CICI:TCR: STARNOVA: Scalable Technology to Accelerate Research Network Operations Vulnerability Alerts

Overview

Cyber attacks, such as ransomware, malware, and denial-of-service (DOS), are persistent threats to the security, reliability, and robustness of scientific cyberinfrastructure (CI). We propose a translational research effort to extend the capabilities of existing NSF-funded Internet measurement infrastructure—UCSD Network Telescope (UCSD-NT)—to immediately improve the robustness, integrity, and resilience of wide range of CI hosted in the San Diego Supercomputer Center (SDSC). Specifically, we propose to develop a novel platform—*Sustainable Technology to Accelerate Research Network Operations Vulnerability Alerts* (STARNOVA)—which will substantially expand the capability to identify targeted attacks against scientific CI at the SDSC.

Our project leverages unsolicited traffic, including Internet-wide scanning campaigns that often provide an an early-warning indicator of new attacks. Such scanning blends in with normal user traffic, but becomes more prominent in darknets. CAIDA operates the largest and longest-running darknet telescope on the Internet, positioning us to effectively translate results of this technology to our hosting CI site, the San Diego Supercomputing Center (SDSC).

We structure the design and implementation of STARNOVA as three tasks. Our first task is to expand the visibility of UCSD-NT by capturing traffic toward SDSC's production networks. We will leverage network/broadcast IP addresses in each SDSC's subnets and the addresses assigned to router interface and point-to-point links to form *greynets*, a collection of dark IP addresses that interspersed with active addresses in the same subnets, to capture unsolicited traffic. We will deploy equipment to mirror the traffic to the modernized UCSD-NT.

Second, we will leverage NSF-funded compute resources (i.e., the Expanse supercomputer at SDSC) to deploy our recent research on machine learning(ML)-based time series analytic methods to detect anomalies in IBR traffic. Our method will efficiently analyze over 200K time series and identify those containing either transient or persistent suspicious pattern changes.

Third, we will automate network flow analysis to examine the time series flagged by our anomaly detection method. We will enhance our current flow data representation, correlate anomalies in different time series, identify potentially affected services, and infer attack origins. We will implement near real-time alerts for operators to make informed defensive actions against potential threats.

Keywords: network telescope, anomaly detection, darknet, traffic, security, cyberinfrastructure.

Intellectual Merit

This project joins the efforts of network engineering and recent advances in cybersecurity research to provide protection to scientific CI. Our novel approach combines advanced traffic capture capabilities, network operations expertise, and ML-based time series analytics to scalably detect traffic anomalies in near real-time. The outcome of this project will provide accurate and custom alerts on emerging cyberattacks for network operators.

Broader Impact

This project is directly responsive to the NSF Cybersecurity Innovation for Cyberinfrastructure Area 3. By translating network telescope measurement instrumentation to research network environments, we will address urgent security management gaps by providing new visibility into threats against SDSC, home of a wide range of NSF-funded scientific cyberinfrastructure. The published design, software, and data will facilitate traffic data collection and information exchange across the network security and scientific CI community, contributing to affordable community-based solutions, especially critical at under-resourced institutions.

1 Introduction

Cyber attacks, such as ransomware, malware, and denial-of-service (DOS), are persistent threats to the security, reliability, and robustness of scientific cyberinfrastructure (CI). Adversaries exploit vulnerabilities to gain unauthorized system access, disrupt services, and retrieve sensitive data through the Internet, severely jeopardizing the integrity of research workflows. Existing defenses, such as regularly updating software and intrusion detection systems can mitigate attacks from known vulnerabilities. However, systems are exposed to attacks during the lag time between applying patches and signature updates. We propose a transitional research effort to extend the capabilities of existing NSF-funded security research instrumentation to immediately improve the robustness, integrity, and resilience of SDSC's wide range of scientific cyberinfrastructure.

Our efforts start with the recognition that attackers have the advantage: they have to discover only one vulnerable host or service, while defenders have to protect all of them. Defenders thus must learn how to leverage all available sources of threat intelligence. Ironically, a powerful source of threat intelligence is an empty but connected network that hosts no devices or services – called a *darknet*. All traffic received by such a network has no active destination endpoint, which implies three properties: it is suspicious, unsolicited, and not privacy-sensitive. In homage to astronomers, researchers call such traffic *Internet background radiation* (IBR). One significant component of IBR traffic is Internet-wide *scanning campaigns*, often conducted to identify vulnerable hosts. Some scans thus represent an early-warning indicator of new attack waves. Such scanning blends in with normal user traffic, but becomes more prominent in darknets.

Over the last two decades, CAIDA (at UCSD) has operated the world's largest network telescope (UCSD-NT) to capture IBR from a darknet. CAIDA's STARDUST platform enables researchers to access the captured IBR traffic data for security studies, e.g., characterizing distributed denial of service attacks (DDoS) [1–4], network censorship [5, 6], and the spread of botnets and malware [7–12]. Given the scarcity of IPv4 address space needed to create such instrumentation, the only realistic way to sustain such data collection is to extract unsolicited traffic from active Internet address space. Any given IP network has at least two IP addresses that should never engage in two-way communication: the network/broadcast IP addresses of each subnet and the addresses assigned to router interfaces or point-to-point links. We use this fact to operationalize security use of *greynets* [13]: networks that capture traffic to their unutilized addresses.

We propose to develop a novel platform—*Scalable Technology to Accelerate Research Network Operations Vulnerability Alerts* (STARNOVA)—which will transform our capability to identify targeted attacks against scientific CI at the San Diego Supercomputer Center (SDSC). Our approach combines advanced traffic capture capabilities, network operations expertise, and innovations in machine learning (ML)-based time series analytics to scalably detect anomalies in near real-time. We structure the design and implementation of STARNOVA as three tasks.

1. **Expand the visibility of UCSD-NT.** We will improve the capability of the UCSD-NT to capture traffic toward SDSC's greynets. The team includes SDSC operational security staff who will deploy and operate equipment to mirror this traffic to our infrastructure.
2. **Anomaly detection for IBR traffic.** We will leverage NSF-funded compute resources (i.e., the Expanse supercomputer at SDSC) to deploy our recent research innovations in machine learning(ML)-based time series analytic methods [14] to detect anomalies in IBR traffic.
3. **Automatic event analysis and alert generation.** We will automate analysis and correlation of anomalies, identify targeted services, and infer attack origins. We will implement integrations with instant-messaging tools (e.g., Mattermost and Slack) to offer real-time alerts to inform operator decisions on applying appropriate defensive actions against potential threats.

This project is directly responsive to the NSF Cybersecurity Innovation for Cyberinfrastructure

Area 3's program goals of *improving the robustness of scientific CI through operational or at-scale deployment, test and evaluation of novel cybersecurity research and techniques.* More specifically, STARNOVA will greatly enhance the capability of UCSD-NT, a well-established infrastructure, to improve the network security of SDSC, which is the home of national NSF-funded scientific CI, such as Expanse [15], Voyager [16], Science Portals & Science Gateways [17], and National Research Platform [18]. SDSC's CI supports important stages in research workflows, particularly compute and analytics, data movement, storage, and sharing, for health IT services, workflow automation, and internet data analysis. Enhancing the robustness and resilience of scientific CI at SDSC could broadly benefit research across science domains.

2 Background

We first provide background on CAIDA's UCSD-NT and STARDUST infrastructures that collect, process, and share IBR traffic data (§2.1). We then provide an overview of security-related analyses of this traffic (§2.2). We also highlight academic and industry efforts to provide threat intelligence (§2.3), and why they are insufficient for the needs of modern scientific cyberinfrastructure.

2.1 UCSD-NT and STARDUST infrastructure

The UCSD-NT captures network traffic sent toward three-quarter of a /8 IPv4 network owned by a non-profit organization, consisting of around 12M IP addresses. Even though most of these IP addresses are *dark* (i.e., not assigned to devices and do not respond to any traffic), UCSD announces legitimate BGP routes to receive the traffic. After filtering traffic from the few legitimate users in the network, the UCSD-NT captures and stores the packets (in pcap format) for off-line analysis. Given current storage constraints, UCSD-NT provides the most recent 60-days of pcap files onsite and sends historical pcap files to NERSC HPSS data archive [19] for long-term storage.

The large data size of this IBR traffic (>100GB per hour) imposes challenges for researchers to perform data analysis. In 2019, CAIDA announced the STARDUST platform, supported by an NSF-funded project that concluded in 2021 (CNS-1730661) "Sustainable Tools for Analysis and Research on Darknet Unsolicited Traffic (STARDUST)", offering three ways to access the data.

1. **STARDUST virtual machines.** STARDUST provides virtual machines (VMs) to researchers to access *live* darknet traffic. UCSD-NT uses multicast to broadcast IBR traffic to the VMs. Researchers bring their code to the VMs to analyze real-time IBR traffic.
2. **FlowTuple files.** The Corsaro software package [20], developed for STARDUST, processes IBR traffic and generates FlowTuple files [21], which are Apache Avro formatted files for compact representation of network flow records, every 5 minutes. Each FlowTuple record represents a sequence of packets sharing features, including source IPs, protocol and destination ports. The Corsaro software package computes traffic statistics of the flows (e.g., distribution/frequency of packet sizes, time-to-live value) and annotates each flow with metadata that facilitates analysis (e.g., prefix-to-AS [22], and IP geolocation).
3. **Traffic time series data.** For near real-time traffic monitoring, STARDUST computes a set of traffic properties (Table 1), yielding over 200K time series. We apply heuristics [23] to identify traffic with spoofed source IP addresses and implement filters to prevent them from compromising our statistical analysis. We use InfluxDB [24], a time series database, to index the data, and Grafana dashboards [25] to publish interactive visualization [26].

We propose three transformative changes to UCSD-NT infrastructure to achieve our cybersecurity innovations: introducing a new traffic source from a stratified sample of SDSC CI's IP address space (§3.1); translating new ML-based methods to support real-time anomaly detection and anal-

Table 1: Traffic metrics, properties, and filters, that when combined, yields over 200K time series. We will use these time series as the input of the anomaly detection method (§3.2).

Properties	Metrics (per minute)	Filters
Origin ASN	# of packets (PPM)	Unfiltered
Geolocation	# of bytes (BPM)	Non-Spoofed
Protocol number	# of unique source IPs	Spoofed (Derived)
TCP/UDP Destination port	# of unique source ASN	
ICMP type & code	# of unique destination IPs	
Spoofing inference		

ysis to this new combined data set (§3.2); and enriching the FlowTuple representations to scale performance of the analysis sufficiently to allow automated notifications and informed operational response by SDSC (§3.3).

2.2 Research to infer security and stability properties of IBR traffic

A vast body of literature has used IBR traffic for studying cybersecurity problems, including distributed Denial-of-Service attacks (DDoS) [1–4], network censorship [5, 6], spread of botnets and malwares [7–12]. Detecting anomalies and classifying the nature of traffic is one of the fundamental goals in darknet research. Researchers leverage fields in packet headers (e.g., TCP/UDP destination ports, TCP flags, TCP sequence numbers, and TTLs) to identify activities. For example, different scanners probe different sets of TCP/UDP ports for vulnerable services [7, 8, 27]. Backscatter induced by DDoS attacks are response packets, such as TCP SYN/ACKs, TCP RSTs, and DNS responses, from victims [2, 28]. Mirai's probe packets use the destination IP addresses as the TCP sequence numbers [9], and ZMap's probe packets have a constant IPID of 54321 [29].

Researchers have recently made progress applying machine learning approaches to study darknet traffic. Balkanli *et al.* [3, 4] used Naive Bayes, Chi-Square and Symmetrical Uncertainty, and the Decision Tree classifier to train supervised learning models to detect DDoS attacks. Time series analyses were applied to infer regional Internet outages from darknet traffic [30, 31]. Gupta [30] analyzed noises in IBR time series data, including traffic bursts and IBR generated by BitTorrent. Chocolatine [31] trained a seasonal Autoregressive Integrated Moving Average (SARIMA) model to predict the number of source IPs appearing in IBR traffic, facilitating detection of Internet outages. Darkvec [10], Kallitsis *et al.* [12], and DANTE [32] leveraged clustering algorithms to identify coordinated scanning campaigns and scanners with similar behavior. However, these models require expensive training with labeled data and are not suitable for real-time analysis. This research will develop novel ML-based methods to achieve near real-time anomaly detection.

2.3 Industry and community efforts to provide threat intelligence

A closely related approach to network telescope is network honeypots, which pretend to be vulnerable hosts to attract attacks from malware and adversaries. That is, a honeypot deployment will respond to unsolicited traffic trying to penetrate a specific IP address, in order to discover the nature of the attempted attack. There are academic and industry deployments of honeypots, such as STINGAR [33], Internet Storm Report [34], GreedyBear [35], and Greynoise [36]. Honeypots require more resources to operate as they must emulate a real operating system.

Some researchers have used the UCSD-NT to create reactive telescope platforms [11, 37], only

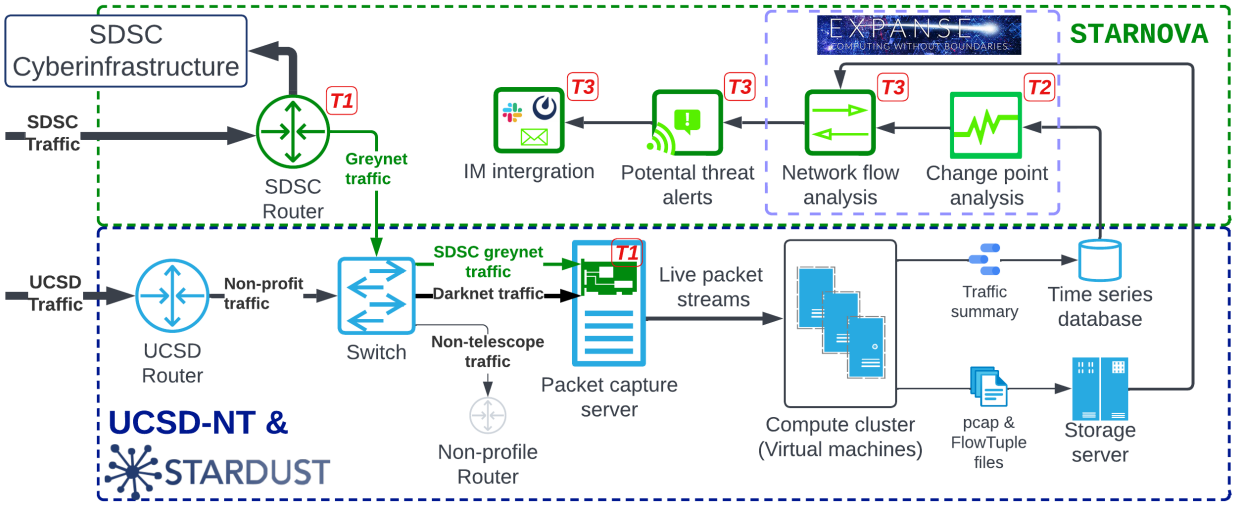


Figure 1: Architecture of UCSD-NT & STARDUST (dashed blue box) and STARNOVA (dashed green box). New components (green) are labeled with corresponding tasks in proposal (red squares).

briefly interacting with the attackers to infer their intent. Although both types of deployment provide intelligence about malicious IP addresses, they require constant updates in order to identify new attacks and maintain instrumentation on unused IP address space. More importantly, someone else's honeypot cannot provide visibility into attacks targeting one's own (in this case SDSC's) CI.

3 Enhancing the cybersecurity of SDSC's CI with STARNOVA

We structure our project into three inter-related tasks (*T1*-*T3* in Fig. 1). First, we will partner with SDSC network engineers (See LoC) to deploy a traffic aggregator that will identify traffic to greynet addresses and forward it to UCSD-NT. Obtaining this traffic data will require tackling unique challenges in monitoring ultra high capacity (400Gbps) networks which SDSC will deploy in 2023. We will also upgrade and modernize UCSD-NT's and STARDUST's software and hardware platforms to handle additional traffic load (§3.1). Second, we will operationalize our recent ML-based change point detection method [14] to identify anomalies in darknet and greynet traffic (§3.2). Third, we will develop methods to automate network flow analysis to extract finer-grain information on anomalies, and correlate traffic data from the two networks to identify threats to SDSC's CI (§3.3). Task 2 and 3 will leverage Expanse, an NSF-funded HPC platform at SDSC, to perform near real-time analysis.

3.1 Task 1. Expanding visibility and capability of UCSD-NT

The UCSD-NT effectively reveals uniformly random Internet-wide scanning campaigns, because it observes over 1/340 of probes in such campaigns. To verify whether the scans impact the CI at SDSC, a straightforward approach would be to deploy an unutilized address block as a production darknet. However, unused IPv4 address blocks are scarce, making darknet deployments unlikely in most networks. To overcome this obstacle, we will leverage assigned IPv4 addresses that have no or low traffic in SDSC's network to expand the coverage of UCSD-NT. Such networks are colloquially called *greynets*, as they have some dark addresses embedded with active ("lit") IP addresses in the same address space [13]. Previous work has opportunistically identified unused addresses [13], but we propose to use three types of addresses to form greynets:

I. Network address refers to the first IP address in the entire subnet [38], which could represent a subnet (except /31 and /32 networks which use their 1-2 addresses for point-to-point connections [39]), or an IP broadcast address. By default, routers discard traffic to these network addresses [38]. Therefore, the characteristics of network addresses are similar to darknet, *i.e.*, we do not expect any traffic to originate from or to these addresses.

II. Broadcast address is the last IP address in a subnet [38] intended for IP broadcasts. Currently, only two protocols (DHCP and BOOTP) use IP broadcast addresses within internal networks. Therefore, packets from the Internet toward broadcast addresses are likely unsolicited.

III. Equipment address is assigned to physical or virtual network devices, such as router interfaces, and an endpoint of a point-to-point connection. These addresses might host management services (e.g., SSH, Telnet, SNMP), send/receive routing messages (e.g., OSPF and STP), and respond to ICMP traffic for network diagnosis, but should not carry any application traffic. Furthermore, these management services are often restricted to be accessible only within internal networks. Therefore, ingress traffic from the Internet is also likely unsolicited.

Using greynets as network telescopes not only saves unused address space, but is less likely to be circumvented (avoided) by attackers for two reasons. First, the unused addresses are deeply embedded in the production network that hosts services or end-users [40]. Second, Classless Inter-Domain Routing (CIDR) [41] defines subnets with arbitrary prefix length; greynet addresses could be at any location in the subnet. The last octets of the addresses are not limited to 0 or 255 in the traditional class A/B/C addressing scheme. SDSC uses CIDR subnets, which obfuscates attempts to identify greynet addresses. Analyzing a week of traffic to each IP in one /24 in the UCSD-NT (Fig. 2), we found no evidence that scanners avoid probing any address in the subnet.

Subnet assignments could change over time. We plan to develop scripts to periodically download and parse network configurations from routers to identify greynet IPv4 addresses. The updated list of greynet addresses will automatically transmit to SDSC's traffic aggregator which is configured to extract and forward traffic from the unutilized addresses. We have identified dozens of greynet IPv4 addresses in SDSC's production network. The corresponding subnets connect multiple CIs and UCSD's campus network.

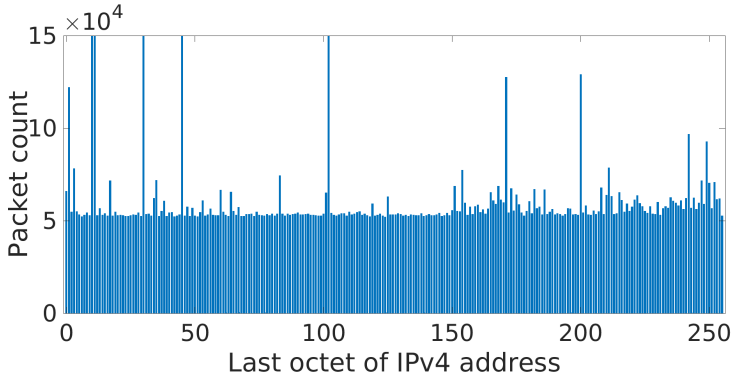


Figure 2: Number of packets send to last octet of IP addresses in a /24 darknet. The first (0) and the last (255) IP addresses in a /24 received a similar number of IBR packets as most other IP addresses in the same subnet, indicating they will provide sound IBR data in greynets.

SDSC will deploy 400Gbps fiber optics links in mid-2023 (Fig. 3). We plan to update SDSC's network taps and traffic aggregator to provide traffic visibility after the upgrade. We will purchase an Arista 7280 series Layer 3 switch and software to support traffic capture and filtering from multiple 400Gbps network taps. The switch will filter traffic to greynet IP addresses identified from the router configurations and forward it to UCSD-NT's packet capture server.

To handle traffic growth, we will upgrade three main components of the UCSD-NT:

1. Packet capture server. We will upgrade the packet capture server and develop new software for processing packets. As the Endace DAG card that UCSD-NT currently uses is no longer

available, the new packet capture server will use programmable SmartNICs (e.g., NVIDA BlueField series [42]) to offload packet processing and filtering from the server's CPU for better performance and reliability. We will use Capsule [43], a Rust-based network function framework that provides an efficient and memory safe programming environment for packet processing, to implement new packet processing functions to run on the SmartNICs.

2. Compute cluster. The computational power of existing VMs that we provide to researchers for data analysis will soon become insufficient. Each research VM currently has 8 CPU cores, 32GBytes RAM and 100GBytes storage. But the processing time of FlowTuple files in the VMs is longer than the time duration that the files cover, inhibiting real-time analysis. (It takes more than an hour to process an hour of data.) Furthermore, our compute cluster is fully utilized at 30 researchers. We plan to purchase new compute servers to offer higher performance VMs to more researchers.

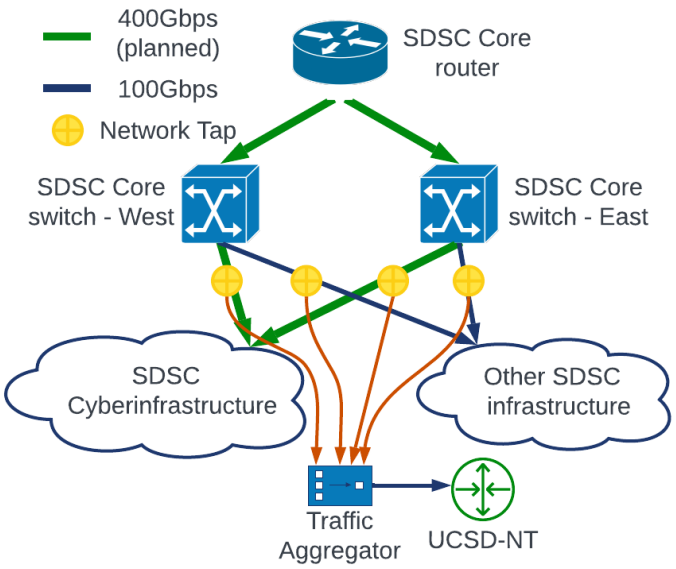


Figure 3: Overview of SDSC core network. SDSC plans to upgrade the network links connected to the CI to 400Gbps (green links). We will use 400Gpbs-compatible network taps and traffic aggregator to passively monitor ingress traffic from the core router to SDSC's networks and forward it to UCSD-NT.

3. Storage. We plan to increase the capacity of our object storage cluster to accommodate the new traffic from UCSD/SDSC and growth of UCSD-NT IBR traffic. We currently provide the most recent 60 days of raw packet traces and all historical FlowTuple files for offline analysis (§3.2). As of February 2023, 60 days of packet traces consume over 200TBytes (and growing). UCSD-NT also produces ≈22TBytes of FlowTuple files per year. As the number of SDSC greynet addresses is much smaller than the darknet, we expect packet and flow data for the new greynet to consume $<$ 1TBytes per year.

3.2 Task 2: Scalable anomaly detection for IBR traffic data

The volume and velocity of historical and live IBR traffic data remains challenging for near real-time classification and analysis of network events. Pre-trained models can reduce the detection speed. However, these models may not be able to quickly adapt to emerging attacks.

We plan to apply our recently developed ML-based framework [14] to detect events in IBR traffic by characterizing traffic dynamics across many time series generated from raw traffic processed by the Corsaro software package every minute (Table 1). Our method has four steps (*i-iv* in Fig. 4).

3.2.1 Time series data preparation

The first step (*i* in Fig. 4) is to query time series from UCSD-NT's InfluxDB, which indexes the data with traffic properties (§2.1 and Table 1). Each combination of values in traffic properties

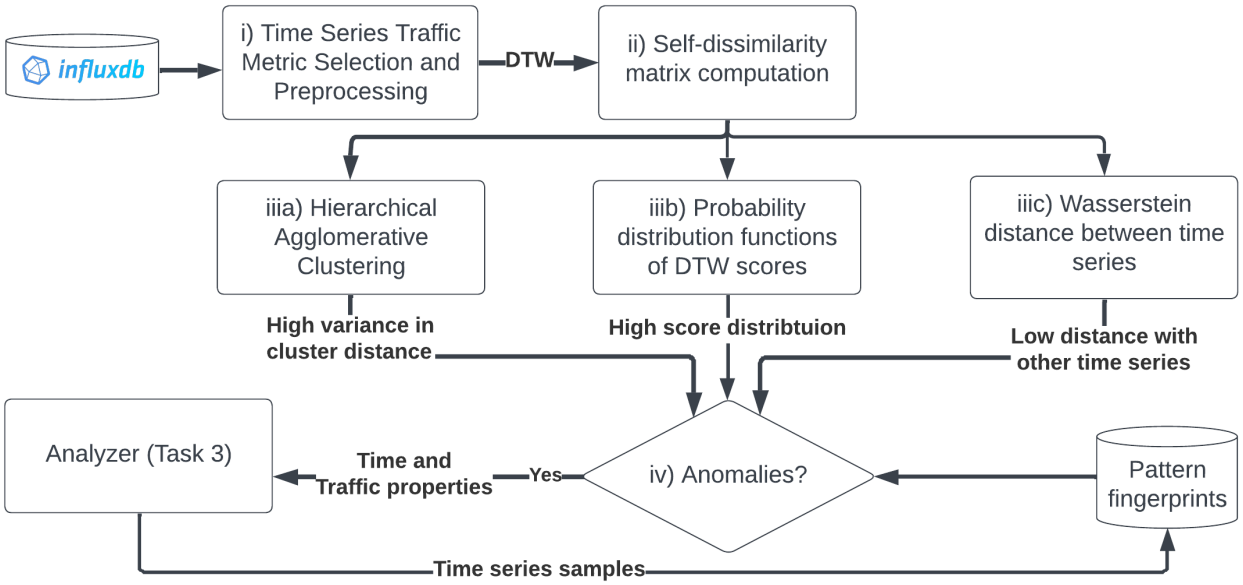


Figure 4: Our anomaly detection method leverages both ML and statistical analyses of time series.

(e.g., country and protocol) and spoofed filters generates 5 time series (e.g., packet per minute). For example, the packet count (PPM) of 500 popular TCP ports of non-spoofed/spoofed traffic from 23 North American countries will result in 11,500(=500×23) matrices. We will query 2-week time series data for all combination of values in traffic properties and traffic metrics from the InfluxDB and transfer to Expanse for analysis.

3.2.2 Self-dissimilarity matrix computation

Our method extracts signals of attacks in time-series statistics that can reveal promising time periods to further investigate (§3.3.2). The main challenge is to detect events from the heterogeneous patterns in the time series. Instead of training models to classify different patterns, we use Dynamic Time Warping (DTW) [44] to quantify similarity between two time-series (*ii* in Fig. 4).

The DTW method compares time series of different lengths or that have been distorted by time-shifting or warping. In such scenarios, using a simple Euclidean distance between time series elements will suffer corruption from noise, e.g., phase-shifts, in the data. DTW overcomes this problem by computing the optimal alignment between two time series that minimizes differences in their shape and timing. In sum, given two time-series Q and S, DTW calculates a dissimilarity score, i.e., squared Euclidean Distance, between each value in Q and *all* other values in S, to form

	(-)	S_0	S_1	S_2	S_3
(-)	0	∞	∞	∞	∞
Q_0	∞	0	16	17	17
Q_1	∞	0	16	17	17
Q_2	∞	16	0	9	25
Q_3	∞	17	9	0 → 1	

Q_0	Q_1	Q_2	Q_3
1	1	5	2

S_0	S_1	S_2	S_3
1	5	2	1

Figure 5: Left: A fully-computed DTW cost-matrix (shortest-cost path in red) between two time series. Right: Sample values for Q and S. DTW captures the high similarity of time series despite that Q and S are not temporally aligned.

a *cost matrix* (Fig. 5) of dissimilarity scores. DTW then computes the overall dissimilarity score as the optimal alignment cost between Q and S, i.e., the least cost path, through the computed cost matrix.

We use DTW to compute a *self-dissimilarity matrix*, $\mathbf{M}$, of each time-series. We partition each time-series, $\mathcal{T}$, into fixed length segments (i.e., $\mathcal{T} = T_0, T_1, ..., T_n$). We then compute DTW dissimilarity score between any two normalized time-series segments to create the matrix (Fig. 6).

$$\mathbf{M} = \begin{bmatrix} DTW(T_0, T_0) & DTW(T_0, T_1) & \dots & DTW(T_0, T_{N-1}) \\ DTW(T_1, T_0) & DTW(T_1, T_1) & \dots & DTW(T_1, T_{N-1}) \\ \vdots & \vdots & \ddots & \vdots \\ DTW(T_{N-1}, T_0) & DTW(T_{N-1}, T_1) & \dots & DTW((T_{N-1}, T_{N-1}) \end{bmatrix}$$

Figure 6: Formation of self-dissimilarity matrix, $\mathbf{M}$, of time-series $\mathcal{T}$, partitioned into N segments. The matrix is symmetric with zero diagonal (blue). Therefore, we consider the lower triangular part of the matrix (red).

3.2.3 Machine-learning and statistical-based analysis

The self-dissimilarity matrices may be noisy. The third step (*iiia-iiic* in Fig. 4) is to apply both *machine-learning* and *statistical* approaches to analyze the self-dissimilarity matrices and identify anomalous subintervals. We will apply hierarchical agglomerative clustering (HAC) to group time-series segments with similar patterns (*iiia* in Fig. 4). Large-size clusters likely contain periodic or recurring activities. However, HAC's results may be over-sensitive, particularly when all values in the matrix are low. To overcome this limitation, we will use the distance between HAC clusters to infer evidence of anomalies. For example, transient pattern changes significantly increase the distance between adjacent clusters. Furthermore, we will examine probability distribution functions (PDFs) of the values in self-dissimilarity matrices (*iiib* in Fig. 4). Bimodal or multimodal distributions in the matrices indicate a change in behavior.

New attacks often only target certain services (ports) and thus only change the traffic patterns of a few time series. We will use the Wasserstein distance [45] metric to compare self-dissimilarity matrices between time series to detect anomalies (*iiic* in Fig. 4). Low Wasserstein distance between two matrices indicates correlated changes in multiple protocols, ports, countries and networks that we should analyze as a whole (§3.2.6).

3.2.4 Anomaly detection

We will combine the signals from Step *iiia-c* to identify time series that might have anomalies, and feed them to our automatic analyzer (§3.3). In addition, we will build a fingerprint database to store samples of time series segments that we have analyzed and labeled. Consulting this database will reduce the number of alerts to known changes in the traffic data.

To achieve near real-time detection, we will update the self-dissimilarity matrices when UCSD-NT creates a new time series segment (a configuration knob currently set for 6 hours per segment). The update process will be quick as it only involves comparing new segments with old ones.

3.2.5 A scalable implementation

Our top-down approach (starting with statistical time series, then flows, then raw packets if needed) requires substantially fewer computational resources than analyzing raw traffic traces. However, comparing many thousands of time series is still computationally intense. We will leverage Expanse, a high-performance computing (HPC) system at SDSC, to compute dissimilarity matrices and perform machine-learning tasks. We have received an Accelerate ACCESS allocation [46] to evaluate and deploy our methods. To fully utilize this allocation, we will query the time series data from UCSD-NT's InfluxDB with CAIDA's compute servers and prepare the time series data for analysis. We will use Dask [47] to distribute compute tasks on SDSC Expanse. The bottleneck of our analysis is computing DTW scores. We will seek to use a GPU implementation of DTW (e.g., cuDTW++ [48]) to reduce computation time.

3.2.6 Preliminary results

We ran pilot tests using 20 days (November 1-20, 2022) of packet per minute (PPM) and bytes per minute (BPM) time series data for traffic from 255 countries and 256 IPv4 protocols under both raw and filtered data (removing likely spoofed traffic), totaling 2044(=2×2×(255+256)) unique time series. We partitioned the time series into 6-hour segments. We computed the self-dissimilarity matrices on a compute box with 32 CPU cores in less than 30 minutes.

Our cross-matrix analysis (Fig. 7) showed there were similar traffic patterns in eight IP protocols, four of which were associated with the Cisco IPv4 Blocked Interface Exploit, present in older versions of Cisco IOS. Carefully crafted packets with these protocol values cause vulnerable routers to incorrectly interpret the destination interface's packet queue as full, resulting in refusal of subsequent packets to the interface [49, 50].

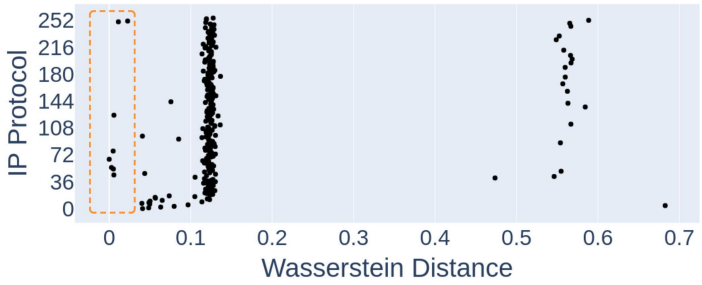
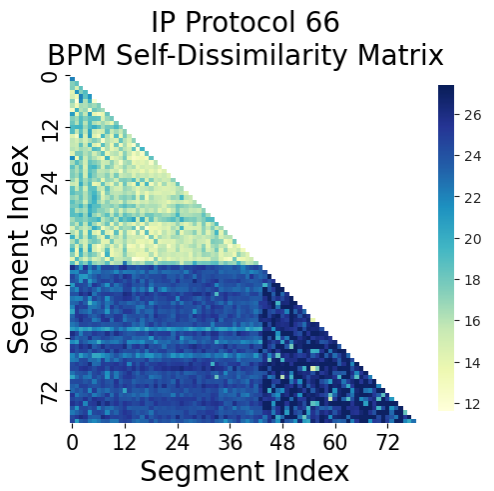


Figure 7: Wasserstein distances between self-dissimilarity matrix of protocol 66 and other IP protocols. Seven protocols had high similarity in traffic volume patterns with protocol 66 (orange rectangle), suggesting a correlated scan.

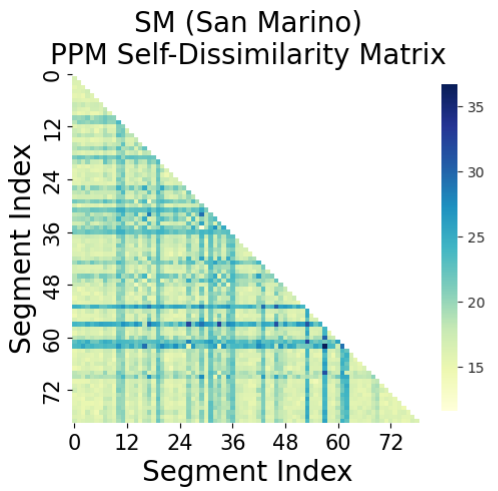
Fig. 8 shows heatmaps of two self-dissimilarity matrices. The darker colors indicate high dissimilarity with other time series segments. We found persistent (Fig. 8a) and temporary (Fig. 8b) changes in time series patterns, both of which could indicate new scanning activities.

3.3 Task 3: Automatic anomaly analysis and threat intelligence generation

Our third task is to leverage the detection results we obtained in Task 2 to accelerate and automate network traffic and flow analyses. We will develop an automatic analyzer to refine the alerts and provide additional information on the potential threats. We will generate threat intelligence in MISP-compatible format [51] to facilitate information sharing with network operators.



(a) Bytes per minute of IP protocol 66 traffic. The dissimilarity score abruptly increased after November 11, 2022, indicating a shift in traffic pattern.



(b) Packets per minute whose source geolocated from San Marino. Four anomalies (segments 53, 57, 61, 62) show high dissimilarity (score > 25) to other segments.

Figure 8: Two Heatmaps of self-dissimilarity matrices characterize two different IBR traffic metrics. Each segment index is a 6-hour time period. The depth of the color represents the dissimilarity score. Both matrices showed sudden changes in characteristics in the time-series.

3.3.1 Enriching FlowTuple data format to accelerate network flow analysis

UCSD-NT generates FlowTuples every 5 minutes, providing a summary of network flows observed in that time interval (Table 2). This information enables characterization of various types of malicious traffic, including scanning campaigns, without the overhead of analyzing raw packets. We will improve the FlowTuple format by providing more information and annotation.

1. *Hostnames* contain rich information about IP geolocation [52] and ASNs [53]. We will use reverse DNS to resolve source IPs to hostnames at the time UCSD-NT creates the FlowTuples.
2. *Acknowledged benign scanners* scan the Internet for research studies or cybersecurity monitoring. They are less likely to be malicious. We will use public collections of known-benign scanning IPs (e.g., [54]) and information provided by the scanner's websites (e.g., [55]) to identify benign scanners by source IPs and hostnames.
3. *Scanner implementation* provides crucial information about the nature of the traffic. The existing FlowTuple format does not provide TCP/IP header values to enable use of heuristics [56] to infer scanner implementation (e.g., Hajime, Zmap, Mirai). Including more header values in FlowTuple will significantly increase the file sizes. Instead, we will compute the heuristics at the time of generating the FlowTuple, and add the inference as a new tag in FlowTuple.
4. *Packet payload samples* could help identify the target and intention of the traffic, such as services or vulnerabilities. Providing representative samples can accelerate analysis of anomalies.

3.3.2 Automating flow analysis on detected anomalies

We will use the enhanced version of FlowTuple files to analyze anomalies we detect in darknet and greynet traffic summary time series. We will build an analyzer to extract and summarize common network characteristics of the anomalies, such as the ports, origin ASNs, countries, and network ser-

Table 2: Summary of current information and proposed new features (bolded) for FlowTuple [21].

Categories	Information
Time	Timestamp of network flows
Network flow information	Source IPs, destination prefix, destination ports, IP Protocol
Summary of traffic properties	destination IPs, TTLs, (TCP/UDP) source ports, TCP flags, packet sizes, TCP flags, **sample payload**
Source IP Annotation	Source IP geolocation (Maxmind and NetAcuity), Prefix-to-AS, **hostname**
Inference	Spoofed packets, Sent using Masscan/**Hajime/Zmap/Mirai**, **Acknowledged scanners**

vices. We will also filter false positives and activities irrelevant to SDSC's CI. For example, changes in scanning patterns from acknowledged scanners are unlikely to cause cybersecurity threats. We will also filter spoofed packets since they have unreliable source IP information.

After screening out these known components of the anomalous traffic, we will classify whether the remaining anomalies are *targeted* or *Internet-wide* activities. Our approach is to match the flow characteristics of anomalies detected by the darknet and greynet telescopes. Internet-wide scanning will likely trigger our detection in both traffic streams. But only the greynet telescope can observe scanning targeted to SDSC's CI.

3.3.3 Publishing threat intelligence

STARNOVA will report the detected anomalies with network flow information for further analysis by SDSC network security experts. We will deliver alerts by email and Mattermost/Slack bots [57]. To facilitate sharing of information with network security communities, we will share the alerts using the MISP core format [51] used to exchange threat intelligence. The format pre-defines a standard list of attributes (e.g., source/destination IPs) and types of network activity. We will publish MISP-formatted data feeds on CAIDA's website, so other networks can use our results.

4 Software licensing approach and justification

The released software produced in the course of the project will have an open source license. Corsaro [20], the software UCSD-NT currently runs, is released using UCSD's Academic Non-Commercial License [58]. We plan to retain this license for future enhancement of this software and related products. For software that we jointly develop with Northwestern University, we will release the software with a license acceptable to both institutions [59], with the goal of maximizing the benefit to the cyberinfrastructure research community to reuse, extend, and continue to share the products we develop.

5 Data Sharing approach

We apply the FAIR principles [60], a guideline for those wishing to enhance the reusability of their data holdings, to conduct and evaluate our data sharing model.

Findable: CAIDA's resource catalog [61] enables users to search and navigate CAIDA's datasets, tools, and related publications. We index individual datasets derived from UCSD-NT data in

the catalog. For traffic and flow data, we organize the data files in UCSD-NT's object storage system by time, such that users can easily locate and download the data.

Accessible: As the data could be sensitive (§6), we require users to apply for data access via CAIDA's website. We will create accounts for accepted users to access the data via their VMs. For SDSC greynet data, we will anonymize SDSC's IP addresses.

Interoperable: We store the traffic data using pcap format, a *de facto* standard for this type of data. The FlowTuple files are based on open-source Apache Avro format. Many open-source libraries in different programming languages can parse the files. We provide the schema of FlowTuple data and sample code on our website [21]. We will publish our threat intelligence using MIST format, which is also an open-source data format with rich documentation.

Reusable: We provide detailed documentation on how the data was captured at the UCSD-NT. We also store historical data files (e.g., traffic filters, and IP geolocation databases) to trace the validity of traffic and relabel the traffic data.

6 Ethical and operational concerns

We take multiple measures to ensure that UCSD-NT does not capture user traffic from the darknet and SDSC greynet address spaces. We work closely with the network administrators of the address space (the non-profit organization that owns the space, UCSD, and SDSC) to update UCSD-NT's traffic filters as IP address assignments in the network change. Keeping the filters up to date is essential to accurately discard user traffic and thus safeguard privacy. For greynet traffic, we will configure the traffic aggregator to only forward traffic from the Internet, but not internal (LAN) and outgoing traffic, to UCSD-NT.

The telescope data may inadvertently reveal victims of ongoing DDoS attacks and vulnerable machines inflected by malware and worms. As such, we restrict the data access from the public. Researchers who wish to access to the compute VMs/traffic data must disclose their purposes, and sign an Acceptable Use Agreement (AUA) [62] to gain access the data. Our project manager/data administrator will review the access every 90 days.

7 Sustainability plan

Over the last two years, CAIDA has actively purused sustainability of the UCSD-NT infrastructure, trying to arrange service agreements to support specific institutions seeking to use the data. In 2023 we signed annual service agreements with Information Sciences Institute (ISI) at the University of Southern California, and Lincoln Laboratory at MIT. We have also received support from the Amateur Radio Digital Communications (ARDC) to support daily operation and data collection of UCSD-NT. We are also working with security companies (e.g., DomainTools) to explore commercial licensing of the darknet data in order to support academic research use of the data.

The new capability of UCSD-NT proposed in this project will bring new opportunities for us to develop new services to sustain STARNOVA beyond the NSF grant. First, the updated UCSD-NT infrastructure will be capable of extending collection of greynet traffic from other networks, including scientific research networks and commercial networks. Therefore, our anomaly detection method can provide tailor-made threat intelligence to participating networks. We will have a workshop to explore how to enable other campus CI deployments to replicate our greynet infrastructure, including ways to leverage the UCSD-NT-derived threat intelligence.

8 Quantitative evaluation metrics

We will evaluate the benefit of STARNOVA with four metrics.

1. *Detection accuracy.* Higher detection accuracy provides better protection to the CI. However, in many cases, it is hard to obtain ground truth about the intent of malicious traffic. We will verify some detection results with SDSC security analyst (senior personnel in this project) to determine whether the malicious activities are also found in traffic toward the CI and impose threats to the systems. Apart from this manual investigation, we will analyze the threats flagged by intrusion detection systems, but not STARNOVA, to refine the detection methods. We will count the number of detected threats that are correctly/incorrectly classified.
2. *Detection delay.* We will evaluate the time delay from the onset of the attacks to the generation of alerts. The accuracy of detection often increases with more data for classification. However, the benefits of alerts diminish when they are triggered after or toward the end of the attacks. STARNOVA will trade off between detection accuracy and detection delay. We will evaluate the detection delay of each alert as a metric for fine-tuning our methods.
3. *Number of users of our threat intelligence.* Apart from measuring the technical aspect of STARNOVA, we will quantify the impact of our project by the number of external users that subscribed to our threat intelligence data feed. The alerts we provided will help enhance the security of users' networks.
4. *Feedback from and uptake of ideas by other campus CI operators.* We will leverage CAIDA's MSRI GMI workshops [63], several of which are focused on traffic data and DDoS measurement efforts, to present and solicit community feedback on our design, deployment, and operational threat intelligence. These meetings are well-attended by network and security researchers, R&E operators, and commercial network operators, and will provide honest feedback on the utility of our results. We will explicitly reach out to campus CI operators to attend one of these meetings to present the status of this project in year 3.

9 Project timeline

Fig. 9 shows our timeline. As supply chain challenges persist, we will order equipment as soon as possible. In parallel, we will develop the anomaly detection method (T2.1) and implement the new FlowTuple version (T3.1). We will deploy all hardware in the first year and receive greynet traffic from SDSC before Q4 2024 (T1.1-1.4). We will present a low-fidelity prototype (M1) upon the deployment of network taps and initial method for anomaly detection. We will refine the method (T2.2-2.4), improve the analysis (T3.2-3.3), and transition the UCSD-NT to the new infrastructure (T1.5). A working high-fidelity prototype will be ready by the end of Y2 (M2). In Y3, we will publish threat intelligence and continue to refine the system (T3.4). In Q2 2026, we will deploy a running demo of STARNOVA (M3).

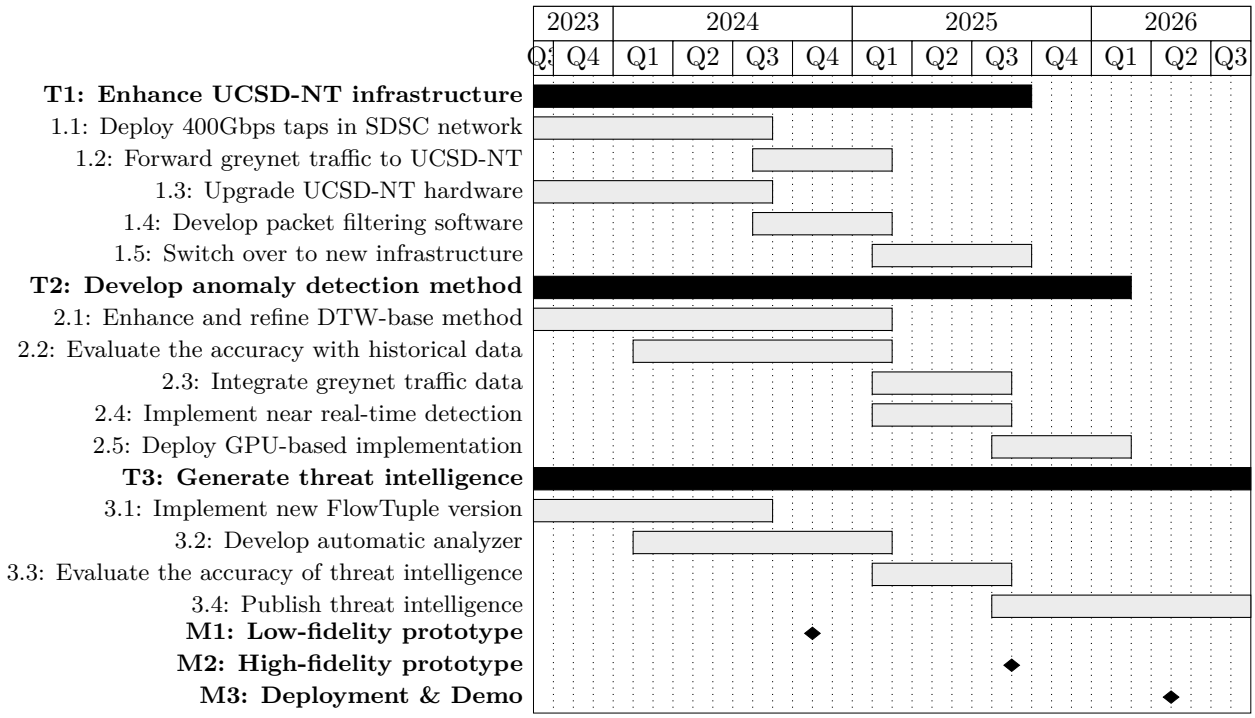


Figure 9: Timeline of project tasks and milestones.

References

[1] R. Sommese, K. Claffy, R. van Rijswijk-Deij, A. Chattopadhyay, A. Dainotti, A. Sperotto, and M. Jonker, "Investigating the Impact of DDoS Attacks on DNS Infrastructure," in *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, (New York, NY, USA), pp. 51–64, Association for Computing Machinery, 2022.

[2] M. Jonker, A. King, J. Krupp, C. Rossow, A. Sperotto, and A. Dainotti, "Millions of Targets under Attack: A Macroscopic Characterization of the DoS Ecosystem," in *Proceedings of the 2017 Internet Measurement Conference*, IMC '17, (New York, NY, USA), pp. 100–113, Association for Computing Machinery, 2017.

[3] E. Balkanli, J. Alves, and A. N. Zincir-Heywood, "Supervised learning to detect DDoS attacks," in *2014 IEEE Symposium on Computational Intelligence in Cyber Security (CICS)*, pp. 1–8, 2014.

[4] E. Balkanli, A. N. Zincir-Heywood, and M. I. Heywood, "Feature selection for robust backscatter DDoS detection," in *2015 IEEE Local Computer Networks Conference Workshops*, pp. 611–618, 2015.

[5] A. Dainotti, C. Squarcella, E. Aben, K. C. Claffy, M. Chiesa, M. Russo, and A. Pescapé, "Analysis of Country-Wide Internet Outages Caused by Censorship," *IEEE/ACM Trans. Netw.*, vol. 22, pp. 1964–1977, dec 2014.

[6] R. Padmanabhan, A. Filastò, M. Xynou, R. S. Raman, K. Middleton, M. Zhang, D. Madory, M. Roberts, and A. Dainotti, "A Multi-Perspective View of Internet Censorship in Myanmar," in *Proceedings of the ACM SIGCOMM 2021 Workshop on Free and Open Communications on the Internet*, FOCI '21, (New York, NY, USA), pp. 27–36, Association for Computing Machinery, 2021.

[7] D. Moore, V. Paxson, S. Savage, C. Shannon, S. Staniford, and N. Weaver, "Inside the Slammer worm," *IEEE Security & Privacy*, vol. 1, pp. 33–39, Jul 2003.
[8] A. Dainotti, A. King, K. Claffy, F. Papale, and A. Pescape, "Analysis of a "/0" Stealth Scan From a Botnet," *IEEE/ACM Transactions on Networking*, vol. 23, pp. 341–354, apr 2015.
[9] M. Antonakakis, T. April, M. Bailey, M. Bernhard, E. Bursztein, J. Cochran, Z. Durumeric, J. A. Halderman, L. Invernizzi, M. Kallitsis, D. Kumar, C. Lever, Z. Ma, J. Mason, D. Menscher, C. Seaman, N. Sullivan, K. Thomas, and Y. Zhou, "Understanding the Mirai Botnet," in *Proceedings of the 26th USENIX Conference on Security Symposium*, SEC'17, (USA), pp. 1093–1110, USENIX Association, 2017.
[10] 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.
[11] M. S. Pour, J. Khoury, and E. Bou-Harb, "HoneyComb: A Darknet-Centric Proactive Deception Technique For Curating IoT Malware Forensic Artifacts," in *Proceedings of IEEE/IFIP Network Operations and Management Symposium*, Apr 2022.
[12] 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.
[13] W. Harrop and G. Armitage, "Defining and Evaluating Greynets (Sparse Darknets)," in *Proceedings of IEEE Conference on Local Computer Networks*, 2005.
[14] M. Gao, R. K. P. Mok, and claffy k, "A Scalable Network Event Detection Framework for Darknet Traffic," in *Proceedings of the 22nd ACM Internet Measurement Conference*, IMC '22, (New York, NY, USA), pp. 738–739, Association for Computing Machinery, 2022.
[15] SDSC, "Expanse,." `https://www.sdsc.edu/services/hpc/expanse/`.
[16] SDSC, "Voyager,." `https://sdsc.edu/support/user_guides/voyager.html`.
[17] SDSC, "Science Portals & Science Gateways,." `https://www.sdsc.edu/services/hpc/science_gateways.html`.
[18] SDSC, "National research platform,." `https://www.sdsc.edu/services/hpc/nrp/index.html`.
[19] N. E. R. S. C. Center, "HPSS data archive." `https://www.nersc.gov`.
[20] "Corsaro." `https://catalog.caida.org/media/2012_dust_corsaro`, 2012.
[21] CAIDA, "FlowTuple." `https://stardust.caida.org/docs/data/flowtuple/`, 2021.
[22] CAIDA, "RouteViews IPv4 Prefix to AS mappings." `https://catalog.caida.org/details/dataset/routeviews_ipv4_prefix2as`, 2021. Accessed: 2022-8-10.
[23] A. Dainotti, K. Benson, A. King, k. claffy, M. Kallitsis, E. Glatz, and X. Dimitropoulos, "Estimating Internet Address Space Usage through Passive Measurements," *SIGCOMM Comput. Commun. Rev.*, vol. 44, pp. 42–49, dec 2014.
[24] influxdata, "InfluxDB." `https://www.influxdata.com`.
[25] Grafana Labs, "Grafana,." `https://grafana.com/grafana/`.
[26] CAIDA, "STARDUST Grafana dashboard,." `https://explore.stardust.caida.org`.
[27] P. Richter and A. Berger, "Scanning the Scanners," in *Proceedings of ACM Internet Measurement Conference*, Oct 2019.
[28] D. Moore, C. Shannon, D. J. Brown, G. M. Voelker, and S. Savage, "Inferring Internet Denial-of-Service Activity," *ACM Trans. Comput. Syst.*, vol. 24, pp. 115–139, May 2006.
[29] Z. Durumeric, M. Bailey, and J. A. Halderman, "An Internet-wide view of Internet-wide scanning," in *Proceedings of USENIX Security Symposium*, 2014.

[30] O. Gupta, "Identifying traffic anomalies interfering with IBR based outage detection," Master's thesis, University of California San Diego, 2018.
[31] A. Guillot, R. Fontugne, P. Winter, P. MÃlrindol, A. King, A. Dainotti, and C. Pelsser, "Chocolatine: Outage Detection for Internet Background Radiation." https://catalog.caida.org/paper/2019_chocolatine.
[32] D. Cohen, Y. Mirsky, M. Kamp, T. Martin, Y. Elovici, R. Puzis, and A. Shabtai, "DANTE: A framework for mining and monitoring darknet traffic," in *Proceedings of European Symposium on Research in Computer Security*, pp. 88–109, 2020.
[33] Duke University, "STINGAR: Threat intelligence for higher education." https://stingar.security.duke.edu.
[34] SANS, "Internet Storm Center,." https://isc.sans.edu.
[35] IntelOwl Project Org, "GreedyBear,." https://github.com/intelowlproject/GreedyBear.
[36] Greynoise, "Turning internet noise into intelligence.,." https://www.greynoise.io.
[37] R. Hiesgen, M. Nawrocki, A. King, A. Dainotti, T. C. Schmidt, and M. W ahlisch, "Spoki: Unveiling a New Wave of Scanners through a Reactive Network Telescope," in *Proceedings of USENIX Security Symposium*, 2022. Accessed: 2023-2-14.
[38] F. Baker, "Requirements for IP version 4 routers." https://datatracker.ietf.org/doc/html/rfc1812, Jun 1995.
[39] A. Retana, R. W. V. Fuller, and D. McPherson, "Using 31-bit prefixes on IPv4 point-to-point links." https://datatracker.ietf.org/doc/html/rfc3021, Dec 2000.
[40] L. Miao, W. Ding, and H. Zhu, "Extracting Internet Background Radiation from RawTraffic Using Greynet," in *Proceedings of IEEE International Conference on Networks*, 2012.
[41] V. Fuller and T. Li, "Classless inter-domain routing (CIDR): The internet address assignment and aggregation plan." https://datatracker.ietf.org/doc/html/rfc4632, Aug 2006.
[42] NVIDIA, "NVIDIA BlueField data processing units." https://www.nvidia.com/en-us/networking/products/data-processing-unit/, 2023.
[43] "Capsule,." https://github.com/capsule-rs/capsule.
[44] D. J. Berndt and J. Clifford, "Using Dynamic Time Warping to Find Patterns in Time Series," in *Proceedings of the 3rd International Conference on Knowledge Discovery and Data Mining*, AAAIWS'94, pp. 359–370, AAAI Press, 1994.
[45] L. V. Kantorovich, "Mathematical Methods of Organizing and Planning Production," *Management Science*, vol. 6, pp. 366–422, Jul 1960.
[46] ACCESS, "Advanced cyberinfrastructure coordination ecosystem: Services & support." https://allocations.access-ci.org, 2023.
[47] Dask Development Team, *Dask: Library for dynamic task scheduling*, 2016.
[48] B. Schmidt and C. Hundt, "CuDTW++: Ultra-Fast Dynamic Time Warping on CUDA-Enabled GPUs," in *Proceedings of Euro-Par 2020: Parallel Processing*, pp. 597–612, 2020.
[49] K. Folga, "Cisco IOS from an Attacker's Point of View." http://www.windowsecurity.com/uplarticle/NetworkSecurity/ios_en.pdf, 2005.
[50] C. Johnson, "The Cisco IPv4 Blocked Interface Exploit," 2003. https://www.giac.org/paper/gcih/655/cisco-ipv4-blocked-interface-exploit/106708.
[51] MISP Threat Sharing project, "MISP - The open source threat intelligence sharing platform." https://www.misp-project.org, 2023.
[52] M. Luckie, B. Huffaker, A. Marder, Z. Bischof, M. Fletcher, and K. Claffy, "Learning to extract geographic information from internet router hostnames," in *Proceedings of ACM CoNEXT*, 2021.
[53] M. Luckie, A. Marder, M. Fletcher, B. Huffaker, and K. Claffy, "Learning to extract and use ASNs in hostnames," in *Proceedings of ACM IMC*, 2020.

[54] M. Collins, "Acknowledged Scanners." https://gitlab.com/mcollins_at_isi/acknowledged_scanners, 2021.
[55] Censys, "Opt Out of Data Collection." https://support.censys.io/hc/en-us/articles/360043177092-Opt-Out-of-Data-Collection, 2022.
[56] A. Tanaka, C. Han, T. Takahashi, and K. Fujisawa, "Internet-wide scanner fingerprint identifier based on TCP/IP header," in *Proceedings of IEEE International Conference on Fog and Mobile Edge Computing*, 2021.
[57] Mattermost, "Using bot accounts,." https://developers.mattermost.com/integrate/reference/bot-accounts/.
[58] UC San Diego, "Copyright Overview,." https://innovation.ucsd.edu/disclose-patent/copyright.html.
[59] UCOP, "UC Guide to Managementing Open Source Software." https://security.ucop.edu/files/documents/resources/guide-to-managing-open-source-software.pdf.
[60] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C. T. Evelo, R. Finkers, A. Gonzalez-Beltran, A. J. Gray, P. Groth, C. Goble, J. S. Grethe, J. Heringa, P. A. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S. J. Lusher, M. E. Martone, A. Mons, A. L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S.-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M. A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao, and B. Mons, "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, mar 2016.
[61] CAIDA, "Resource catalog,." https://catalog.caida.org.
[62] CAIDA, "CAIDA master acceptable use agreement (AUA)." https://www.caida.org/about/legal/aua/.
[63] CAIDA, "CAIDA Global Measurmeent Infrastructure (GMI) Meetings." https://www.caida.org/workshops/, 2023.