CIRC: New: iVoyager: Internet Voyager for gathering cyber threat intelligence

1 Introduction

For years, researchers have deployed network telescopes on unused IPv4 address spaces [1,2] and/or public cloud infrastructure [3,4] to passively capture incoming unsolicited traffic, known as *Internet Background Radiation (IBR)* [5,6], to identify victims of denial-of-service attacks [7–11] and malicious Internet activities [12–16]. With consistent U.S. government support, including the now-concluded CI-SUSTAIN project (CNS-1730661), CAIDA has sustained operation of the world's largest IPv4 network telescope for over two decades, helping hundreds of CISE researchers to study Internet-wide cybersecurity incidents and produce datasets for cybersecurity education.

However, this passive approach has limitations, as it captures only a few types of security events, and malicious actors evolve their tactics to evade detection. Researchers have also deployed *honeypots* (e.g., [17–22]) which react to unsolicited traffic to lure further engagement by attackers, yielding attack fingerprints, victim identification, and malware samples. Both telescope and honeypots face a daunting challenge with the growing use of IPv6. Scanning the vast IPv6 address space, or even small networks, is practically infeasible, so scanners must strategically target likely-active networks, which requires innovative algorithms to generate target "hitlists" (e.g., [23–29]). Furthermore, most existing honeypot implementations support one IPv4 address per instance. Running multiple instances to monitor a large blocks of IPv6 addresses is resource-prohibitive.

We propose iVoyager, a transformative cyberinfrastructure (CI) designed to enable CISE researchers to effectively explore the landscape of Internet threats by scalably gathering cyber threat intelligence. Specifically, iVoyager will provide three capabilities:

1. A flexible virtualized environment for researchers to facilitate the rapid development and scalable deployment of distributed **dual-stack (IPv4 and IPv6) telescopes and honeypots**.
2. A **proactive IPv6 telescope** that applies novel techniques to attract malicious IPv6 network activities, overcoming the visibility limitations of previous attempts.
3. Deployment of **geographically distributed dual-stack vantage points** in public clouds.

We will operationalize our reference design of iVoyager to collect longitudinal datasets that include baseline IPv6 IBR, practively-triggered IPv6 IBR, and geographically distributed IBR, as well as active measurements of topology and latency from our VPs to the senders of IBR. These datasets will enable better characterization of IP spoofing and other malicious Internet activities, but also facilitate use of machine learning/artificial intelligence (ML/AI) for cyber threat hunting, anomaly detection, and malware analysis. Importantly, iVoyager will not replace existing network telescopes or honeypots — it will complement these CIs by providing additional datasets for more comprehensive cyber threat analysis.

This project aligns with CIRC's goal to provide *new research opportunities for a broad community pursuing a focused research agenda.* Successful execution of iVoyager will overcome engineering challenges, navigate policy barriers to data-sharing, and foster a robust community advancing both network security research and operational IPv6 cybersecurity expertise. The datasets we collect will facilitate the use of ML/AI in addressing cybersecurity and Internet measurement challenges.

2 Research opportunities and community need

The success of iVoyager will unleash new research opportunities by bridging data and infrastructure gaps in broad research areas of cybersecurity, Internet topology, and network anomaly detection in IPv6 networks. Infrastructure to support IPv6-related infrastructure research has fallen significantly behind the adoption of IPv6. Notably, the research community has no scalable and longitudinal deployment of IPv6 network telescopes or honeypot infrastructure. Research thus far has used either one-off small-scale temporary deployments (e.g., [6, 30, 31]) or private datasets collected but not shared by industry [32, 33].

New IPv6 functions and configuration idiosyncracies introduce security and privacy risks [34,35]. Lack of instrumentation to provide situational awareness in IPv6 networks thus poses imminent threats to national security [36].

iVoyager will create a transformative impact on CISE community by:

1. providing *comprehensive, flexible, and scalable* infrastructure to conduct both IPv4 and IPv6 experiments on unsolicited traffic,
2. collecting *publicly accessible longitudinal* datasets to capture trends and events of the Internet,
3. offering tools and software for researchers to effectively use and integrate data collected by iVoyager and other CAIDA Internet measurement cyberinfrastructure on NSF-funded high-performance computing resources.

Six research teams investigating a broad range of CISE-related research topics have committed to early use of the infrastructure, datasets, or services provided by iVoyager, for research they would not otherwise be able to undertake. The three datasets we will collect will provide high-quality data for network anomaly detection and distributed denial-of-service (DDoS) attacks characterization. The unique ability to conduct *both* active and passive measurements on iVoyager will enable researchers to perform realistic evaluations of IPv6 target generation algorithmsand effective IPv6 blocklist generation.

3 Infrastructure Description

3.1 Fundamental infrastructure

This project will design and implement the iVoyager distributed infrastructure (Figure 1), consisting of two parts: iVoyagerStar and iVoyagerSatellite. *iVoyagerStar* is a comprehensive platform that supports researchers to deploy *complex* cybersecurity experiments on unused address blocks, including honeypots and malware tracking tools, and collect data across different layers of the Internet protocol stack (e.g., BGP, DNS, and NTP) to identify potential information sources of scanners. We will deploy *iVoyagerSatellite* in geographically distributed cloud locations to capture IBR and conduct *lightweight* topology measurements, which will inform analysis of collected traffic. Geographical diversity will enable researchers to study how targeting of malicious activity varies across address space [3, 4].

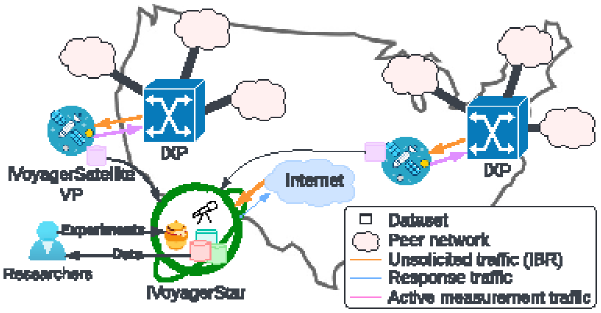

Figure 1: Overview of iVoyager infrastructure. iVoyagerStar will provide researchers with a testing ground for new experiments, while distributed iVoyagerSatellite vantage points will collect unique unsolicited traffic data from diverse locations, and support active measurements to inform analysis of such traffic.

3.2 iVoyagerStar: A dual-stack cybersecurity sensor

iVoyagerStar will allow researchers to experiment with new methods to collect data to detect Internet malware and cyberattacks. Figure 2 illustrates an overview of iVoyagerStar. We will host iVoyagerStar at the San Diego Supercomputer Center (SDSC), co-located with Expanse, a NSF-funded high performance computing resources. We will allocate virtual machines (VMs) with access to supporting services to researchers on iVoyagerStar infrastructure to analyze *live* IPv4/IPv6 IBR traffic. iVoyagerStar will comprise six major components, offering a comprehensive pipeline for experiment execution, data collection, storage, and analysis.

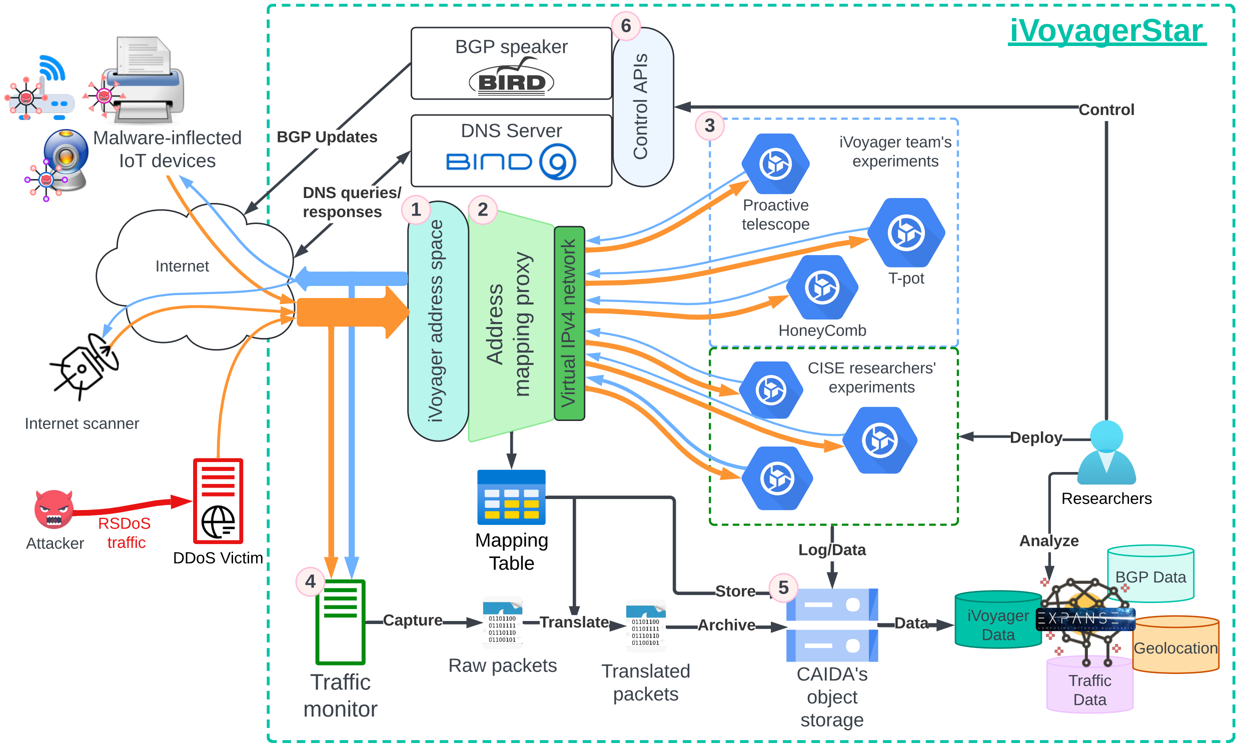

Figure 2: iVoyagerStar infrastructure. Numbers in pink circles refer to infrastructure components as numbered in the subsections of §3.2.

Component 1: Acquire and deploy IP address space for research experiments.

New, unused blocks of IP address space are essential for capturing IBR, as previously used address space can alter the behavior of scanners for extended periods of time [4]. We are in the process of acquiring a new allocation of IPv6 space from the American Registry for Internet Numbers (ARIN) to support iVoyager's users. For IPv4 address space, which is essentially unavailable from address registries, we have established a collaboration with the owner of the address space used by the UCSD network telescope (UCSD-NT), for which the owner has lent us a /22 for this three-year project.

Component 2: Implement address mapping proxy.

To lower the complexity of honeypot deployment and implementation, we will design and implement a novel address mapping proxy to dynamically translate both contiguous or non-contiguous blocks of

IPv4/IPv6 address into one private IPv4 address. Existing honeypot implementations, particularly high-interaction ones (e.g., [37, 38]), operate as system services to engage with IBR traffic [39], reusing the system's network stack. However, this approach is difficult to scale when monitoring large address blocks. One solution is to implement the tools with raw sockets or `libpcap` (e.g., [21]), which requires building an entire TCP/UDP stack rather than relying on the kernel's, thus limiting the interaction level. Our approach will allow researchers to implement experiments using conventional POSIX sockets without having to handle traffic directed toward multiple IPs. More importantly, this approach will ensure compatibility with legacy honeypot implementations, and significantly reduce the time and complexity involved in setting up new experiments.

Scaling IPv4 honeypots. For traffic toward iVoyagerStar's IPv4 address blocks monitored by one of the VMs, our proxy will forward it to the private destination IP assigned to the VM. We will leverage the source port in the IP header as the identifier, to create the mapping from the actual to the private IPs.

Scaling IPv6 honeypots. The proxy will convert all incoming IPv6 traffic into IPv4, so experimenters do not need to support IPv6. We will create a transient mapping from IPv6 source-destination addresses pair to an IPv4 source address-source port pair, using private (10/8) or reserved (240/4) IPv4 address space to avoid confusion with IPv4 traffic. Most scanning flows are brief, so timing out this mapping soon after the connection closes is critical to avoiding mapping collisions. Therefore, the proxy will monitor the state of TCP flows. Different from standard NAT64, the proxy will map traffic destined for a block of destination IPs to one IPv4 address. We will design a new algorithm that incorporates incorporates sequence numbers of TCP flows into the mapping function, to minimize the collision probability in the mapping. We will configure the proxy to handle stateless ICMP traffic, which dominates IPv6 scanning traffic. Since these flows do not require interaction from the honeypot, having the proxy reply on behalf of the honeypot will mitigate the load on the VM hosting the honeypot, improving scalability.

To enable researchers to conduct fine-grained analysis on the traffic to/from their VMs, the proxy will update the mapping table in databases, allowing iVoyagerStar to isolate packet captures for each experiment.

Component 3: Create virtualized experiment environment.

iVoyagerStar will offer researchers virtual machines (VMs) to deploy software that responds to incoming traffic. Our systems administrator will modify the configuration of the address mapping proxy (Component 2) to redirect traffic destined for specific address block(s) to the interface IP of the VM. The software can respond to the incoming traffic as if the VM were an end-host. The proxy will rewrite the IP headers according to the mapping, and send the packets back to the source.

Component 4: Build SmartNIC-base traffic monitor.

We will leverage an NVIDIA ConnectX SmartNIC to capture traffic to/from the infrastructure. Use of a SmartNIC will enable us to acquire precise packet timestamps without impairing performance. Due to the address translation mechanism, the VMs cannot directly observe actual packets exchanged with the Internet. To address this, we will provide a traffic monitor to capture traffic at the entry/exit point of iVoyagerStar (*Raw packets* in Figure 1) and use the mapping table to separate traffic to/from different VMs (*Translated packets* in Figure 1). We will store both versions of packet traces in our object storage and share them with researchers.

Component 5: Expand Object storage.

Experiments on iVoyager will generate various types of data, including activity logs and packet traces. We will expand CAIDA's storage system by 360TB to meet experimental data storage needs. CAIDA currently uses Swift, an open-source S3-compatible object storage server. Researchers will have access to virtual machines to store their data. One key advantage of local object storage for researchers is that our storage cluster is co-located with SDSC's Expanse high-performance cluster [40], which researchers can access through ACCESS-CI [41] for complex data analysis.

Component 6: Deploy peripheral services.

We will deploy BGP and DNS servers to support more complex experiments. For example, the IPv6 proactive telescope we will deploy as a reference design (§3.4.3) will require the use of domain names and issuance of TLS certificates. The DNS server will act as an authoritative name server, handling DNS queries for the domains and the zone. Issuance of TLS certificates will require the creation of temporary TXT records in the zone file. The BGP speaker will handle BGP announcements of the IPv6 prefixes of iVoyager, which will enable researchers to introduce variables in BGP as part of its experiment. We will set up an NTP server instance that can join the NTP pool, providing timing sources for the system. Additionally, users of the IPv6 NTP pool have proven to be a valuable source for collecting live IPv6 hosts [29, 30].

We will implement a REST APIs for researchers to use these peripheral services, providing control knobs in their experiments. To shed light on how activation of these services correlate with discovery of the network by scanners, we will store logs/records into the object storage.

3.3 iVoyagerSatellite: A distributed sensor infrastructure

iVoyagerStar consists of multiple vantage points (VPs) at different geographical and network loactions. Each VP can perform lightweight active measurements (e.g., traceroute) and monitor small ranges of unused addresses to capture IBR. We will deploy at least five *iVoyagerSatellite* VPs in cloud locations.

Each VP will announce a dedicated prefix that we acquired (§3.2) to simulate networks with distributed geographic presence. The maximum length of globally routable prefix for IPv4 and IPv6 is /24 and /48, respectively (i.e., 256 IPv4 addresses and 2^{80} IPv6 addresses). Since we need only a small portion of these address blocks to manage the iVoyager infrastructure and establish peering sessions, the remaining unused address space will be *dark*, i.e., unused. We will capture IBR toward these addresses to form a distributed network telescope. Additionally, we will install `scamper` [42] on each VP to perform active measurements of topology and latency.

Cloud-based deployment. Our team will leverage the expertise of Dr. Mattijs Jonker's research group at the University of Twente, which has experience setting up virtual measurement VPs on Vultr [43], a cloud provider that allows customers to make their own BGP announcements (see LoC). We will build on their expertise in using the Vultr platform for extensive measurement of anycast [44] to deploy iVoyagerSatellite VPs to capture IBR.

3.4 Tools, resources, and datasets

During the course of this project, our team will develop new open-source software tools, provide networking, compute, and storage resources for researchers, and generate publicly accessible datasets to support research on the broad topics of Internet measurement, cybersecurity, and ML/AI.

3.4.1 Software Tools

We will release three software tools designed for implementing iVoyager:

1. **Address mapping proxy.** This core software will enable the deployment of *dual-stack* network telescopes and honeypots at scale.
2. **Traffic Monitor.** Integrated into iVoyagerStar, this tool will translate packet traces in near real-time using the mapping table. We will develop new tools to support this functionality.
3. **Proactive IPv6 Telescopes.** This toolset will integrate different components of iVoyagerStar, including the BGP server, DNS server, and external TLS services.

We will support and enhance existing two community software libraries for this project:

1. **Libtrace:** A highly efficient software used for capturing packets in UCSD-NT. We will enhance it to integrate with iVoyager for high-performance packet capturing and processing.
2. **Scamper:** Widely used by the Internet measurement community for active measurements such as traceroute, alias resolution, and latency measurements. Our recent advances in domain-specific language have lowered the barrier to programming complex active measurements [45].

We will use GitHub to disseminate these tools, accompanied by usage documentation and development guides.

3.4.2 Resources

This project will offer rich and unique resources for researchers to perform both active and passive measurements. iVoyagerStar will provide researchers with **virtual machines** (VMs), capable of capturing and responding to incoming traffic in unused IP address space. Researchers will also have access to **production application services** (e.g., DNS and NTP) to build their experiments. We will offer **object storage** space in iVoyagerStar for data storage. For vetted researchers, we will provide access to **execute traceroutes** on iVoyagerSatellite.

3.4.3 Datasets

Our team will leverage iVoyager to conduct three longitudinal measurement campaigns to provide seed datasets for the research community. We will index these datasets using CAIDA's catalog [46]. Researchers can access the data from Swift containers we will extend for this project. We will release the source code of the experiments as a use case for other iVoyager researchers.

Dataset 1. Captured IBR from IPv6 darknet telescopes.

We will collect IBR traffic sent to unused address spaces of iVoyagerStar and the deployed iVoyagerSatellite to create a dataset comparable to that generated by the IPv4-based UCSD-NT. Our preliminary investigation and prior literature [30] based on short-term datasets show that IPv6 darknets currently attract little traffic [6, 47]. Collecting longitudinal data will be valuable for observing long-term trends in the IPv6 IBR ecosystem (the result from Internet-wide scanning, malware, and misconfigurations) and serve as a baseline against which to compare innovations in building IPv6 measurement instrumentation such as proactive telescopes.

Dataset 2. Captured IBR from Proactive IPv6 telescope.

Our team is currently collaborating with the student author of [30] (USENIX Security 2023) at the University of Iowa, who concluded his summer internship at CAIDA in 2023. We are extending his work to improve IPv6 IBR visibility by introducing multiple triggering actions to attract IPv6 scanners. These actions require announcing more specific prefixes via BGP, responding from the entire targeted subnet [23], modifying DNS zone files, signing TLS certificates, and deploying honeypots at different levels of interactivity. We built a prototype proactive IPv6 telescope within a transit ISP and found that our triggers do indeed induce both impulse and long-term effects on the amount of unsolicited traffic. In particular, deployment of a high-interaction honeypot led to a more than 100x increase in traffic compared to the baseline (Figure 3).

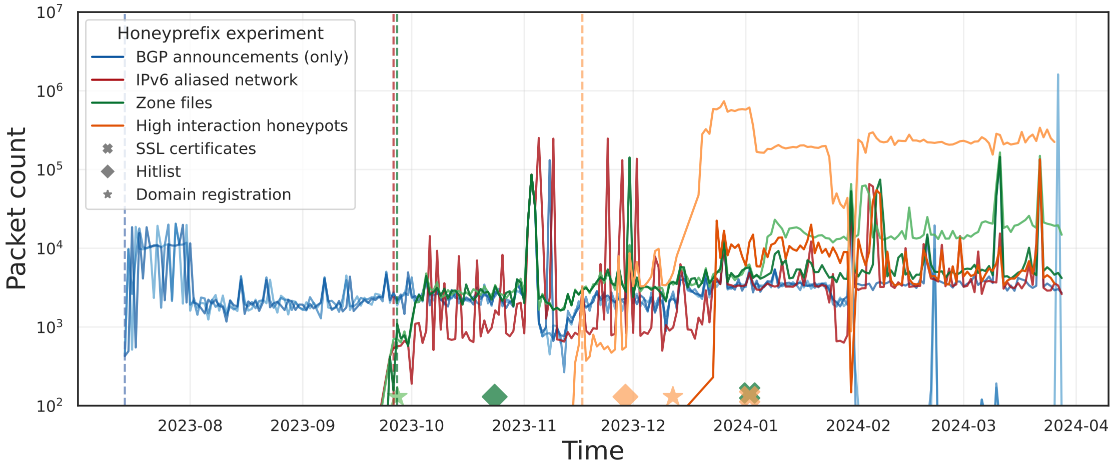

Figure 3: The traffic volume (in packet counts) significantly surged after we introduced each trigger, illustrating the importance of deploying triggers to solicit IPv6 traffic to network telescopes. iVoyagerStar will support researchers to configure these triggers for their experiments.

Our IPv6-only current prototype relied on our commercial ISP partner's network, which prohibited us from sharing the datasets and infrastructure. The research process helped us identify programming and operational limitations to deploying tools to monitor vast blocks of IPv6 address space that our team will address in iVoyagerStar. Since this experiment will use most of the iVoyagerStar components, it will serve as an example for researchers on how to use the infrastructure. The datasets collected will provide insights into how the responsiveness of different network components attracts scanning and other malicious behavior.

Dataset 3. Tracking malware propagation.

Our team at LSU will deploy their honeypot software, HoneyComb [22], specifically designed to monitor the spread of Mirai malware [48], which was first discovered in 2016. Mirai exploits Internet of Things (IoT) devices, such as IP cameras and home routers, creating botnets for launching DDoS attacks. HoneyComb uses packet fingerprints captured by the network telescope to identify and respond to Mirai scanning attempts. HoneyComb engages with infected hosts to obtain malware binaries and visualizes trends in malware spread. This experiment will record the list of victim IPs and create a longitudinal sample of binaries, aiding researchers in studying the evolution of the malware. Figure 4 shows the dashboard of the HoneyComb platform that visualizes the spread of Mirai across the globe.

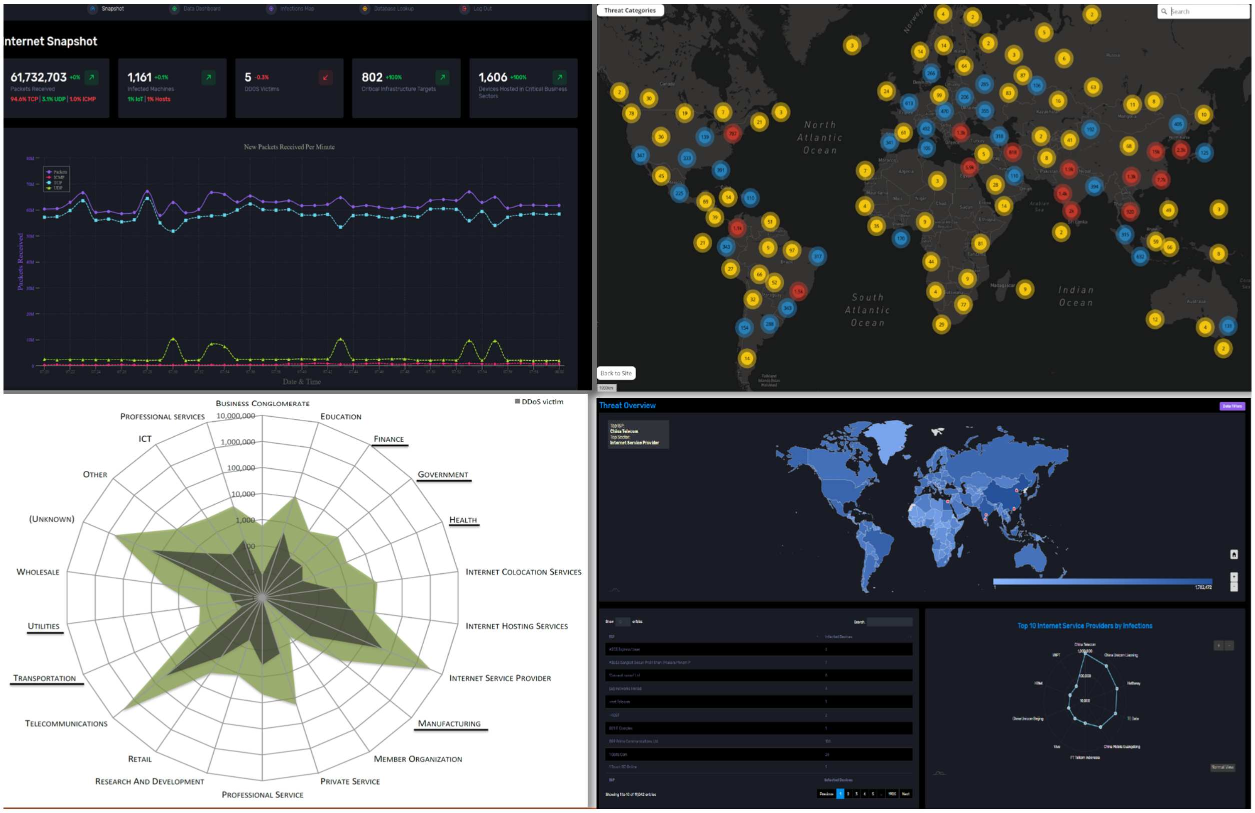

Figure 4: Dashboard of HoneyComb platform, visualizing the trend of the number of inflected hosts (top-left), their geographical distribution (top-right), their network types (bottom-left), and their ISPs (bottom-right).

3.5 User services

Our proposed iVoyager infrastructure will support a variety of services, as described in §3.1 and shown in Figure 1. On iVoyagerStar, we will provide unused IP subnets and virtual machines for researchers to run experiments that capture and respond to unsolicited traffic. iVoyagerStar will also offer common application services (e.g., DNS and NTP) and BGP servers, allowing researchers to apply various network configurations. iVoyagerStar's traffic monitor will capture packet with high precision timestamps, enabling researchers to perform in-depth traffic analysis.

Researchers will have the opportunity to conduct topology and other active measurements on iVoyagerSatellite VPs using CAIDA's domain-specific language [45]. Specifically, we will provide access to our VP controller, hosted at CAIDA, for executing measurement scripts written in this language.

3.6 Community engagement

Our community engagement efforts will involve the CISE research community in the use and management of iVoyager, including the following components.

Involve CISE researchers in infrastructure development We will establish a Community Advisory Board with experts in cybersecurity, ML/AI, Internet measurements, networking, etc. The board will meet annually and will guide iVoyager's development, evaluate its functionalities and ensure that it aligns with community needs. We will provide mechanisms, including our channelized chat server, to collect user feedback and use it to improve infrastructure. We will allow users to contribute new tools, software and data to the platform.

Build an interdisciplinary community for CS education We will provide educational resources (e.g. online tutorials, webinars, office hours) to help users learn about iVoyager's capabilities. We will host annual hackathons targeting a wide-range of audience aimed at the use of iVoyager to solve real-world problems. We will invite researchers and experts from different areas (e.g. ML/AI, cybersecurity, operators) to participate in interdisciplinary collaborative projects using the iVoyager infrastructure.

Supporting Community-driven Research We will organize and facilitate regular meetings focused on pressing operational challenges (e.g. IPv6 security, ML/AI for threat detection, and anomaly detection) that iVoyager can support. We will encourage collaboration between different academic and industry groups by helping them to secure funding and lead multi-institution research projects using iVoyager. We will promote and facilitate sharing of research results, data and tools by indexing publications, software and dataset in CAIDA's catalog and disseminating information in blogs and newsletters.

3.7 Community outreach

Outreach to the CISE research community is a critical component of iVoyager's success. We will conduct outreach to expand awareness and encourage the broader research community to adopt and use iVoyager. Below we summarize our planned outreach activities.

Disseminate information about iVoyager We will engage in our long-standing dissemination activities: presenting infrastructure status and research results at research meetings and workshops, and organizing workshops and tutorials where researchers can share their experiences, present their work, and discuss improvements for the infrastructure. We will develop a project website for documentation, and updates on capabilities and research findings. We will maintain discussion and support forums using our MatterMost platforms, and publish annual newsletters detailing the datasets collected, research collaborations formed, and known publications and findings generated through iVoyager. We will provide a support team to assist new users and to help with troubleshooting.

Integrate feedback into infrastructure evolution We will use the channels described above to solicit feedback on user satisfaction with the infrastructure and data products, which we will use to refine and improve subsequent infrastructure operations. We will announce these meetings and Mattermost access to other CISE community channels including NSF and ACM Slack teams. We will attend academic conferences and workshops where many researchers are already using CAIDA's data to socialize new data products as they are available. We will announce new infrastructure developments, tools, and data sets on CAIDA's data-announce@caida.org mailing list that has over 3,000 subscribed users.

4 Focused research agenda: scalably gathering threat intelligence

Consistent with the goals of the CIRC, we propose a bold research direction, which this infrastructure will facilitate and catalyze: *effectively exploring the landscape of Internet threats by scalably* *gathering cyber threat intelligence.* iVoyager will facilitate the rapid development and scalable deployment of distributed, dual-stack (IPv4 and IPv6), reactive telescopes. The resulting infrastructure will collect longitudinal datasets that facilitate use of machine learning/artificial intelligence (ML/AI) for cyber threat hunting, anomaly detection, and malware analysis.

As the adoption of IPv6 continues to increase, iVoyager will provide critical and unique datasets to open opportunities to answer a broad range of IPv6 research questions in four CISE communities: ML/AI for anomaly detection, target generation algorithms, blocklist generation, and DDoS characterization. Users of our existing data sets, including the UCSD telescope, have made clear they would pursue additional research opportunities given the new capabilities we are proposing.

4.1 ML/AI for network anomaly detection.

The longitudinal datasets that iVoyager will collect (§3.4.3) will provide training data for ML/AI applications for detecting network anomalies generated by malicious actors, such as scanners, malware, and denial-of-service (DoS) attacks. Recent work has adopted different ML/AI algorithms to identify clusters of senders with similar characteristics (e.g., DarkVec [49], DANTE [50], DarknetSec [51], and Kallitsis et al. [52]) and detect changes in traffic time series (e.g., Bou-Harb et al. [53], Dark-TRACER [54], and DarkSim [55]) using IPv4 network telescopes, including UCSD-NT.

Given the vast size difference in IPv6 address space and unique characteristics of IPv6 scanner behavior [30, 31, 56], iVoyager's new datasets will enable researchers to develop new ML/AI algorithms to answer research questions including: How do scanners discover targets in IPv6 networks? What are the traffic characteristics and scanning patterns of different scanning tools? What are the attack vectors commonly exploited by attackers in IPv6 networks?

Researchers can conduct joint analyses using iVoyager's datasets with network flow data and packet traces collected by other CISE community infrastructure, including the NETAI4ALL proposal and CC* project [57] led by Dr. Ram Durairajan at University of Oregon (see LoC) and ILANDS project led by co-PI Claffy [58]. Combining the IBR data with user traffic data will shed light on the prevalence of cyberattacks in IPv6 networks and the effectiveness of trained ML/AI models for detecting unsolicited activities in production networks or the Internet core.

4.2 Evaluation for IPv6 target generation algorithms

Current technology (e.g., Zmap [59–61] and Yarrp [62, 63]) can efficiently scan the entire IPv4 address space in less than a hour, but such an approach is infeasible in IPv6 due to its vast address space. An emerging area of IPv6 research involves developing Target Generation Algorithms (TGAs) (e.g., Gasser et al. [23] 6GAN [24], 6Hit [25], 6Tree [26], AddrMiner [27], 6Sense [28], and Rye and Levin [29])), to compile lists of responsive IPv6 addresses, also known as IPv6 hitlists.

These TGAs leverage publicly accessible datasets (e.g., historical topology data [64], CTLog [65], DNS names [66], and TLD zone files [67]) and/or set up services (e.g., NTP clients [29]) to record end-host IPs as seed addresses to generate lists of candidate IPs to probe, curating a list of responsive hosts as hitlists. Without ground truth, the primary evaluation metric for these efforts is the size of the lists (i.e., true positives).

iVoyager will serve two functions in enhancing the evaluation of TGAs. First, as a network telescope and honeypot, iVoyager will be able to capture the probe packets generated by TGAs. Analyzing this traffic can validate the TGA's actual probing characteristics against the intended behavior. Second, iVoyager will provide a testing ground by emulating different network characteristics, such as the distribution of responsive hosts, routers, and services. This environment will allow us to evaluate TGA algorithms by compiling hitlists of iVoyager's monitored network address blocks, and investigating why certain algorithms fail to discover live hosts.

4.3 Blocklist generation

IP blocklists have been an effective threat intelligence tool for network operators to block malicious actors and malware-infected hosts from intruding into IPv4 networks [68, 69]. Blocklist providers often crowdsoure information from operators (e.g., AbuseIPDB [70]) or use network telescopes and honeypots to capture IPs involved in malicious activities.

In IPv6, the recommended practice is to assign a subnet block (e.g., /64, as recommended by RFC6177 [71], or /56 for AWS EC2 VMs) rather than an individual address (i.e., /128) to each user. Attackers can dynamically change source addresses within the block, rendering blocklisting individual addresses ineffective. The IBR data collected by iVoyager will enable researchers to develop new methods to characterize the distribution of source addresses from each network and infer subnet sizes.

4.4 IPv6-based DDoS characterization

UCSD-NT provides a large aperture to effectively capture backscatter generated by DDoS victims responding to requests sent using randomly spoofed source IP addresses [7]. Honeypots (e.g., AmpPot [18], Hopscotch [72]) attract attackers exploiting the honeypot's services for reflection attacks. Our recent study [73] offered a comprehensive review and comparison of long-term trends in IPv4 DDoS activities and visibility into attack events from different vantage points.

5 Relationship with existing resources

iVoyager will provide new and unique resources to the CISE community while complementing existing NSF-funded resources to enable new research opportunities. We have identified four major CISE infrastructures and resources relevant to this project:

1. **UCSD-NT** [1] is the world's largest IPv4 network telescope hosted by CAIDA, capturing over 1 TB of IPv4 IBR daily. Unlike UCSD-NT, iVoyager will capture IPv6 IBR and will also be capable of responding to IBR for further analysis. Researchers will be able to conduct joint analyses using datasets from both infrastructures.
2. **ILANDS** [58] builds 100Gbps traffic monitors to capture traffic from the Internet backbone. Datasets gathered by iVoyager can help train models to identify malicious activities in user traffic data.
3. **PEERING** [74] allows users to make BGP announcements to manipulate the routing control plane, similar to iVoyagerSatellite. PEERING aims to support various short-term BGP-related network experiments, whereas iVoyager will include longitudinal data collection to capture long-term trends suitable for ML/AI model training.
4. **FABRIC** [75]**, CloudLab** [76]**, and Chameleon** [77] are configurable testbed platforms designed for controlled networking experiments. iVoyager, on the other hand, aims to study the global Internet and collect real-world active and traffic measurement.
5. **Expanse** [40] is an NSF-funded high-performance cluster at SDSC, connected to CAIDA's storage cluster via a high-speed network. Researchers can efficiently transfer data from iVoyager to Expanse for complex ML/AI computations.

Expanding the user community of our existing telescope data set will immediately benefit from the new data sets. Between January 2021 and September 2024, we received ≈90 requests for access to restricted but downloadable UCSD-NT datasets, ≈100 requests for access to UCSD-NT data that requires analysis via a virtual machine (VM) at CAIDA. During this period, around 400 users downloaded publicly available UCSD-NT datasets. Use cases for the data included: (a) machine learning applications for DDoS attack detection, anomaly/intrusion detection, and other Internet security threats; (b) network traffic measurement and behavior analysis; and (c) IoT and Edge computing research.

References

[1] CAIDA, "STARDUST. UCSD network telescope," 2021.

[2] Merit, "ORION network telescope." https://www.merit.edu/initiatives/orion-network-telescope/, 2023.

[3] E. Pauley, P. Barford, and P. McDaniel, "DSCOPE: A cloud-native Internet telescope," in *Proceedings of USENIX Security Symposium*, 2023.

[4] L. Izhikevich, M. Tran, M. Kallitsis, A. Fass, and Z. Durumeric, "Cloud watching: Understanding attacks against cloud-hosted services," in *Proceedings of ACM Internet Measurement Conference*, 2023.

[5] R. Pang, V. Yegneswaran, P. Barford, V. Paxson, and L. Peterson, "Characteristics of internet background radiation," in *Proceedings of ACM Internet measurement conference*, 2004.

[6] J. Czyz, K. Lady, S. G. Miller, M. Bailey, M. Kallitsis, and M. Karir, "Understanding ipv6 internet background radiation," in *Proceedings of ACM Internet measurement conference*, pp. 105–118, 2013.

[7] 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, p. 115–139, may 2006.

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

[9] 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*, 2017.

[10] M. Jonker, A. Pras, A. Dainotti, and A. Sperotto, "A first joint look at DOS attacks and BGP blackholing in the wild," in *Proceedings of ACM Internet Measurement Conference*, pp. 457–463, 2018.

[11] 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 ACM Internet Measurement Conference*, 2022.

[12] 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.

[13] S. Torabi, E. Bou-Harb, C. Assi, M. Galluscio, A. Boukhtouta, and M. Debbabi, "Inferring, characterizing, and investigating internet-scale malicious IoT device activities: A network telescope perspective," in *Proceedings of IEEE/IFIP International Conference on Dependable Systems and Networks*, jun 2018.

[14] C. Fachkha, E. Bou-Harb, A. Keliris, N. Memon, and M. Ahamad, "Internet-scale probing of CPS: Inference, characterization and orchestration analysis," in *Proceedings of Network and Distributed System Security Symposium*, Internet Society, 2017.

[15] F. Shaikh, E. Bou-Harb, J. Crichigno, and N. Ghani, "A machine learning model for classifying unsolicited IoT devices by observing network telescopes," in *Proceedings of International Wireless Communications & Mobile Computing Conference*, IEEE, jun 2018.

[16] S. Torabi, E. Bou-Harb, C. Assi, E. B. Karbab, A. Boukhtouta, and M. Debbabi, "Inferring and investigating IoT-generated scanning campaigns targeting a large network telescope," *IEEE Transactions on Dependable and Secure Computing*, vol. 19, pp. 402–418, Jan. 2022.

[17] F. Zhang, S. Zhou, Z. Qin, and J. Liu, "Honeypot: a supplemented active defense system for network security," in *Proceedings of the Fourth International Conference on Parallel and Distributed Computing, Applications and Technologies*, pp. 231–235, 2003.

[18] L. Kramer, J. Krupp, D. Makita, T. Nishizoe, T. Koide, K. Yoshioka, and C. Rossow, "AmpPot: Monitoring and Defending Amplification DDoS Attacks," in *Proceedings of the 18th International Symposium on Research in Attacks, Intrusions and Defenses*, 2015.

[19] W. Han, Z. Zhao, A. Doupe, and G.-J. Ahn, "HoneyMix: Toward SDN-based intelligent honeynet," in *Proceedings of ACM SDN-NFV Security*, Mar. 2016.

[20] D. Telekom Security GmbH, "T-Pot - The all in one multi honeypot platform." https://github.com/telekom-security/tpotce.

[21] 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.

[22] 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.

[23] O. Gasser, Q. Scheitle, P. Foremski, Q. Lone, M. Korczyński, S. D. Strowes, L. Hendriks, and G. Carle, "Clusters in the Expanse: Understanding and Unbiasing IPv6 Hitlists," in *Proceedings of the Internet Measurement Conference 2018*, 2018.

[24] T. Cui, G. Gou, G. Xiong, C. Liu, P. Fu, and Z. Li, "6GAN: IPv6 multi-pattern target generation via generative adversarial nets with reinforcement learning," in *Proceedings of IEEE INFOCOM*, 2021.

[25] B. Hou, Z. Cai, K. Wu, J. Su, and Y. Xiong, "6Hit: A reinforcement learning-based approach to target generation for Internet-wide IPv6 scanning," in *Proceedings of IEEE INFOCOM*, 2021.

[26] Z. Liu, Y. Xiong, X. Liu, W. Xie, and P. Zhu, "6Tree: Efficient dynamic discovery of active addresses in the IPv6 address space," *Computer Networks*, vol. 155, pp. 31–46, May 2019.

[27] G. Song, J. Yang, L. He, Z. Wang, G. Li, C. Duan, Y. Liu, and Z. Sun, "AddrMiner: A comprehensive global active IPv6 address discovery system," in *Proceedings of USENIX Annual Technical Conference*, 2022.

[28] G. Williams, M. Erdemir, A. Hsu, S. Bhat, A. Bhaskar, F. Li, and P. Pearce, "6Sense: Internet-wide IPv6 scanning and its security applications," in *Proceedings of USENIX Security Symposium*, 2024.

[29] E. Rye and D. Levin, "IPv6 hitlists at scale: Be careful what you wish for," in *Proceedings of ACM SIGCOMM*, Sept. 2023.

[30] H. B. Tanveer, R. Singh, P. Pearce, and R. Nithyanand, "Glowing in the dark: Uncovering IPv6 address discovery and scanning strategies in the wild," in *Proceedings of USENIX Security Symposium*, pp. 6221–6237, 2023.

[31] L. Zhao, S. Kobayashi, and K. Fukuda, "Exploring the discovery process of fresh IPv6 prefixes: An analysis of scanning behavior in darknet and honeynet," in *Proceedings of Passive and Active Measurement Conference*, 2024.

[32] P. Richter and A. Berger, "Scanning the scanners: Sensing the internet from a massively distributed network telescope," in *Proceedings of ACM Internet measurement conference*, Oct 2019.

[33] P. Richter, O. Gasser, and A. Berger, "Illuminating large-scale IPv6 scanning in the internet," in *Proceedings of ACM Internet Measurement Conference*, 2022.

[34] E. Rye, R. Beverly, and K. C. Claffy, "Follow the scent: defeating IPv6 prefix rotation privacy," in *Proceedings of ACM Internet Measurement Conference*, Nov. 2021.

[35] S. J. Saidi, O. Gasser, and G. Smaragdakis, "One bad apple can spoil your IPv6 privacy," *ACM SIGCOMM Computer Communication Review*, vol. 52, pp. 10–19, Apr. 2022.

[36] NSA, "IPv6 Security Guidance." https://media.defense.gov/2023/Jan/18/2003145994/-1/-1/0/CSI_IPV6_SECURITY_GUIDANCE.PDF, Jan. 2023.

[37] cowrie, "Cowrie honeypot." https://github.com/cowrie/cowrie, 2023.

[38] "Endlessh: an SSH tarpit." https://github.com/skeeto/endlessh, Apr. 2019.

[39] N. Ilg, P. Duplys, D. Sisejkovic, and M. Menth, "A survey of contemporary open-source honeypots, frameworks, and tools," *Journal of Network and Computer Applications*, vol. 220, p. 103737, Nov. 2023.

[40] SDSC, "Expanse," 2020.

[41] ACCESS, "Advanced cyberinfrastructure coordination ecosystem: Services & support." https://allocations.access-ci.org, 2023.

[42] CAIDA, "Scamper."

[43] VULTR, "VULTR: The Everywhere Cloud." http://www.vultr.com/. [Last accessed: September 11, 2024].

[44] R. Sommese, L. Bertholdo, G. Akiwate, M. Jonker, R. van Rijswijk-Deij, A. Dainotti, K. Claffy, and A. Sperotto, "Manycast2: Using anycast to measure anycast," in *Proceedings of ACM Internet Measurement Conference*, 2020.

[45] M. Luckie, "CAIDA Blog. Towards a Domain Specific Language for Internet Active Measurement." https://blog.caida.org/best_available_data/2024/01/16/towards-a-domain-specific-language-for-internet-active-measurement/, 2024.

[46] CAIDA, "CAIDA Resource Catalog." https://catalog.caida.org/.

[47] J. Ronan and D. Malone, "Revisiting and revamping an IPv6 network telescope," in *Proceedings of Irish Signals and Systems Conference (ISSC)*, 2023.

[48] 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 USENIX Conference on Security Symposium*, 2017.

[49] 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.

[50] 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.

[51] J. Lan, X. Liu, B. Li, Y. Li, and T. Geng, "DarknetSec: A novel self-attentive deep learning method for darknet traffic classification and application identification," *Computers & Security*, vol. 116, p. 102663, 2022.

[52] 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.

[53] E. Bou-Harb, M. Debbabi, and C. Assi, "A time series approach for inferring orchestrated probing campaigns by analyzing darknet traffic," in *Proceedings of International Conference on Availability, Reliability and Security*, aug 2015.

[54] 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.

[55] M. Gao, R. Mok, E. Carisimo, E. Li, S. Kulkarni, and kc claffy, "DarkSim: A similarity-based time-series analytic framework for darknet traffic," in *Proceedings of ACM Internet measurement conference*, 2024.

[56] K. Fukuda and J. Heidemann, "Who Knocks at the IPv6 Door?: Detecting IPv6 Scanning," in *Proceedings of ACM Internet Measurement Conference*, 2018.

[57] R. Durairajan, "CC* Integration-Large: Bringing Code to Data: A Collaborative Approach to Democratizing Internet Data Science." https://www.nsf.gov/awardsearch/showAward?AWD_ID=2126281, 2021.

[58] CAIDA, "Integrated library for advancing network data science - (ILANDS)." https://www.caida.org/funding/ccri-ilands/, 2021.

[59] Z. Durumeric, E. Wustrow, and J. A. Halderman, "ZMap: Fast internet-wide scanning and its security applications," in *Proceedings of USENIX Security Symposium*, pp. 605–620, 2013.

[60] Z. Durumeric, M. Bailey, and J. A. Halderman, "An Internet-wide view of Internet-wide scanning," in *Proceedings of USENIX Security Symposium*, 2014.

[61] "The ZMap Project." https://zmap.io/, 2024. Accessed 2024-05-15.

[62] R. Beverly, "Yarrp'ing the Internet: Randomized high-speed active topology discovery," in *Proceedings of ACM Internet Measurement Conference*, Nov. 2016.

[63] "Yarrp (Yelling at Random Routers Progressively)," 2019. https://www.cmand.org/yarrp/.

[64] CAIDA, "Ark IPv6 Topology Dataset." https://catalog.caida.org/dataset/ipv6_allpref_topology.

[65] "Certificate transparency." https://certificate.transparency.dev.

[66] CAIDA, "Ark IPv6 Topology DNS Names." https://catalog.caida.org/dataset/ark_ipv6_dns_names.

[67] ICANN, "About zone file access." https://www.icann.org/resources/pages/zfa-2013-06-28-en, 2013.

[68] V. G. Li, M. Dunn, P. Pearce, D. McCoy, G. M. Voelker, and S. Savage, "Reading the tea leaves: A comparative analysis of threat intelligence," in *Proceedings of USENIX security symposium*, 2019.

[69] V. G. Li, G. Akiwate, K. Levchenko, G. M. Voelker, and S. Savage, "Clairvoyance: Inferring blocklist use on the internet," in *Proceedings of Passive and Active Measurement*, 2021.

[70] "AbuseIPDB." https://www.abuseipdb.com. Accessed: 2024-05-16.

[71] T. Narten, G. Huston, and L. Roberts, "RFC 6177: IPv6 address assignment to end sites." https://www.rfc-editor.org/rfc/rfc6177.html, Mar. 2011.

[72] D. R. Thomas, R. Clayton, and A. R. Beresford, "1000 days of UDP amplification DDoS attacks," in *Proceedings of APWG Symposium on Electronic Crime Research (eCrime)*, Apr. 2017.

[73] R. Hiesgen, M. Nawrocki, M. Barcellos, D. Kopp, O. Hohlfeld, E. Chan, R. Dobbins, C. Doerr, C. Rossow, D. R. Thomas, M. Jonker, R. K. P. Mok, X. Luo, J. Kristoff, T. Schmidt, M. Wählisch, and K. Claffy, "The age of DDoScovery: An empirical comparison of industry and academic DDoS assessments," in *Proceedings of ACM Internet Measurement Conference*, 2024.

[74] "PEERING: The BGP testbed." https://peering.ee.columbia.edu/, 2024.

[75] "FABRIC testbed." https://fabric-testbed.org, 2024.

[76] "CloudLab." https://www.cloudlab.us.

[77] "Chameleon." https://www.chameleoncloud.org, 2024.