expanding its use beyond security research [17–19]. These funds have supported the router and data collection infrastructure at UCSD, as well as the sharing of raw and processed data with researchers, leading to hundreds of peer-reviewed publications [20].

The UCSD-NT has serendipitous origins. Since the 1990s UCSD has provided Internet transit for the Amateur Packet Radio Network (AMPRnet [21]) address space, which meant the UCSD network received most packets destined for this /8 IPv4 address space. The amateur radio community received this legacy address block in 1981 to support experimentation with connecting amateur radio infrastructure to Internet infrastructure, and to facilitate IP usage in radio communication. One of AMPRNet's founding members, Brian Kantor (WB6CYT), worked at UCSD for 47 years.

In 2011, Brian Kantor led the creation of a non-profit organization – Amateur Radio Digital Communications (ARDC) – to support AMPRNet and related activities [21], including continued allocation of the subnets within the larger /8 network, and evolution of their use. ARDC implemented a list of which subnets were allocated, namely the *filter list*; UCSD synchronized its configuration with this list daily to ensure that UCSD-NT would not monitor allocated subnets [16]. In 2015, increasing interest in AMPRNet address allocations [22] motivated ARDC to create a database and APIs to support this function. In 2019, ARDC sold the last quarter of the original /8 subnet (i.e., one /10) to a large cloud provider [23], further increasing the utilization of the remaining address space. After the sale, the UCSD-NT was limited to the remaining /9 and /10.

We identified two major challenges that stemmed from this historical setup:

- Similar to the hostname-to-IP mapping problem before the adoption of Domain Name System (DNS), the elevated allocation activities and the updates of the AMPRNet system resulted in the instrumentation repeatedly falling out of synchronization with the filter list. In mid-2023, scientific cybersecurity researchers using the UCSD-NT data discovered both unexpected

traffic as well as the absence of expected traffic to certain parts of the darknet.

- As AMPRNet currently does not have a plan to adopt routing security features, such as Resource Public Key Infrastructure (RPKI), the address space is vulnerable to BGP hijack attacks. Also, users who receive temporary allocation sometimes continue to announce the prefix(es) after their agreement expired. Although the hijacked or squatted prefixes may have no actual users, these events can redirect IBR away from UCSD-NT, reducing its coverage.

Fig. 1 shows the prefix utilization of the AMPRNet address space. The used address blocks (■ and ■) are non-contiguous. This visualization also identifies prefixes that may have been affected by BGP hijacking (■). Although AMPRNet operators confirmed that these prefixes were announced by members with expired allocations rather than due to malicious activity, the absence of IBR to these blocks could mislead researchers and their AI models.

Figure 1: Hilbert curve of original /8 AMPRNet address space on January 10, 2025, illustrating increased dynamics of address usage that require transformative innovations to protect integrity of the UCSD-NT data. Dark pixels (■) highlight telescope IP addresses. Blue pixels (■) are IP addresses assigned to AMPRNet users potentially getting transit via UCSD. Green pixels (■) are leased addresses covered by a BGP announcement. Red pixels (■) are potentially hijacked telescope addresses. The orange (■) /10 block was sold in 2019.

Figure 2: The existing packet capturing process of the UCSD-NT has shown packet loss in VMs during traffic surges. Since the packets received by the VMs from the multicast streams are jumbo frames, each lost packet may encapsulate hundreds of smaller packets, significantly compromising data integrity (§2.2).

Discovery of these incidents ultimately led to remedial actions by CAIDA. In the process we realized the urgency of the need for continuous monitoring, to reduce the time to detect and remediate these threats to the *integrity, provenance, and authenticity* of the data.

2.2 System artifacts in UCSD-NT

The volume of IBR has grown over the last two decades, and has reached 3.86TB per day on average [24]. Short-term traffic surges have also become more prominent. Fig. 3 shows an example of traffic spikes similar to the one detected by our ML-based anomaly detection method [24]. The surge lasted about 23 minutes and increased the peak traffic rate to over 3M packets/min. We geolocated the source of the traffic to the United States (Fig. 3a). Our investigation found that the simultaneous drop in traffic from other countries was likely due to the high volume of traffic which overwhelmed the system and induced severe packet loss (Fig. 3b). The frequency and intensity of these surges are increasing. We recorded 166 similar events in 2024. These events compromise the integrity of the data, posing risks to AI models that depend on the data.

Fig. 2 shows the current UCSD-NT infrastructure that CAIDA researchers developed in 2017 [15]. The packet capture machine is a bare-metal server equipped with an Endace DAG card to capture packets from a split of the fiber to the AMPRNet gateway. The packet capturer encapsulates multiple received packets into one UDP jumbo frame, leveraging multicast to distribute the frames to different virtual machines (VMs), including one that stores the traffic into PCAP files. Our investigation found that the VM's kernel was dropping jumbo frames during traffic surges.

Currently, UCSD-NT does not provide external telemetry data to monitor packet loss or inform researchers about potential traffic loss. This data gap makes it difficult

(a) Top 10 countries by total packet count. Traffic sourced by IPs geolocated to the U.S. (blue curve) surged, impairing packet collection that resulted in loss of packets from other countries, which distorts the UCSD-NT signal.

(b) Traffic volume (top), number of packets dropped by network interface (middle), and number of UDP receive buffer errors (bottom) recorded on the packet capturing VM (Figure 2). The surge resulted in a traffic rate of ≈ 6Gbps, which overwhelmed the network and the kernel.

Figure 3: Red areas highlight a 23-minute traffic surge in UCSD-NT in February 2025.

for researchers to assess the integrity of the data.

2.3 Why other approaches to collection of IBR are not sufficient

Given IPv4 scarcity, researchers have developed alternatives to traditional telescopes, including renting cloud instances to collect IBR [25]. Unfortunately, a disproportionately high volume of packets to previously-active service ports is likely on cloud IP addresses [26], potentially impacting system using these cloud IP addresses. Researchers have also explored collecting traffic in transit to dark (unassigned) IP addresses on large production networks [27,28] or at IXPs [29]. Traditional network telescope deployment is still necessary to provide accurate baseline of IBR without any interference [30].

2.4 IBR data for ML/AI use

Researchers have developed ML/AI applications based on UCSD-NT (or similar network telescope deployment) to detect malicious Internet activities (e.g., [5,6,31–40]). These methods analyzed the relationships and dynamics between source IPs and destination ports to extract insights from the IBR. We categorize this prior work by the fundamental ML/AI techniques they use.

Multivariate statistical analysis ccreates matrix representations using source IPs and/or destination ports in IBR data. Due to the sparse nature of these matrices, researchers have developed dimension reduction techniques (e.g., Dark-NMF [32], and Dark-Tracer [33]) to extract meaningful insights, such as identifying synchronous scanning behavior and detecting abnormal changes in traffic patterns.

Graphical approaches model IBR traffic as graphs (e.g., [37–40]) and apply clustering or community detection algorithms to identify similar nodes. However, these methods often face scalability challenges due to the large size of the graphs.

Word2Vec is a natural language processing technique that generates vector representations of words. DANTE [4] and DarkVec [3] leveraged this method to encode the scanning sequences of each source IP, enabling the clustering of scanners with similar behaviors.

Time series methods analyze traffic statistics (e.g., DarkSIM [24], [36]) to detect potential anomalies in IBR data, helping to narrow the scope of traffic analysis.

The evaluation of these studies relies on network telescopes deployed in various locations, including Japan [33], Italy [40], France [39], and the U.S. [24]. These telescopes vary significantly in size, ranging from a single /24 subnet [3] to over 10 million IPs [24]. Such disparities make direct comparisons between different methods nearly impossible.

Furthermore, unlike honeypots, network telescopes do not interact with senders, making it difficult to collect definitive ground truth regarding the intent of the senders. This lack of interaction poses a challenge in accurately evaluating the effectiveness of detection methods.

Except for time series methods, most other work analyzed raw IBR traffic in PCAP format. UCSD-NT provides data access in three formats.

PCAP. The *de facto* format for sharing network traffic. Due to the storage requirements (several terabytes per day), UCSD-NT keeps only the last 30 days of data on-premise, and archives older files into NERSC's HPSS archive system [41].

FlowTuple [7]. UCSD-NT provides this Apache avro-based format files that store aggregated traffic statistics between source IPs and destination /16 subnets every 5 minutes. We annotate each flow record with IP geolocation and ASNs.

Traffic timeseries. UCSD-NT indexes three metrics (packet count, traffic volume, and number of unique source IPs) for hundred of thousands of combinations of features, including protocols, port numbers, source IP geolocation, and ASNs. Researchers can access these timeseries via Grafana dashboards and perform exploratory data analysis to discover which flow records of packet captures they need to analyze in more detail [42].

During our process of reproducing some of these methods (more details in §3.3), we identified shortcomings in FlowTuple for AI applications, such as limited time resolution and insufficient metrics collected, which necessitated re-processing raw traffic data.

To conduct a thorough scientific evaluation of ML/AI algorithms for analyzing IBR, it is crucial to assess different methods on a common foundation and perform cross-validation across them.

3 Proposed Research and Development Agenda: CANIS

CANIS will consist of a suite of modules to improve the UCSD-NT infrastructure for data acquisition, preprocessing, and data analytics of typical cybersecurity research workflows. These enhancements will contribute to our overall objective: scaling the collection and sharing of IBR datasets that will facilitate the transparent and accountable use of ML/AI in cybersecurity research. To achieve this goal, this project has three major thrusts: 1) enhancing the data integrity of UCSD-NT; 2) building an AI-ready data processing pipeline; 3) crafting reference datasets for ML/AI training and evaluation. Fig. 4 shows the overview of the proposed CANIS architecture.

Figure 4: The overview of CANIS. Dashed rectangle boxes encloses modules and deliverables

3.1 Task 1: Improving the data integrity of UCSD-NT

We will design and implement three transformative enhancements to UCSD-NT to address the data integrity problems that arise from packet loss and routing anomalies.

1. Improving packet capturing reliability and robustness. We will re-architect the packet capturing process in UCSD-NT to reduce system overhead and increase the reliability of traffic collection. As most UCSD-NT data users perform offline analyses, our new design will prioritize the integrity of the traffic collection process (Fig. 5). We will move the PCAP generation from a VM subscribing to packet multicast streams to a container on the UCSD-NT packet capture machine. More specifically, we will use Open vSwitch [43] to poll packets from the SmartNIC using Data Plane Development Kit (DPDK) [44] and mirror them to two Docker containers.

Figure 5: Software components of the proposed packet capture server.

The *traffic collection* container will capture the traffic into PCAP files and store them in CAIDA's storage cluster. Open vSwitch does not support DPDK for the virtual ports delivering the mirrored traffic, so we will employ kernel bypass network primitives, such as PF_RING, that boost the packet capture performance. The *traffic streaming* container will run the existing software for encapsulating and multicasting packets into the VM network to continue supporting the live streaming of IBR and other non-critical post-processing tasks.

Telemetry monitoring. We will capture network statistics from the SmartNIC, Open vSwitch, and the network interface in the traffic collection container. We will deploy `telegraf` to gather and record the number of received and dropped packets at these components. Also, we will monitor common system metrics, such as CPU usage, disk I/O, and memory utilization. The collected data will stream to UCSD-NT's InfluxDB instance (Fig. 4, Task 1).

2. External monitoring from the Internet. A number of cybersecurity companies (e.g., Censys, Shadowserver, Alpha Strike Labs, and Leitwert) and academic (e.g., University of Michigan and Technical University of Munich) periodically launch Internet-wide scanning campaigns to measure the liveness, detect open services, and identify vulnerabilities on end-hosts and servers. UCSD-NT captures probe packets from these campaigns. As the behavior of these campaigns are predictable with data access, we will leverage them as independent tests to audit the integrity of UCSD-NT without generating additional overhead or noise into the traffic signal.

Figure 6: Packets/minute from three scanners' ASes (Table 1) to UCSD-NT from 18-25 Mar 2025. All three scanners showed persistent scanning traffic that we can leverage to audit UCSD-NT data integrity.

Table 1 lists the three scanners whose scanning we observe in UCSD-NT and have access to their scanning results. These scanners have different spatial and temporal scanning strategies, complementing each other. Fig. 6 shows the traffic rate in packets/min from the source ASes of these three campaigns. All three scanners persistently sent probing traffic to various segments of the UCSD-NT address space.

Table 1: External scanning projects used in this work to identify anomalies.

Scanner	Traffic type	Scanning strategy
Alpha Strike	TCP SYN port scans	All IPv4 addresses every 4h
Leitwert	TCP Paris Traceroutes	One IPv4 address within every /22 every day
Censys	TCP/UDP popular ports	All IPv4 addresses at least once a day

To leverage this scanning traffic for assessing the integrity of UCSD-NT, we will first manually cross-validate the scanning results with existing UCSD-NT data to ensure that the spatial (address space scanned) and the temporal (scanning frequency) characteristics match their public documentation. We will implement software to analyze UCSD-NT data daily to examine if UCSD-NT is capturing the expected traffic in all of the addresses and prefixes. Missing packets or unexpected packets to some addresses/prefixes indicates potential configuration errors, such as out-of-sync subnet filters.

3. Active measurements from Internet vantage points. Passive monitoring of known scanning campaigns only allows us to examine the connectivity from the scanners' networks to the UCSD-NT. We will use CAIDA's active measurement programming primitives that we recently published [45] to conduct traceroutes from CAIDA's Ark vantage points to UCSD-NT. Thus, we can collect the forward paths from many networks to the UCSD-NT. As UCSD-NT currently does not

respond to any incoming traffic, traceroute measurements could take seconds to wait for timeout.

We will deploy a *response server* to reply to these traceroute probes to speed up the measurements and ensure the packets reach the UCSD-NT. However, a responding subnet tends to attract scanning traffic [46]. We will restrict this behavior to a dedicated /24 subnet in the UCSD-NT address space to isolate it from the pure unsolicited IBR. Apart from testing for connectivity, this measurement monitors the packet loss rate in UCSD-NT by comparing the number of probe packets sent to UCSD-NT and the total number of probes received.

4. Control plane monitoring. In addition to data plane measurements (traceroute and scanning), we will monitor the BGP routing tables of AMPRNet prefixes using BGP data collected by RouteViews [47]. We will track three metrics to detect BGP anomalies.

I. We will monitor the global visibility of the AMPRNet prefixes' routes in different BGP monitors. Therefore, we can ensure that BGP announcements of AMPRNet's prefixes correctly propagate across the Internet.

II. We will periodically compare the prefixes and their lengths with those reported by the AMPRNet address API, ensuring consistency between assigned prefixes and global BGP announcements.

III. Since AMPRNet cannot currently deploy routing security measures such as RPKI, we will seek to employ BGP hijack detection techniques (e.g., [48]) to identify common hijack attacks that could redirect traffic from UCSD-NT. For this task we hope to leverage infrastructure that may be developed in a parallel CICI collaboration (proposed by CAIDA and Internet2) to verify the routing integrity of R&E network fabrics.

The deliverable of this task will significantly increase the integrity of the UCSD-NT data, improving its verifiability and providing an authentic foundation for ML/AI training and modeling. More broadly, this task will facilitate scientific research to improve the security and resilience of critical infrastructure.

3.2 Task 2: Creating AI-ready IBR datasets

Our second task is to build a pipeline to enhance and enrich the labeling of IBR data while disseminating it in an AI-friendly format. The curated data will facilitate ML/AI applications by enabling researchers to focus on modeling and analysis rather than data preparation.

In our survey of prior work that used IBR datasets (§2.4), we identified two major ML/AI applications that would benefit from enhanced labeling: scanner clustering and anomaly detection. To support these use cases, we will develop labeling functions that leverage four types of information:

i) Blocklists. We will map source IPs observed in UCSD-NT to IP blocklists, such as FireHOL IP lists [49] and AbuseIPDB [50]. IPs listed in these databases often show evidence of malicious activities, including scanning and spamming. Researchers can leverage these labels for ML/AI training and as features in their models.

ii) Known scanners. In Task 1, we leveraged three Internet-wide scanning campaigns to examine the integrity of UCSD-NT. These campaigns are often benign, and identifying them can help reduce the complexity of AI models for detecting cyber threats. However, some scanners do not explicitly disclose the IP ranges they use, and manually curated lists (e.g., [51]) may be outdated. We will employ large language models (LLMs) to identify the IP ranges and ASNs associated with these campaigns, as LLMs can efficiently summarize publicly available information. We will validate the output of the LLM.

iii) Malware-infected hosts. Some malware families, such as Mirai and Crackonosh, perform Internet-wide scanning. Such malware often injects packet fingerprints or algorithmically computes destination ports. We will leverage these characteristics to identify malware-generated

packets. These labels will be used to train ML/AI-based scanner clustering algorithms.

iv) Malicious activity engagement. We will apply scanner detection logic similar to Zeek [52] to identify source IPs engaging in horizontal and vertical scanning activities. Since scanners may behave differently in a darknet environment compared to active networks, we will refine the detection algorithms and thresholds to distinguish heavy hitters from ethical scanning performed for research or network diagnostics.

Providing AI-ready data format and metrics. The legacy UCSD-NT uses a FlowTuple format, built on Apache Avro [53], which uses a compact, row-based structure. However, this format forces users to read entire files even when only a few features are needed, slowing down data exploration, searching, and subsetting for researchers. To overcome this, CANIS will adopt Apache Parquet [54], a column-oriented format that blends storage efficiency with rapid retrieval. Widely supported in data science languages like Python and R, and compatible with Apache Spark, Parquet enables fast access to high-performance computing resources like SDSC Expanse [55]. Its encoding and metadata allow selective extraction—such as specific columns or filtered records—without loading full files into memory. CANIS will produce two Parquet file types for ML/AI use:

Label Mapping Files. These files will map source IPs to labels, aiding algorithms like [56] in clustering potentially malicious IPs and simplifying models. Given the daily updates of some blocklists, labels remain stable short-term, so we'll refresh these files hourly.

Traffic Statistics Files These files will contain packet counts and unique destination counts indexed by source IP–destination port pairs, preserving the sequence of observations. Multiple ML/AI algorithms (e.g., [4, 32, 33, 56]) rely on such data to construct high-dimensional matrices or word embeddings of IP-port pairs to detect anomalies. We will generate these files every five minutes, aligning with existing FlowTuple updates.

Our survey found that some algorithms require a finer time resolution than five minutes. We will leverage Parquet's page structure to segment the data into one-minute pages, allowing researchers to recover data at a one-minute resolution from the same files.

The new AI-friendly IBR data format directly addresses the needs of ML/AI algorithms in this domain. It also facilitates the use of national HPC resources, such as those provided by NAIRR.

3.3 Task 3: Curating reference datasets

We will curate a set of reference datasets that are critical for AI training and evaluation. We will select time periods that capture representative and well-documented events, such as the onset of malware outbreaks, scanning campaigns targeting well-known vulnerabilities, and large-scale DDoS attacks. To provide more fine-grained information for AI training, the reference datasets, along with the data we will continuously generate in Task 2, will preserve PCAP files and provide detailed descriptions of the events. More importantly, we will use these reference data sets to replicate published ML/AI research, and publish our findings. Having such reference data sets is essential for transparent and accountable assessments of the performance and accuracy of newly developed models and comparisons to state-of-the-art approaches.

We will include representative algorithms employing different types of models and applications, such as the ones we surveyed (§2.4). These algorithms will execute on SDSC Expanse [55], an NSF-funded HPC resource available to researchers through ACCESS-CI [57]. During execution, we will record runtime, CPU usage, and memory consumption. More importantly we will share the results, such as clusters of source IPs, word embedding from Word2Vec-based methods, and detected events. Additionally, we will share our code to facilitate replication, enabling researchers to benchmark the time and space complexity of their models under the same platform.

Some algorithms, when evaluated on a smaller-scale network telescope, may not scale to handle the traffic volume of UCSD-NT. To address this limitation, we will provide scripts to downsample the UCSD-NT data to an appropriate size. The sampling process is critical to ensuring reproducibility. For example, selecting a portion of the darknet where UCSD-NT does not have full visibility could introduce bias into AI models.

Preliminary Results. We reproduced DarkVec [3] using IBR data collected by UCSD-NT in August 2016, at the onset of Mirai—a malware that creates botnets and remains active to this day. The original evaluation was conducted using a small network telescope covering a single /24 subnet [3]. However, the code published by the author was not efficient for use with UCSD-NT data. We re-implemented an HPC-compatible version, sampled five /16 subnets from UCSD-NT as input data, and executed it on SDSC Expanse.

Figure 7: Computational performance of DarkVec against three metrics (counts of packets, bytes, and unique senders) using subsets of UCSD-NT data. Top row shows runtime per observation period. Bottom row shows the memory consumption. Capturing these metrics enables us to characterize the complexity of the analysis methods and compare them with newly developed ML/AL models.

We recorded the runtime and memory usage while varying the length of the observation period, b (Fig. 7). We found that DarkVec's time and space complexity are linearly correlated with the number of packets or bytes in each observation period. However, the algorithm requires analyzing a larger number of observation periods. Its runtime also positively correlated with the number of source IPs observed in an observation period, as the size of the word embedding increases.

Figure 8: Visualization of scanner clusters reported by DarkVec [56] during the onset of Mirai in 2016. Each color represents one cluster. We used the parameters suggested in the original work (b=5, and 6 clusters). Our reference datasets will include these results, facilitating comparisons with the outcomes of other ML/AI methods.

Fig. 8 shows DarkVec's clustering results. We set the algorithm to create six clusters, as suggested in [3]. At the onset of the Mirai botnet, we expected no variants to be present, meaning all infected hosts should have grouped into a single cluster. However, using Mirai's packet fingerprint (i.e., *TCP sequence number==source IP*), we found that the infected hosts were distributed evenly across clusters (ranging from 12.1% to 22.5% of hosts per cluster), failing to reproduce the published results. We are currently using this dataset to try to replicate other published techniques (e.g., [2]) to evaluate their accuracy in identifying Mirai-infected hosts.

Throughout the project, we will create similar AI-ready reference datasets and share them with cybersecurity and AI researchers. The open-source tools developed by CANIS will lower the barrier for researchers to utilize NSF-funded AI resources.

4 Software license

Contributing open-source software not only benefits the cybersecurity and AI communities but also enables third-party researchers to validate and improve the work, fostering a mutually beneficial environment. The software developed during this project will be released under an open-source license. Following the suggestion from the UCSD campus, we will adopt UCSD's Academic Non-

Commercial License [58]. We plan to retain this license for future enhancements to the software and related products.

5 Data sharing approach

We will adopt the FAIR guiding principles [59] to share the data and related resources, aiming to increase the machine-actionability of the datasets.

Findable. We will index all datasets created in CAIDA's resource catalog. Each catalog entry includes metadata, such as tags and keywords, to help users search for data. The catalog supports custom search queries [60], allowing users to refine search results further. We have already created catalog entries for the UCSD-NT data released so far [61] and will continue this practice for datasets generated by this project.

Accessible. Due to the sensitive nature of the data, users must complete a data request form on CAIDA's website. Our data administrator will create accounts for approved users, granting access based on different use cases—for example, password-protected web pages for downloading standalone datasets or shell access to research VMs for longitudinal datasets.

Interoperable. UCSD-NT stores traffic data in PCAP (or PCAP-NG) format, a de facto standard for this type of data. The new labeled and reference datasets will use the Apache Parquet format, which is widely adopted in the data science community. Numerous open-source tools and libraries are available to parse this format, and major cloud providers (e.g., AWS and Azure) also support it.

Reusable. We will regularly update documentation on accessing and parsing the IBR data captured by UCSD-NT and the reference datasets. We will release code recipes in CAIDA's resource catalog, sharing best practices developed by our team for analyzing the data.

6 Sustainability plan

Over the past two years, CAIDA has been actively working to ensure the sustainability of the UCSD-NT infrastructure by securing service agreements with institutions interested in using the data. In the past few years, we established paid service agreements with the Information Sciences Institute (ISI) at the University of Southern California, Lincoln Laboratory at MIT, and Louisiana State University. Additionally, we received funding from Amateur Radio Digital Communications (ARDC) to support the daily operations and data collection of UCSD-NT. We are also collaborating with established and start-up security firms to explore commercial licensing of UCSD-NT data to help sustain its availability for academic research. CANIS will significantly enhance the reliability and robustness of UCSD-NT, potentially aligning with industry availability standards.

The new dataset features and AI-ready reference datasets are in high demand for ML/AI-based cybersecurity research. We believe this project can foster new collaboration opportunities with both academic researchers and industry partners. Additionally, we will collect user feedback through workshops and GitHub, fostering a dedicated AI research community that can continue supporting and expanding this work beyond the project's duration.

7 Ethics and operational concerns

We implement multiple safeguards to ensure that UCSD-NT does not capture user traffic from the AMPRNet address spaces. The network interface for packet capturing physically connects only with the fiber optics for the ingress direction. Therefore, UCSD-NT's capturing server cannot observe egress traffic from the AMPRNet infrastructure, by design.

We will collaborate closely with the network administrators of these monitored address blocks, including UCSD and ARDC, to automatically update UCSD-NT's traffic filters as IP address assignments change. Maintaining up-to-date filters is crucial for accurately discarding user traffic and protecting privacy. We will share our new monitoring capabilities with ARDC's operators, providing them additional information about their network.

Since the telescope data can unintentionally expose victims of ongoing DDoS attacks and devices compromised by malware or worms, we impose strict access controls and vetting of users. The data is not publicly available, and researchers seeking access to compute VMs or traffic data must disclose their intended use and sign an Acceptable Use Agreement (AUA) [62]. We also restrict users from distributing the raw data without CAIDA's written authorization. Our project manager and data administrator review and renew access permissions every 90 days.

8 Quantitative evaluation metrics

We will evaluate the success of CANIS with metrics in four categories.

1. Dataset curation, accessibility, and adoption in AI research. We will evaluate the project based on the number of reference datasets released. Researchers benefit from a variety of high-quality datasets to improve model accuracy. To quantify the impact, we will track:

- The number of users and institutions accessing our datasets
- The number of downloads per dataset
- The number of research publications citing our datasets
- The specific usage of labels (Task 2) and reference datasets (Task 3) in AI research

2. System performance. We will assess the *peak traffic rate* that the new packet-capturing server can sustain without significant packet loss. The validation process will involve:

- Replay testing using historical traffic traces in a testbed environment
- Running the new and existing systems in parallel to compare captured traffic
- Evaluating the effectiveness of external scanning campaigns in detecting packet loss using internal telemetry

We will also evaluate the data retrieval performance when using Parquet-formatted data, comparing it to the previous FlowTuple format.

3. Data coverage and integrity.

- The number of labels generated in Task 2 and the ratio of labeled to unlabeled traffic will measure the coverage of our annotations in IBR data.
- We will track the number of anomaly events detected by CANIS and assess their impact on UCSD-NT's visibility of IBR traffic.

4. Community engagement.

- We will dedicate sessions at CAIDA's annual AIMS workshop to discussions on UCSD-NT, creating a platform for researchers to share their experiences.
- At these workshops we will offer tutorials to engage potential data users.
- We will use workshop attendance, engagement, and feedback collected via exit surveys as additional metrics of project impact.

9 Broader impacts

By providing high-quality, AI-ready network telescope datasets, CANIS will accelerate the development of ML/AI techniques for cybersecurity. Researchers will have access to curated datasets that facilitate anomaly detection, threat intelligence, and attack mitigation, ultimately strengthening global cybersecurity efforts.

CANIS will leverage federally funded cyberinfrastructure for AI training and modeling. The AI-ready datasets and associated tools will lower barriers for cybersecurity researchers to utilize these resources, promoting broader adoption of federally funded AI investments.

The recently approved School of Computing, Information, and Data Sciences (SCIDS) at UCSD aims to train a future workforce that will drive AI advancements. The AI-ready cybersecurity datasets released through this project will support SCIDS's educational initiatives in cybersecurity and AI. Educators can integrate these datasets into coursework, helping to train the next generation of cybersecurity experts and AI practitioners.

In the Winter 2023 quarter, for Dr. Claffy's Internet Data Science and Cybersecurity course, PI Mok and co-PI Claffy extracted data collected by UCSD-NT to create a hands-on assignment to analyze a darknet traffic sample. This exercise significantly increased student interest in network traffic analysis. This project will facilitate developing such data-driven assignments for use in SCIDS and other departments.

References

[1] CAIDA, "The UCSD network telescope." https://www.caida.org/projects/network_telescope/.

[2] M. Kallitsis, R. Prajapati, V. Honavar, D. Wu, and J. Yen, "Detecting and interpreting changes in scanning behavior in large network telescopes," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3611–3625, 2022.

[3] L. Gioacchini, L. Vassio, M. Mellia, I. Drago, Z. B. Houidi, and D. Rossi, "Darkvec: automatic analysis of darknet traffic with word embeddings," in *Proceedings of the 17th International Conference on Emerging Networking EXperiments and Technologies*, CoNEXT '21, (New York, NY, USA), p. 76–89, Association for Computing Machinery, 2021.

[4] D. C. et al., "DANTE: A framework for mining and monitoring darknet traffic," in *Computer Security – ESORICS 2020*, (Cham), pp. 88–109, Springer International Publishing, 2020.

[5] M. Zakroum, A. Houmz, M. Ghogho, G. Mezzour, A. Lahmadi, J. FranCois, and M. E. Koutbi, "Exploratory data analysis of a network telescope traffic and prediction of port probing rates," in *2018 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pp. 175–180, IEEE.

[6] M. Zakroum, J. François, I. Chrisment, and M. Ghogho, "Monitoring network telescopes and inferring anomalous traffic through the prediction of probing rates," *IEEE Transactions on Network and Service Management*, vol. 19, no. 4, pp. 5170–5182, 2022.

[7] CAIDA, "FlowTuple." https://stardust.caida.org/docs/data/flowtuple/.

[8] I. Sharafaldin, A. Habibi Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," *Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP)*, pp. 108–116, January 2018.

[9] A. Shiravi, H. Shiravi, M. Tavallaee, and A. A. Ghorbani, "Toward developing a systematic approach to generate benchmark datasets for intrusion detection," *Computers and Security*, vol. 31, pp. 357–374, May 2012.

[10] David Moore and Geoffrey Voelker and Stefan Savage, "NSF Proposal: NSF-01-160: Quantitative Network Security Analysis," 2002. https://www.caida.org/funding/nsftrust/.

[11] kc claffy, "NSF Proposal: A Real-time Lens into Dark Address Space of the Internet," 2010. https://www.caida.org/funding/cri-telescope/.

[12] Alberto Dainotti and kc claffy, "IODA: Detection and analysis of large-scale Internet infrastructure outages," 2011. https://www.caida.org/funding/ioda/.

[13] Kimberly Claffy, "DHS Proposal: PREDICT/IMPACT - Supporting Research and Development of Security Technologies through Network and Security Data Collection," 2012. https://www.caida.org/funding/impact/.

[14] Alberto Dainotti and Marina Fomenkov, "IODA-NP: Multi-source Realtime Detection of Macroscopic Internet Connectivity Disruption," 2017. https://www.caida.org/funding/paridine-iodanp/.

[15] Alberto Dainotti and Alistair King, "NSF Proposal: STARDUST: Sustainable Tools for Analysis and Research on Darknet Unsolicited Traffic," 2017. https://www.caida.org/funding/stardust/.

[16] Ricky Mok and kc claffy and Fabian Bustamante, "NSF Proposal: Scalable Technology to Accelerate Research Network Operations Vulnerability Alerts," 2023. https://www.caida.org/funding/cici-starnova/.

[17] Dainotti, A and Squarcella, C and Aben, E and claffy, k and Chiesa, M and Russo, M and Pescapè, A, "Analysis of Country-wide Internet Outages Caused by Censorship," in *Internet Measurement Conference (IMC)*, pp. 1–18, ACM, November 2011.

[18] K. Benson, A. Dainotti, k. claffy, and E. Aben, "Gaining Insight into AS-level Outages through Analysis of Internet Background Radiation," in *ACM SIGCOMM Conference on emerging Networking EXperiments and Technologies (CoNEXT) Student Workshop*, pp. 63–64, ACM, December 2012.

[19] O. Gupta, *Identifying Traffic Anomalies Interfering with IBR Based Outage Detection*. PhD thesis, UC San Diego, June 2018. `https://catalog.caida.org/paper/2018_identifying_traffic_anomalies`.

[20] CAIDA, "Papers by non-caida authors that used caida data." `https://catalog.caida.org/search?query=types%3Dpaper%20links%3Dtag%3Aused_caida_data%20!links%3Dtag%3Acaida%20`.

[21] Wikipedia, "AMPRNet," 2024. `https://en.wikipedia.org/wiki/AMPRNet`.

[22] Amateur Radio Digital Communications, "ARDC 44Net Assessment Results," 2024. `https://www.ardc.net/wp-content/uploads/ARDC-44net-Survey-Assessment-Results-sm.pdf`.

[23] Gene Takagi, 2020. `https://www.ardc.net/wp-content/uploads/Courtesy-Notice-to-AG-Signed-ARDC.pdf`.

[24] M. Gao, R. Mok, E. Carisimo, k. claffy, E. Li, and S. Kulkarni, "DarkSim: A Similarity-Based Time Series Analytic Framework for Darknet Traffic," in *ACM Internet Measurement Conference (IMC)*, November 2024.

[25] E. Pauley, P. Barford, and P. McDaniel, "DScope: A Cloud-Native Internet Telescope," in *Proceedings of the 32nd USENIX Security Symposium (USENIX Security 2023)*, (Anaheim, CA), USENIX Association, Aug. 2023.

[26] E. Pauley, R. Sheatsley, B. Hoak, Q. Burke, Y. Beugin, and P. McDaniel, "Measuring and mitigating the risk of ip reuse on public clouds," in *2022 IEEE Symposium on Security and Privacy (SP'22)*, pp. 558–575, IEEE, 2022.

[27] J. Li, F. Xhafa, J. Weng, R. Wang, Z. Liu, M. Tao, and L. Zhang, "Identifying internet background radiation traffic based on traffic source distribution," *Journal of High Speed Networks*, vol. 21, no. 2, pp. 107–120, 2015.

[28] L. Miao, W. Ding, and H. Zhu, "Extracting internet background radiation from raw traffic using greynet," in *18th International Conference on Networks (ICON)*, IEEE, 2012.

[29] D. Wagner, S. A. Ranadive, H. Griffioen, M. Kallitsis, A. Dainotti, G. Smaragdakis, and A. Feldmann, "How to operate a meta-telescope in your spare time," in *Proceedings of the 2023 ACM on Internet Measurement Conference*, IMC '23, (New York, NY, USA), p. 328–343, Association for Computing Machinery, 2023.

[30] E. Bou-Harb, M. Husák, M. Debbabi, and C. Assi, "Big data sanitization and cyber situational awareness: A network telescope perspective," *IEEE Transactions on Big Data*, vol. 5, no. 4, pp. 439–453, 2019.

[31] C. Han, J. Shimamura, T. Takahashi, D. Inoue, M. Kawakita, J. Takeuchi, and K. Nakao, "Real-time detection of malware activities by analyzing darknet traffic using graphical lasso," in *2019 18th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/13th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*, pp. 144–151, 2019.

[32] C. Han, J. Takeuchi, T. Takahashi, and D. Inoue, "Automated detection of malware activities using nonnegative matrix factorization," in *2021 IEEE 20th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom)*, pp. 548–556, 2021.

[33] C. Han, J. Takeuchi, T. Takahashi, and D. Inoue, "Dark-tracer: Early detection framework for malware activity based on anomalous spatiotemporal patterns," *IEEE Access*, vol. 10, pp. 13038–13058, 2022.

[34] E. Bou-Harb, M. Debbabi, and C. Assi, "Behavioral analytics for inferring large-scale orchestrated probing events," in *2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pp. 506–511, IEEE.

[35] E. Bou-Harb, M. Debbabi, and C. Assi, "A time series approach for inferring orchestrated probing campaigns by analyzing darknet traffic," in *2015 10th International Conference on Availability, Reliability and Security*, pp. 180–185, IEEE.

[36] E. Bou-Harb, M. Husak, M. Debbabi, and C. Assi, "Big data sanitization and cyber situational awareness: A network telescope perspective," vol. 5, no. 4, pp. 439–453.

[37] S. Lagraa and J. Francois, "Knowledge discovery of port scans from darknet," in *2017 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pp. 935–940, IEEE.

[38] S. Lagraa, Y. Chen, and J. François, "Deep mining port scans from darknet," *International Journal of Network Management*, vol. 29, no. 3, p. e2065, 2019. e2065 nem.2065.

[39] L. Evrard, J. François, and J.-N. Colin, "Attacker behavior-based metric for security monitoring applied to darknet analysis," in *2019 IFIP/IEEE Symposium on Integrated Network and Service Management (IM)*, pp. 89–97, 2019.

[40] F. Soro, M. Allegretta, M. Mellia, I. Drago, and L. M. Bertholdo, "Sensing the noise: Uncovering communities in darknet traffic," in *2020 Mediterranean Communication and Computer Networking Conference (MedComNet)*, pp. 1–8, 2020.

[41] National Enery Research Scientic Computing Center (NERSC), "HPSS data archive." `https://www.nersc.gov/systems/hpss-data-archive/`.

[42] CAIDA, "UCSD-NT Grafana dashboards." `https://explore.stardust.caida.org/d/AlX8okkMz/home?orgId=1`.

[43] Open vSwitch Community, "Open vswitch." Online, 2025. Version 3.5.90, Accessed: 2025-04-01.

[44] DPDK Project, "Data plane development kit (dpdk)." Online, 2025. Accessed: 2025-04-01.

[45] M. Luckie, S. Hariprasad, R. Sommese, B. Jones, K. Keys, R. Mok, and k. claffy, "An Integrated Active Measurement Programming Environment," in *Passive and Active Measurement Conference (PAM)*, December 2024.

[46] C. Fachkha and M. Debbabi, "Darknet as a source of cyber intelligence: Survey, taxonomy, and characterization," vol. 18, no. 2, pp. 1197–1227.

[47] University of Oregon, "Route Views Project." `http://www.routeviews.org/`, 2017.

[48] B. Al-Musawi, P. Branch, and G. Armitage, "Bgp anomaly detection techniques: A survey," vol. 19, no. 1, pp. 377–396.

[49] FireHOL, "All cybercrime ip feeds." `https://iplists.firehol.org`.

[50] "AbuseIPDB." `https://www.abuseipdb.com`.

[51] M. Collins, "Acknowledged Scanners." `https://gitlab.com/mcollins_at_isi/acknowledged_scanners`, 2021.

[52] "Zeek." `https://zeek.org`.

[53] Apache, "Avro."

[54] Apache, "Parquet." `https://parquet.apache.org`.

[55] San Diego Supercomputer Center, "Expanse." `https://www.sdsc.edu/systems/expanse/index.html`.

[56] L. Gioacchini, L. Vassio, M. Mellia, I. Drago, Z. B. Houidi, and D. Rossi, "DarkVec: Automatic Analysis of Darknet Traffic with Word Embeddings," in *Proceedings of the 17th International*

Conference on Emerging Networking EXperiments and Technologies, CoNEXT '21, (New York, NY, USA), pp. 76–89, Association for Computing Machinery, 2021.

[57] "ACCESS." https://access-ci.org.

[58] UCSD, "Copyright overview." https://innovation.ucsd.edu/disclose-patent/copyright.html.

[59] GO FAIR, "FAIR principles." https://www.go-fair.org/fair-principles/.

[60] Center for Applied Internet Data Analysis, "CAIDA Catalog View: Papers known to use UCSD Telescope Data (294 on 25 Jan 2025)," 2025. https://catalog.caida.org/search?query=types%3Dpaper%20links%3Dcollection%3Aucsd_telescope_datasets.

[61] CAIDA, "Resource catalog: telescope datasets." https://catalog.caida.org/search?query=types%3Ddataset%20telescope.

[62] CAIDA, "CAIDA master acceptable use agreement (AUA)." https://www.caida.org/about/legal/aua/.