Community Datasets

PAM 2026 Panel

Community Labeling and Sharing of Security and Networking Test datasets (CLASSNET)

Jelena Mirkovic, John Heidemann, Wes Hardaker (USC/ISI)
Bob Stovall (Merit Network, Inc.)

Work funded by the NSF award #CRI-8115780

merit

User Portal: https://comunda.isi.edu

Project Web page: https://ant.isi.edu/classnet/

Project Goals

- Provide labeled datasets for cybersecurity
  - Both curated (snapshots) and ongoing
  - Current, relevant, labeled
- Build a community around datasets
  - Enable users to rate and review datasets
  - Enable users to contribute alternative/additional labels
- Act as middleman between data providers and consumers
  - Akin to DHS IMPACT but lower friction
  - Provide mechanisms for easy onboarding of new data providers

Darknet data from Merit
Census, hitlists, outages, B-root DNS from USC/ISI
Netflow from FRGP and FIU and DDoS attack alerts
Vehicle CAN bus data from PIVOT project

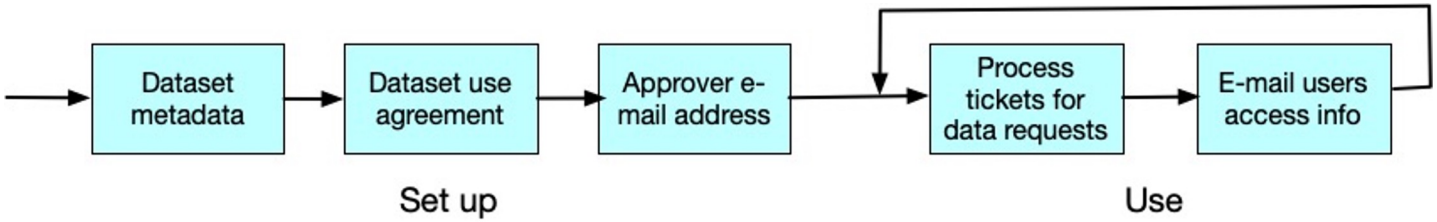
USC INFORMATION SCIENCES INSTITUTE

merit

User Portal: https://comunda.isi.edu

Project Web page: https://ant.isi.edu/classnet/

Provider Workflow



- No loss of privacy, data and decision-making stays with provider
- Can offload ticket processing to us, if desired
- Standardized metadata, ready to use platform
- Reach a wider audience
- COMUNDA provider page will produce statistics of dataset requests ready for reporting (how many requests, by whom, for what purpose)

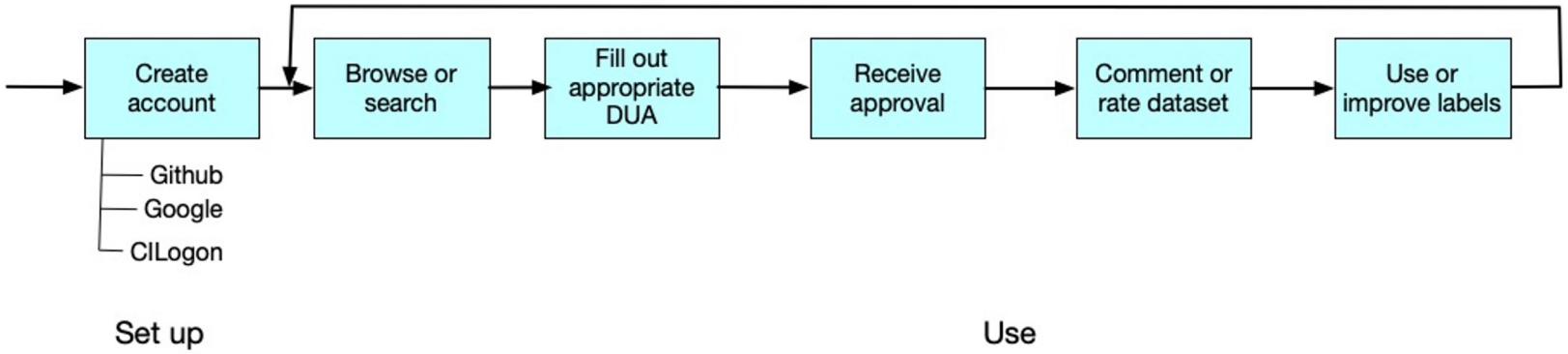
USC INFORMATION SCIENCES INSTITUTE

merit

User Portal: https://comunda.isi.edu

Project Web page: https://ant.isi.edu/classnet/

User Workflow



- A variety of datasets, some labeled and ready for ML
- Easy, click-through request process, especially for multiple datasets
- Can rate and comment on datasets, can submit label corrections or alternative labels

USC INFORMATION SCIENCES INSTITUTE

merit

User Portal: https://comunda.isi.edu

Project Web page: https://ant.isi.edu/classnet/

Current Status

- Live Web portal for users and for providers (about a year old)
  - Log in with your institutional credentials, Github or Google
  - Search, request, rate and comment on datasets, upload or download labels
  - Around 100 users
- Smooth data ingest and release
  - Mostly automated, a few manual steps only around setting up a new provider or DUA
  - Provider statistics page
- Continued outreach to providers and users
  - Building a user base takes time

LaSIC: Labeled Security Information Capture

Project provides labeled cybersecurity datasets to the community collected at academic ISPs

- AmLight (Miami, FL)
- FRGP (Denver, CO)
- Supports continuous flow and on-demand packet data
- Nearly complete network coverage
- Access to ISP personnel

Datasets and Tools:

- Netflow (anonymized and raw)
- Alerts from two commercial IDS deployments
  - Tool to correct alert timestamps
- Intel Telemetry data
- Unsampled flow data via custom tool
- Short bursts of packet data, potentially including payload
- On-site access to raw data

Work funded by NSF award ##2232864

The AmLight Network

AmLight is an International Research & Education Network built to enable collaboration among Latin America, Africa, and the U.S.

- Members: FIU, AURA, Vera Rubin Observatory, RNP, Rednesp, RedClara, REUNA, FLR, SANReN, TENET, and Internet2

Supported by several NSF OAC awards since 2010.

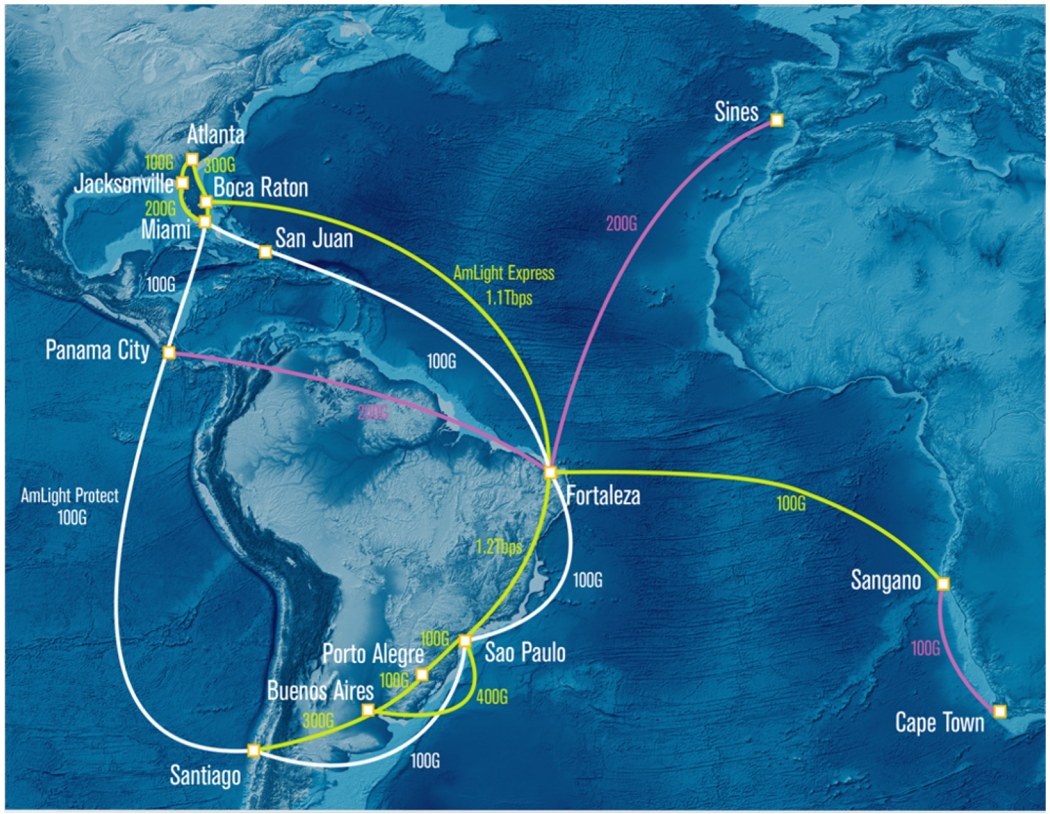
5.1+ Tbps of total connectivity

Landings in four continents

NAPs: Florida(3), Atlanta, Brazil(2), Chile, Puerto Rico, Argentina, Panama, and South Africa

Infrastructure managed by a our SDN controller: Kytos-ng (github.com/kytos-ng)

https://www.amlight.net/

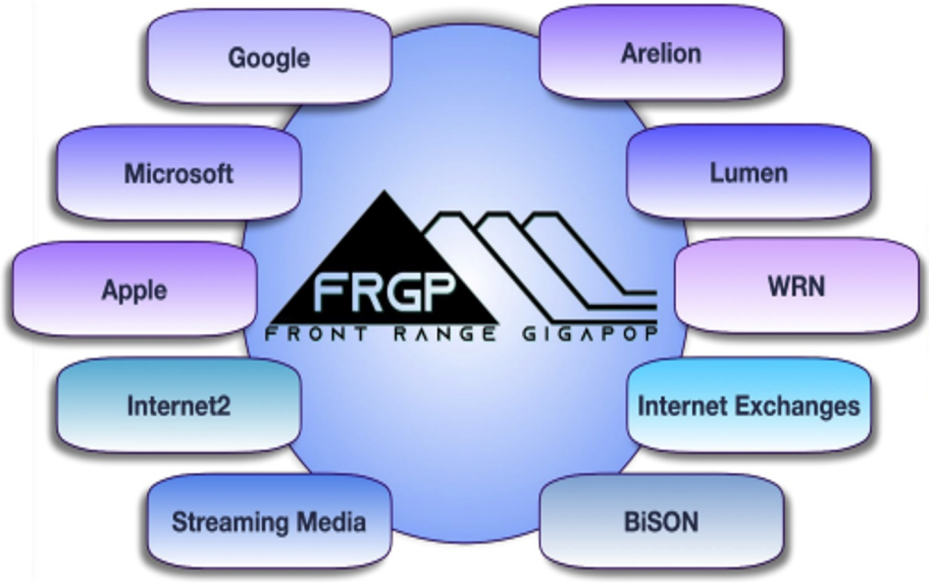


How AmLight Supports LaSIC

- AmLight provides LaSIC with data from three sources
  - Raw IP packets at line rate
  - NetFlow data from its routers
  - IDS alerts from its commercial ISPs when there is a DDoS attack traversing the commercial ISP's network
- FIU archives telemetry data related to DDoS attacks on a storage appliance with 96TB of storage capacity
  - Packets unrelated to the DDoS attack are discarded
  - Datasets and metadata from the DDoS tools are archived for use

FRGP - Front Range GigaPop

The FRGP is a consortium of more than thirty higher education institutions, research organizations, non-profit corporations, government agencies, cities, counties, states, and K-12s that cooperate to share wide-area networking in the Rocky Mountain region. The FRGP is owned and controlled by the research and education communities to advance and enable science and academia.



UCAR
UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

From the high plains to the high peaks and beyond
25Th ANNIVERSARY
FRGP
FRONT RANGE GIGAPOP

FRGP Netflow Information

- Netflow covers the entire FRGP network:
  - Participant Interfaces, Provider and Peer Interfaces, and Core Backbone Router Connections
- There are 4 Core FRGP Routers:
  - Two routers have a sample rate of 1 in 100 Packets
  - Two routers have a sample rate of 1 in 4096 packets
  - This is due to limitations of specific hardware
- FRGP Participants/Customers can view Netflow data through an internal application before the data is anonymized on a VM

UCAR

UNIVERSITY CORPORATION FOR ATMOSPHERIC RESEARCH

IDS Alerts & Alert Timestamp Correction

Source	Alerts	Period
FIU	~550	Jan 2024 – Mar 2026
FRGP	~3,600	Dec 2019 – Feb 2026

IDS Alert
- Target IP address
- Attack type & intensity
- Start & End time
- Duration

- IDS alert timestamps are often misaligned with actual attack traffic in NetFlow
- **Timestamp Correction Tool** - Scans the NetFlow traffic around each alert to find the true attack window
  - **Corrected Start time:** Searches backward from IDS reported start → **ground truth traffic ≥ alert intensity**
  - **Corrected End time:** Searches forward from corrected start time → traffic **drops below alert intensity** → Applies Hierarchical clustering to pinpoint true end

Result: Corrected timestamps → Accurate NetFlow labeling → Significantly better DDoS detection

NetFlow Integration with OpenSearch

Repository-backed overview of ingestion flow, OpenSearch pipelines, and the ECS-shaped data pushed into OpenSearch

What this deck covers

How flow files move from nfdump into OpenSearch bulk ingestion

What each pipeline adds: normalization, enrichment, scoring, and detections

Which ECS branches and raw netflow fields are stored in the index template

Implementation anchors

Ingestion script: netflow_to_opensearch.py

Pipelines: opensearch-pipelines/*.json

Schema contract: opensearch-templates/netflow-index-template.json

Operational highlights

Bulk upload with configurable parallelism and gzip-ready HTTP path

Timezone-aware timestamp normalization using collector metadata

Designed for OpenSearch 2.x data streams and lifecycle-managed storage

End-to-End Integration

How records are extracted, transformed, and indexed

Source and transport

The script calls nfdump directly and streams JSON objects without jq or curl helper stages.

Each bulk action targets index_name netflow-data-default and pipeline netflow-data-default-enhanced by default.

Collector metadata can add collector_timezone and collector_name to every document.

OpenSearch integration points

Index templates create a data stream-ready schema and attach the lifecycle policy.

Ingest pipelines centralize transformation so the Python path stays thin and fast.

GeoIP and threat logic run inside OpenSearch, keeping the upload side stateless.

Resulting operating model

All collectors can land in the same data stream while still using timezone-aware pipelines.

Dashboards and queries read ECS-shaped fields such as source.ip, destination.port, and flow.risk_level.

Raw netflow.* fields remain available for exporter-specific analysis and future schema growth.

Enhanced Pipeline Sequence

What happens inside the ingestion path

1. **Timestamp guard**
   Drops invalid 1970-01-01 received values from flow-tools exports.

2. **Timezone normalization**
   Uses collector_timezone when no offset is present and converts all timestamps to UTC.

3. **ECS mapping**
   Merges IPv4 and IPv6 fields, maps ports, bytes, packets, and protocol metadata.

4. **Date processors**
   Parses @timestamp, event.start, and event.end with ignore_failure safeguards.

5. **GeoIP**
   Adds source.geo and destination.geo when GeoIP databases are available.

6. **Scoring and tagging**
   Calculates risk bands, behavior patterns, investigation tags, and event categories.

7. **Cleanup**
   Removes raw helper fields and strips null or empty values before indexing.

Threat Detection Logic

Detections are additive and stored alongside the normalized flow

Flooding and volumetrics

small_packet_flood when packets are high but bytes per packet stay low

syn_flood when TCP SYN is present without ACK on short, packet-heavy flows

udp_flood when UDP traffic is packet-heavy within short windows

Scanning and recon

TCP SYN, FIN, and NULL scan heuristics for low-byte, low-packet probes

UDP probe detection for minimal-response scan behavior

Indicators accumulate into threat.indicators rather than replacing one another

Exfiltration and tunnels

large_data_exfiltration and persistent_data_exfiltration for outbound high-volume sessions

Suspicious outbound transfer when high-volume data leaves on non-standard ports

dns_tunneling, icmp_tunneling, and http_tunneling for covert-channel style behavior

Severity output

threat.score sums weights across all matched indicators

threat.severity becomes low, medium, high, or critical

event.type and tags are updated for alerting and dashboards

Example Indexed Document

Representative field layout after transformation and enrichment

ECS-style flow document


```
{
  "@timestamp": "2026-02-06T07:04:43.008Z",
  "event": {
    "dataset": "netflow",
    "kind": "event",
    "category": ["network"],
    "type": ["connection"],
    "ingested": "2026-02-06T07:04:45.100Z"
  },
  "flow": {
    "direction": "internal-external",
    "traffic_type": "web_https",
    "duration": 53248,
    "bytes_per_second": 43344.2,
    "risk_score": 15,
    "risk_level": "medium"
  },
  "source": {"ip": "192.168.1.100", "port": 54321},
  "destination": {"ip": "203.0.113.1", "port": 443},
  "network": {"protocol": "tcp", "bytes": 288500, "packets": 1400},
  "tcp": {"flags": {"syn": true, "ack": true}},
  "netflow": {"exporter_id": "collector-a"}
}
```


What changed from raw nfdump

src4_addr and dst4_addr collapse into source.ip and destination.ip.

proto becomes both a readable protocol name and network.iana_number.

Raw timestamps become event and flow dates that OpenSearch can sort and aggregate.

Query-ready benefits

Risk and threat fields are available at ingest time for dashboards and alerts.

GeoIP fields enable map visualizations and ASN-based pivots.

The same schema works across the original 35-field sample and the 31-field flow-tools sample.

PIVOT Dataset Platform

Secure cybersecurity dataset sharing — storage and access control layer

Problem

Comunda users need a secure, authenticated method to share cybersecurity datasets, including large, versioned collections hosted at their own institutions

Approach

Extend COMUNDA with a dedicated storage and access layer using the S3 protocol as a universal abstraction over institutional or cloud-hosted object stores

Additions

A Go microservice to provide/generate/validate endpoints for STS-based temporary credential issuance and fine-grained access control per dataset

Motivation

Why cybersecurity datasets need a dedicated storage and access layer

Large, versioned datasets

Cybersecurity datasets grow continuously and require version control. Researchers need to cite specific snapshots and reproduce experiments across versions.

Distributed ownership

Data lives at researchers' institutions, university clusters, or cloud providers, not in a single repository. Any solution must work across all of these.

Automated data producers

Network sensors, telescopes, and monitoring devices produce datasets continuously. They need to auto-upload to buckets without manual intervention.

S3 as a universal wire format

The S3 protocol is supported by every major self-hosted and cloud object store. By targeting S3, the PIVOT storage layer remains backend-agnostic and researchers can use whatever fits their institution.

Critically, modern S3-compatible stores support:

- Dataset versioning
- STS temporary credentials
- Fine-grained access policies
- Bucket-level permission scoping

S3-Compatible Storage Backends

Any backend supporting STS, versioning, and fine-grained access control is a valid target

Solution	Deployment	STS	Versioning	Fine-grained ACL
MinIO	Self-hosted	✓	✓	Bucket & prefix policies
Ceph RGW	Self-hosted	✓	✓	IAM-compatible policies
Garage	Self-hosted	✓	✓	Key-scoped bucket ACLs
SeaweedFS	Self-hosted	✓	✓	IAM policy engine
Cloudflare R2	Cloud (edge)	✓	✓	Token permission scopes
AWS S3	Cloud (AWS)	✓	✓	Full IAM + bucket policy
Backblaze B2	Cloud	✓	✓	Application key scoping

All options listed implement the AssumeRoleWithWebIdentity STS endpoint, object versioning, and bucket/prefix-level IAM-style policies — the three requirements for this integration.

COMUNDA Owns the Workflow, S3 owns the Data

Storage and access control were intentionally left out of scope — and that is the right design

COMUNDA's design is deliberately focused on community workflows, not storage infrastructure — a principled separation of concerns that makes extension straightforward.

Dataset discovery & search

Rich metadata catalog with search across all contributed datasets. Researchers can find relevant data by topic, format, time range, and contributor.

Data Use Agreement (DUA) management

COMUNDA handles the full DUA lifecycle: providers upload their agreement, researchers sign electronically, approvers receive signed copies — fully automated.

Access request workflow

Structured request and approval process with notifications. Providers define criteria (affiliation, faculty status, region); COMUNDA orchestrates the flow.

Intentionally out of scope

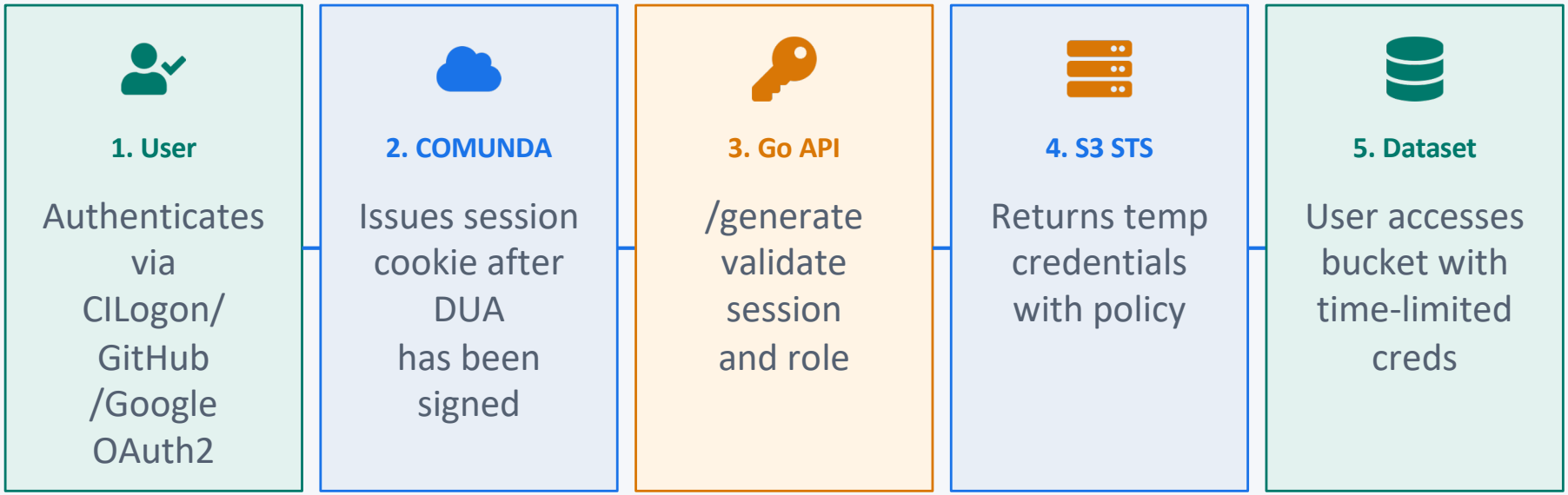
COMUNDA deliberately does not handle:

- Object storage backends
- Dataset file hosting
- Byte-level access control
- STS credential issuance
- Bucket permission policies and enforcement

These belong to infrastructure. COMUNDA's strength is the community, discoverability and governance layer.

Authentication Architecture

How COMUNDA sessions are extended to grant time-limited, role-scoped storage access



/validate

Called before every protected download. Confirms active COMUNDA session, signed DUA, and user role. Returns: allowed, role, message.

/generate

Exchanges a valid COMUNDA session cookie for short-lived S3 credentials scoped to the appropriate bucket policy (viewer, downloader, or uploader).

Access Roles & Permissions

Fine-grained control over who can view, download, or contribute datasets

Viewer

- Browse dataset metadata
- Read bucket object listings
- No download access

Granted after DUA signing for public-index datasets

Downloader

- All viewer permissions
- GetObject on approved buckets
- Scoped to specific prefix/version

Standard role after provider approves access request

Uploader / Producer

- All downloader permissions
- PutObject on assigned bucket
- Automated device or researcher upload

Issued to sensors, instruments, or trusted contributors

Credentials are always temporary

STS tokens are short-lived (configurable TTL). No long-term secrets are issued to users. Revocation happens passively as tokens expire.

Devices auto-upload to buckets

Network sensors and data-producing instruments receive uploader-scoped STS credentials and push datasets directly to designated buckets without any manual step.

Policies enforced at the storage layer

Bucket and prefix policies are evaluated by the S3 backend — independent of the API. Even if the Go API is bypassed, unauthorized access is rejected.

Merit ORION Telescope

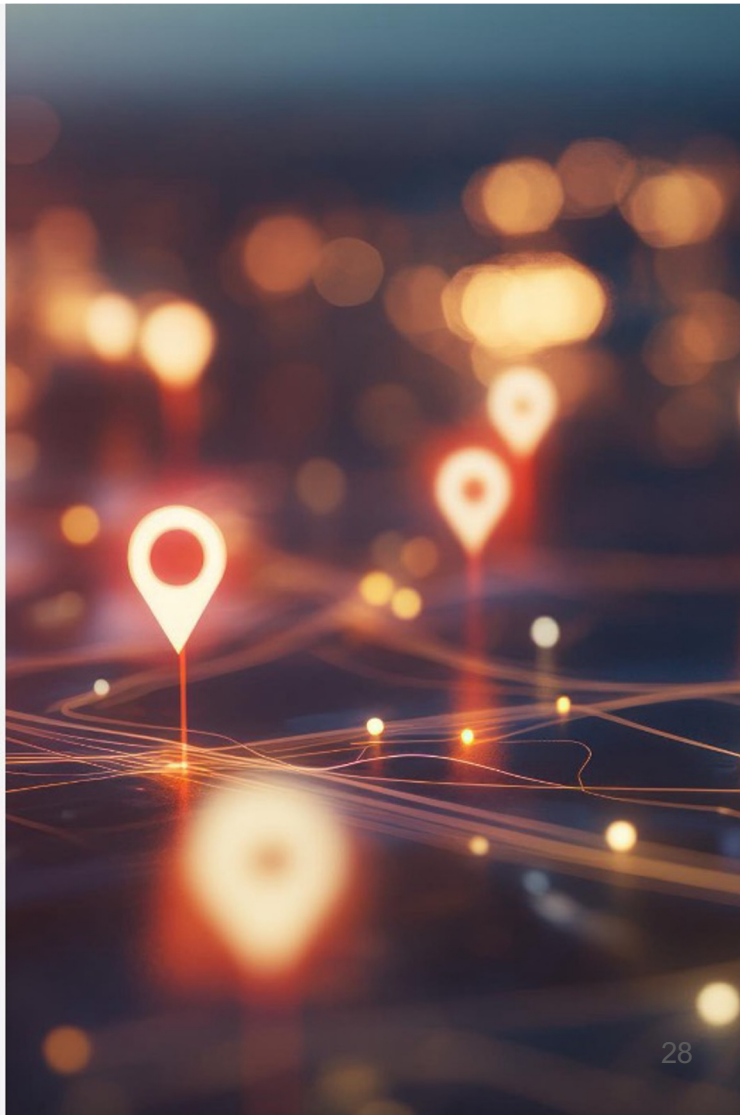
Passive and Active Measurement Conference 2026

March 24, 2026

Bob Stovall – Merit Network

About the telescope

- Collecting since 2005
- Monitors approximately 475,000 unused IPv4 addresses
- 474TB raw data + 76TB processed events
- Global coverage:Traffic from over 229,000 unique /24 networks worldwide
- Rich context: GeoIP, ASN enrichment, automated event detection, and DNS data
- Multiple formats: Raw PCAPs for deep analysis, structure JSON for ML/analytics

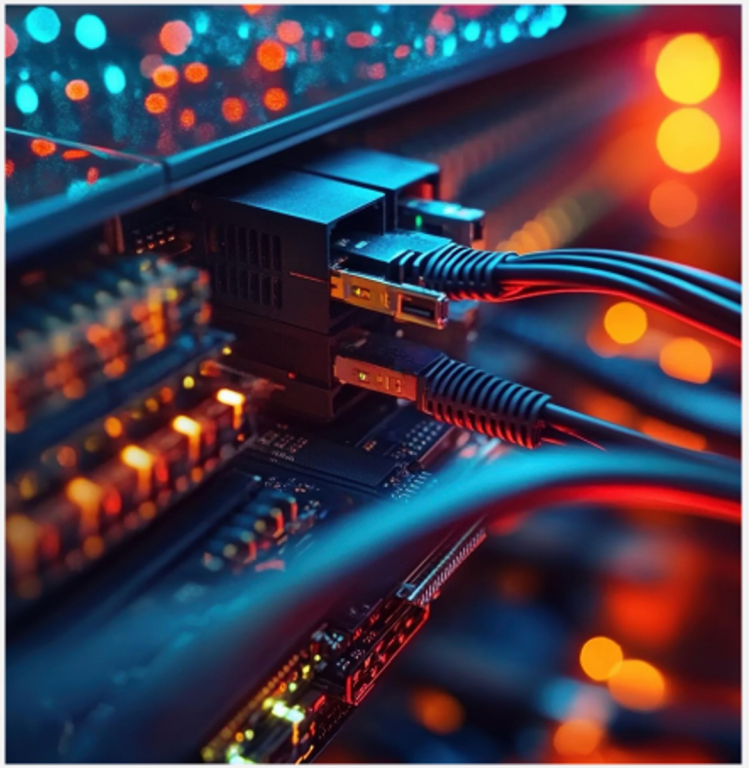


Telescope Operations

- Network telescopes monitor large blocks of unallocated IPv4/IPv6 address space (darknets)
- Any traffic destine to these unused addresses is inherently supicious – no legitimate services exist there
- Passive monitoring captures all incoming packets without responding
- Traffic is routed to collection infrastructure via BGP announcements of monitored prefixes

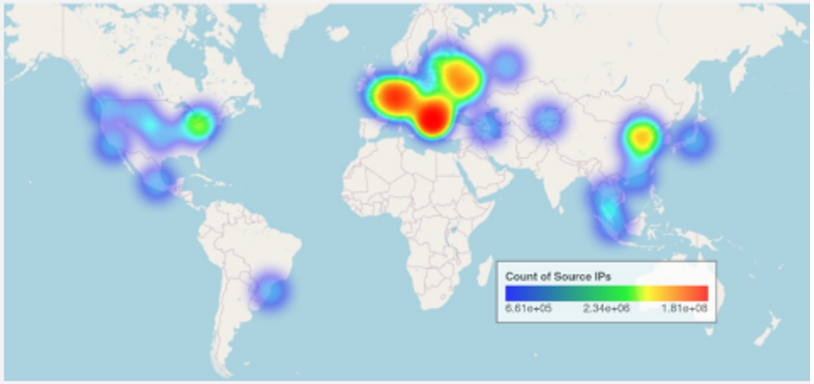
What Gets Captured:

- Backscatter from DDoS attacks – response packets from spoofed source IPs
- Scanning activity – automated bots probing for vulnerable systems
- Worm propagation attempts – self-replicating malware seeking new hosts
- Misconfigured systems – legitimate traffic sent to wrong IP addresses
- Internet Background Radiation (IBR) – continuous low-level malicious traffic

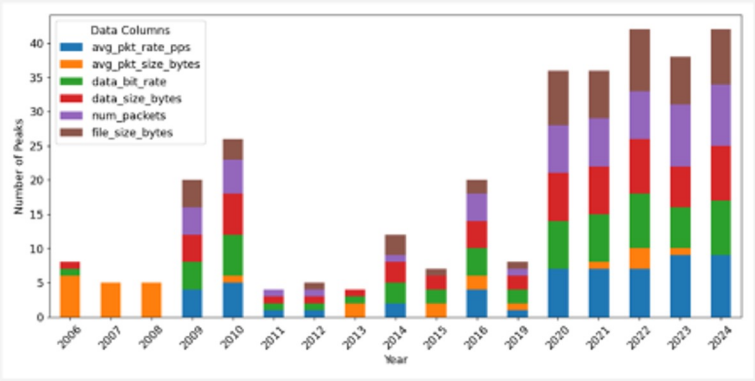


Current Research

- Machine Learning for Cyber Threat Detection
  - Apply supervised/semi-supervised/unsupervised ML models
  - Detect scanners, botnets, and malware patterns as well as emerging “never-seen before” threats
- Large-Scale Data Analysis
  - Analyzed high-volumes of unsolicited traffic across multiple darknet blocks
  - Build scalable pipelines for flow level, packet-level and metadata extraction
- Longitudinal Studies of Threat Evolution
  - Perform multi-year analysis of scanning campaigns, malware families, and actor infrastructure, internet-scale threats change over time
  - Modeling the evolution of Iot malware, CVE exploitation trends and C2 behavior
- Dynamic Network Telescope
  - Rotate prefixes automatically to avoid blacklisting and increase measurement diversity
  - Sustain meaningful monitoring even with IPv4 space becomes increasingly limited.



- **2024 Top 10 Source Ips**



- **Number of peaks per year 2006 to 2024**

THIRTY INSTITUIONAL, STATE, + INTERNATIONAL PARTNERS

(and counting)

- Data/IPv4 Provider
- Governance Board
- Data Consumer (for research or as a tool for teaching and learning)
- Data Set Distribution
- Data Transfer Partner