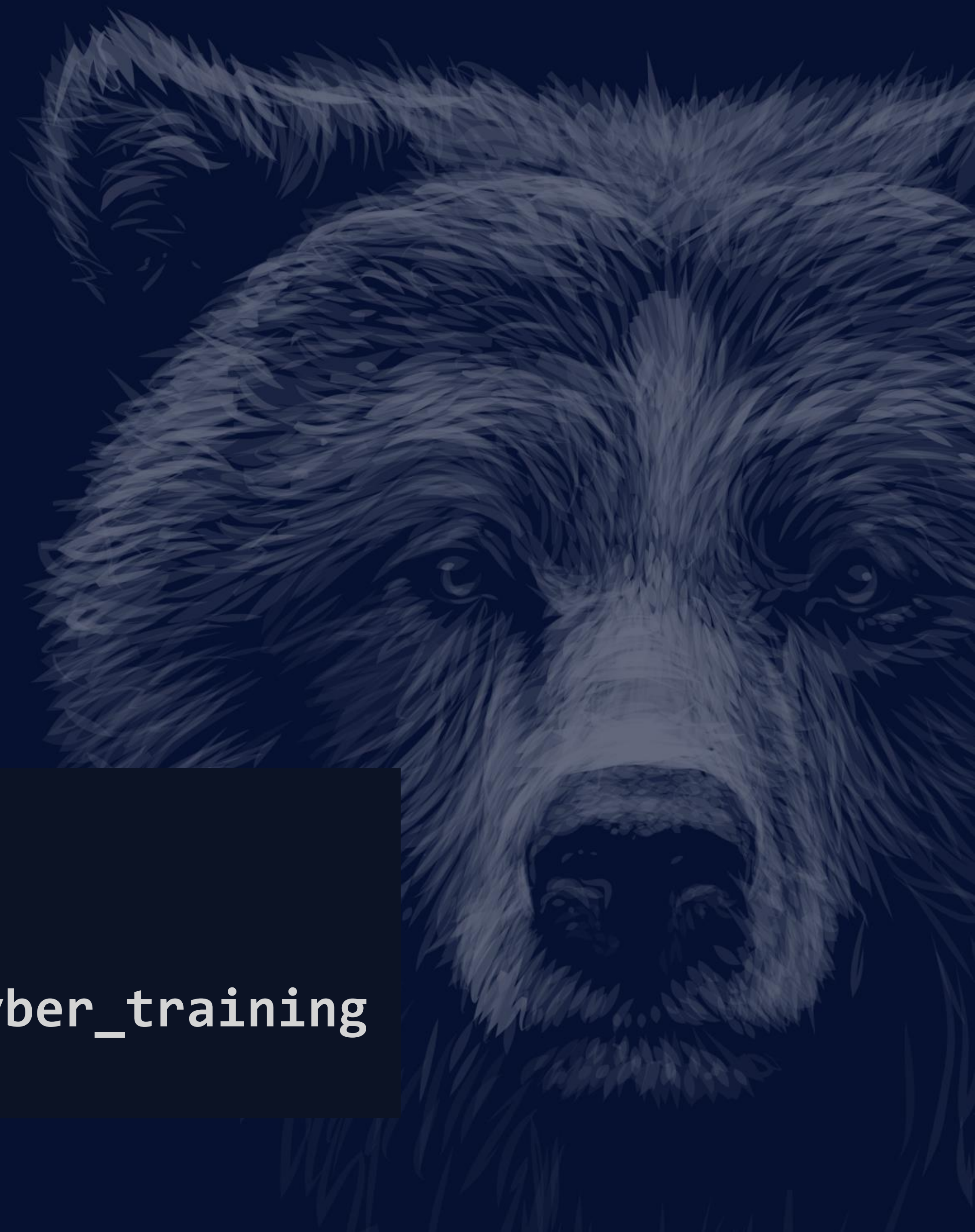
Training Track 1 – Intro to AI/ML

AI & ML Immersion Day

1. Https://colab.research.google.com/
2. Sign-in to Google
3. GitHub URL:

https://github.com/ChandlerProvence/cyber_training



BIGBEAR.AI

Welcome

AI & ML Immersion Day

BIGBEAR.AI

Chandler Provence
Data Scientist

Charles Ramsay
Principal Technologist and Lead Data Scientist

Joe Davis
Cybersecurity Research Scientist

AI and ML Immersion Day

Applied AI/ML Solutions for Cyber Data

Register Using the Link in the Post!

misi Maryland Innovation & Security Institute

Preparations

- Environment in Google Colab
- Includes:
  - Python
  - Jupyter notebook
  - scikit-learn 1.0
- Gitlab for the notebooks

Steps:

1. https://colab.research.google.com/

2. Sign-in to Google

3. GitHub URL:

https://github.com/ChandlerPrvence/cyber_training

AM Agenda

1. Introduction to AI
   - Data: Iris
   - Notebooks
     - 1_MachineLearning.ipynb
     - 2_DeepLearning.ipynb

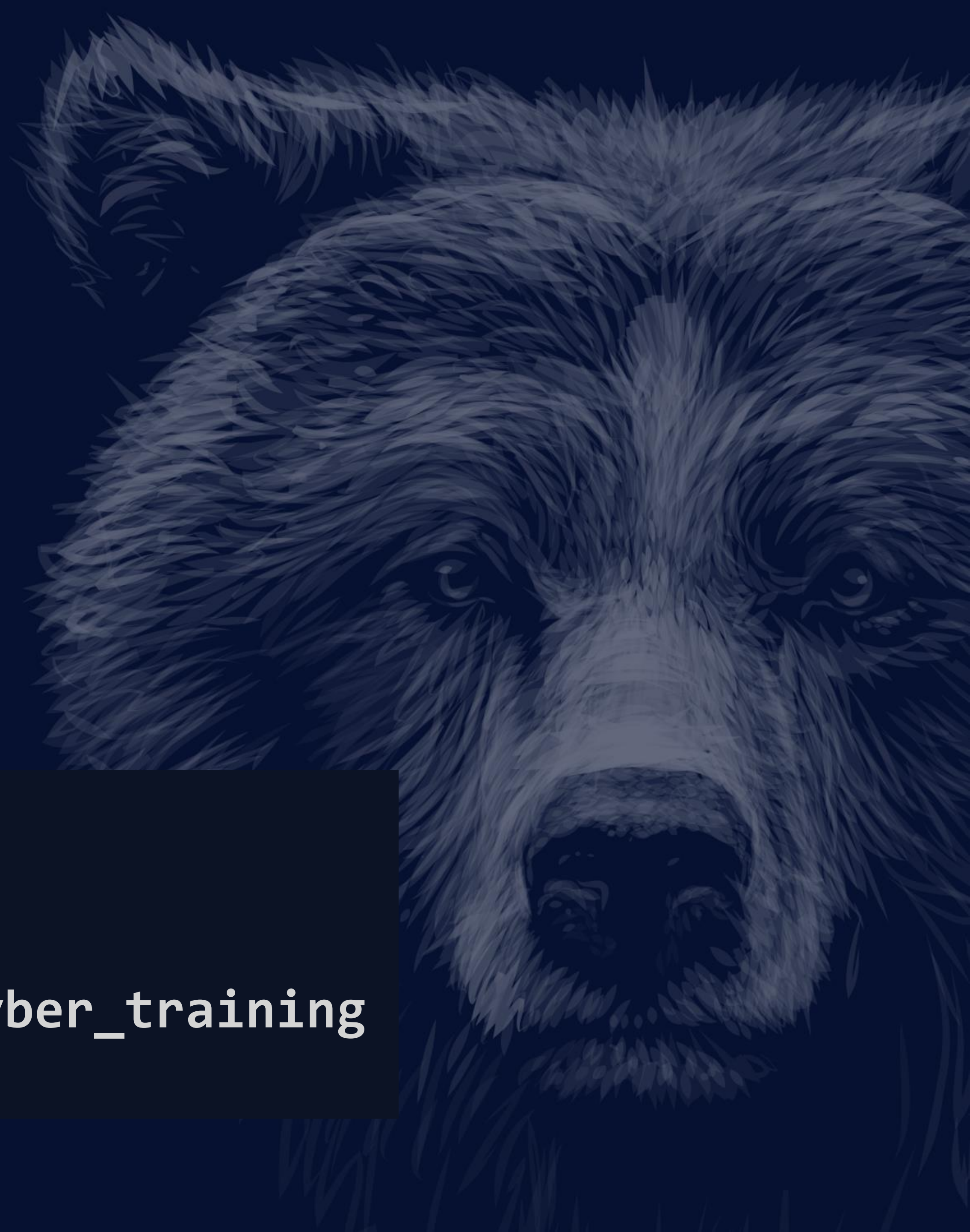
Training

AI & ML Immersion Day

Training Track 1 – Applied AI/ML Solutions for Cyber Data

AI & ML Immersion Day

1. Https://colab.research.google.com/
2. Sign-in to Google
3. GitHub URL:
https://github.com/ChandlerProvence/cyber_training



BIGBEAR.AI

Chandler Provence
Data Scientist

Charles Ramsay
Principal Technologist and Lead Data Scientist

Joe Davis
Cybersecurity Research Scientist

AI and ML Immersion Day

Applied AI/ML Solutions for Cyber Data

Register Using the Link in the Post!

misi Maryland Innovation & Security Institute

Preparations

- Environment in Google Colab
- Includes:
  - Python
  - Jupyter notebook
  - scikit-learn 1.0
- Gitlab for the notebooks

Steps:

1. https://colab.research.google.com/

2. Sign-in to Google

3. GitHub URL:

https://github.com/ChandlerPrvence/cyber_training

Afternoon Agenda

1. Traditional Cyber Analysis Tools and Techniques
2. Cyber Analysis Data Engineer and Exploratory Data Analysis: Use Case Walkthrough
3. Applied ML/AI Solutions for Cyber Data
   - Data: KDD 99 Updated
   - Notebooks
     - 3_IntrusionDetection. ipynb

Traditional Cyber Analysis Tools and Techniques

AI & ML Immersion Day

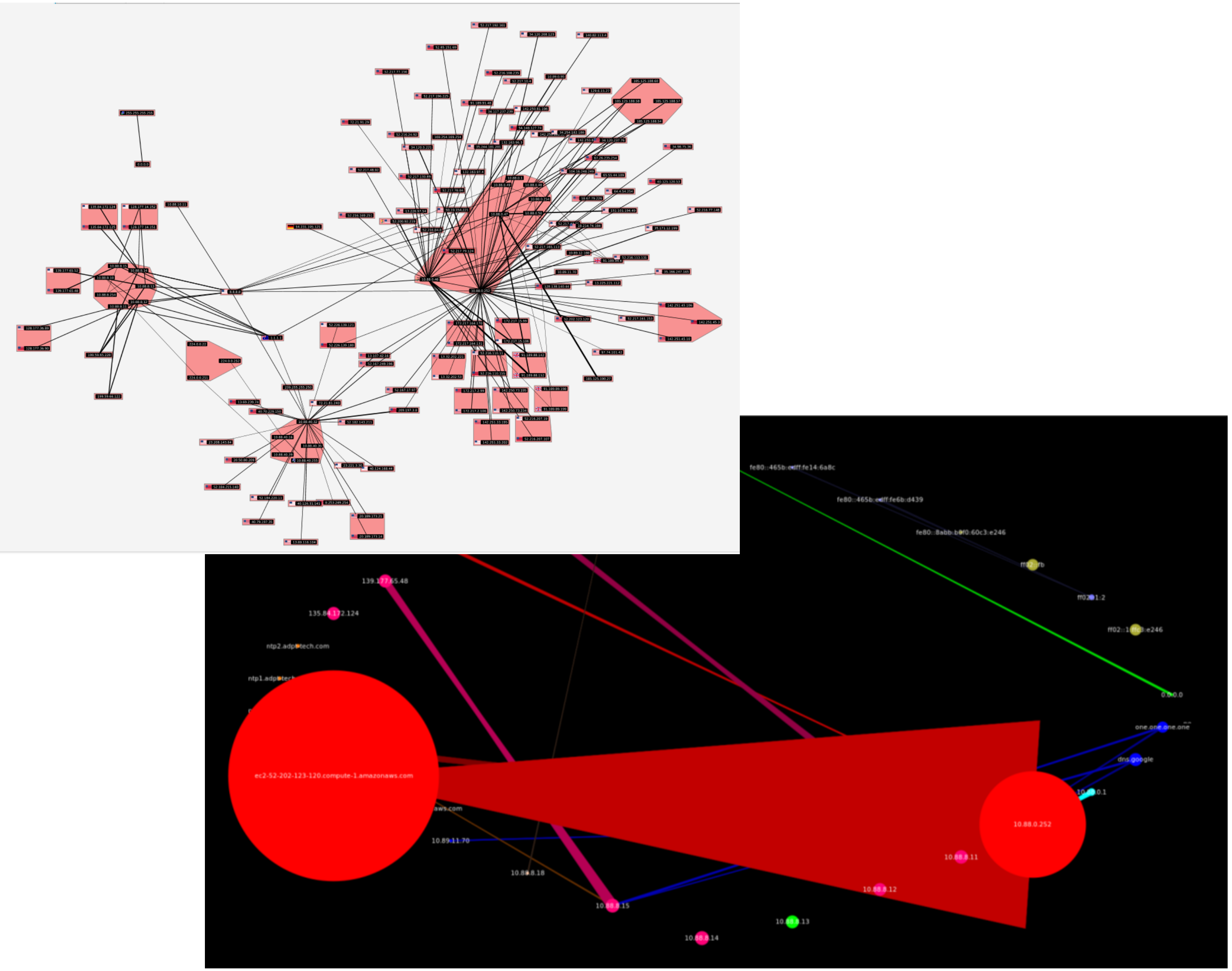
Dataset Used

~3.7G of known bad network traffic out of ~370G PCAP

```
known_bad-SF_GENERIC_SIGNATURE_ALERT_20220323_15324_mimikatz.pcap
known_bad-SF_GENERIC_SIGNATURE_ALERT_20220324_15845_mimikatz_dcsync.pcap
known_bad-SF_SIGNATURE_ENGINE_ALERT_20220323_15204_directory_traversal.pcap
known_bad-SF_SMB_EXPLOITATION_ATTEMPT_MS17_10_WANNACRY_20220323_15194_eternalblue.pcap
known_bad-htp-sniffer-1647948592.pcap
known_bad-htp-sniffer-1647988985.pcap
known_bad-htp-sniffer-1647988985_SMBBruteForce.pcap
known_bad-htp-sniffer-1648062056.pcap
known_bad-htp-sniffer-1648062056_SiemensStop.pcap
known_bad-htp-sniffer-1648072144_SiemensSolenoidAttack.pcap
known_bad-htp-sniffer-1648073282.pcap
known_bad-htp-sniffer-1648073501.pcap
known_bad-htp-sniffer-1648075793.pcap
known_bad-htp-sniffer-1648140792.pcap
known_bad-htp-sniffer-1648142445.pcap
known_bad-htp-sniffer-1648142445_PCModbusWrite.pcap
```

Cyber Analysis Tools

1. Wireshark
2. Strings
3. GrassMarlin
4. EtherApe



Use Case: Dreamport PCAP Exploratory

AI & ML Immersion Day

Processing

1. joincap to join 16 known bad PCAP files into 1
2. Zeek used to vectorize PCAP into log files, separated by PCAP filename
   - Include Geolocation information
   - Add MAC addresses
3. 0_Zeek_Dreamport_Known_Bad_ETL to transform Zeek log files into pandas dataframes
4. 1_Zeek_Dreamport_Known_Bad_EDA to cleanup and export a network for visualization
5. 2_Zat_Dreamport_Known_Bad_EDA to perform unsupervised ML
   1. IsolationForest
   2. Kmeans
6. Visualize results in Power BI, Gephi and Brim

Visualization and Analysis Tools

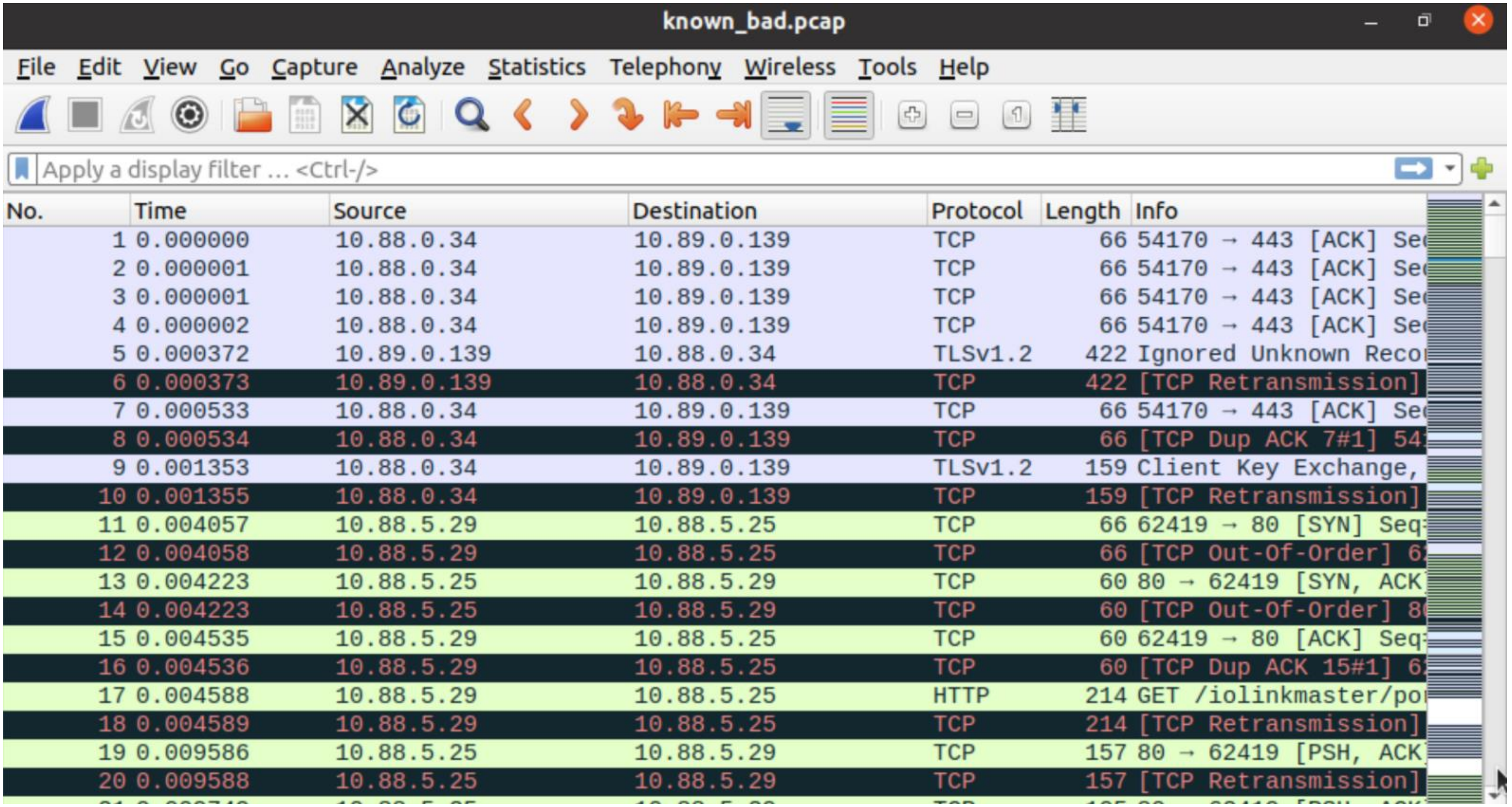
AI & ML Immersion Day

Wireshark

Used widely, the workhorse of PCAP forensic analysis.

https://www.wireshark.org/

Displaying 3.7G of Dreamport’s Known Bad PCAP



known_bad.pcap
File Edit View Go Capture Analyze Statistics Telephony Wireless Tools Help
Apply a display filter … <Ctrl-/>

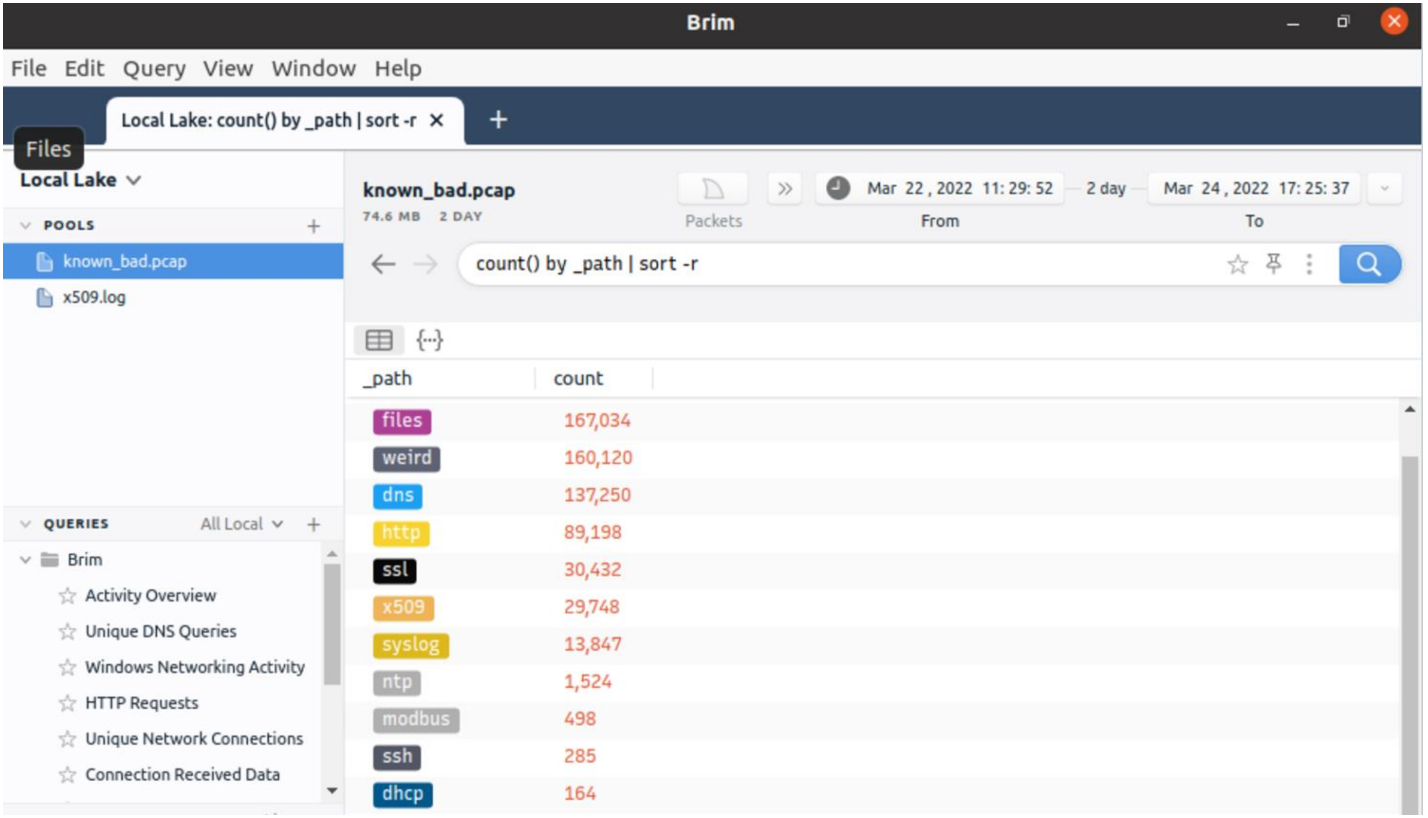
No.	Time	Source	Destination	Protocol	Length	Info
1	0.000000	10.88.0.34	10.89.0.139	TCP	66	54170 → 443 [ACK] Se
2	0.000001	10.88.0.34	10.89.0.139	TCP	66	54170 → 443 [ACK] Se
3	0.000001	10.88.0.34	10.89.0.139	TCP	66	54170 → 443 [ACK] Se
4	0.000002	10.88.0.34	10.89.0.139	TCP	66	54170 → 443 [ACK] Se
5	0.000372	10.89.0.139	10.88.0.34	TLSv1.2	422	Ignored Unknown Reco
6	0.000373	10.89.0.139	10.88.0.34	TCP	422	[TCP Retransmission]
7	0.000533	10.88.0.34	10.89.0.139	TCP	66	54170 → 443 [ACK] Se
8	0.000534	10.88.0.34	10.89.0.139	TCP	66	[TCP Dup ACK 7#1] 54
9	0.001353	10.88.0.34	10.89.0.139	TLSv1.2	159	Client Key Exchange,
10	0.001355	10.88.0.34	10.89.0.139	TCP	159	[TCP Retransmission]
11	0.004057	10.88.5.29	10.88.5.25	TCP	66	62419 → 80 [SYN] Seq
12	0.004058	10.88.5.29	10.88.5.25	TCP	66	[TCP Out-Of-Order] 6
13	0.004223	10.88.5.25	10.88.5.29	TCP	60	80 → 62419 [SYN, ACK
14	0.004223	10.88.5.25	10.88.5.29	TCP	60	[TCP Out-Of-Order] 8
15	0.004535	10.88.5.29	10.88.5.25	TCP	60	62419 → 80 [ACK] Seq
16	0.004536	10.88.5.29	10.88.5.25	TCP	60	[TCP Dup ACK 15#1] 6
17	0.004588	10.88.5.29	10.88.5.25	HTTP	214	GET /iolinkmaster/po
18	0.004589	10.88.5.29	10.88.5.25	TCP	214	[TCP Retransmission]
19	0.009586	10.88.5.25	10.88.5.29	TCP	157	80 → 62419 [PSH, ACK
20	0.009588	10.88.5.25	10.88.5.29	TCP	157	[TCP Retransmission]

Brim

"Brim is an open source desktop application for security and network specialists. Brim makes it easy to search and analyze data…"

https://github.com/brimdata/brim

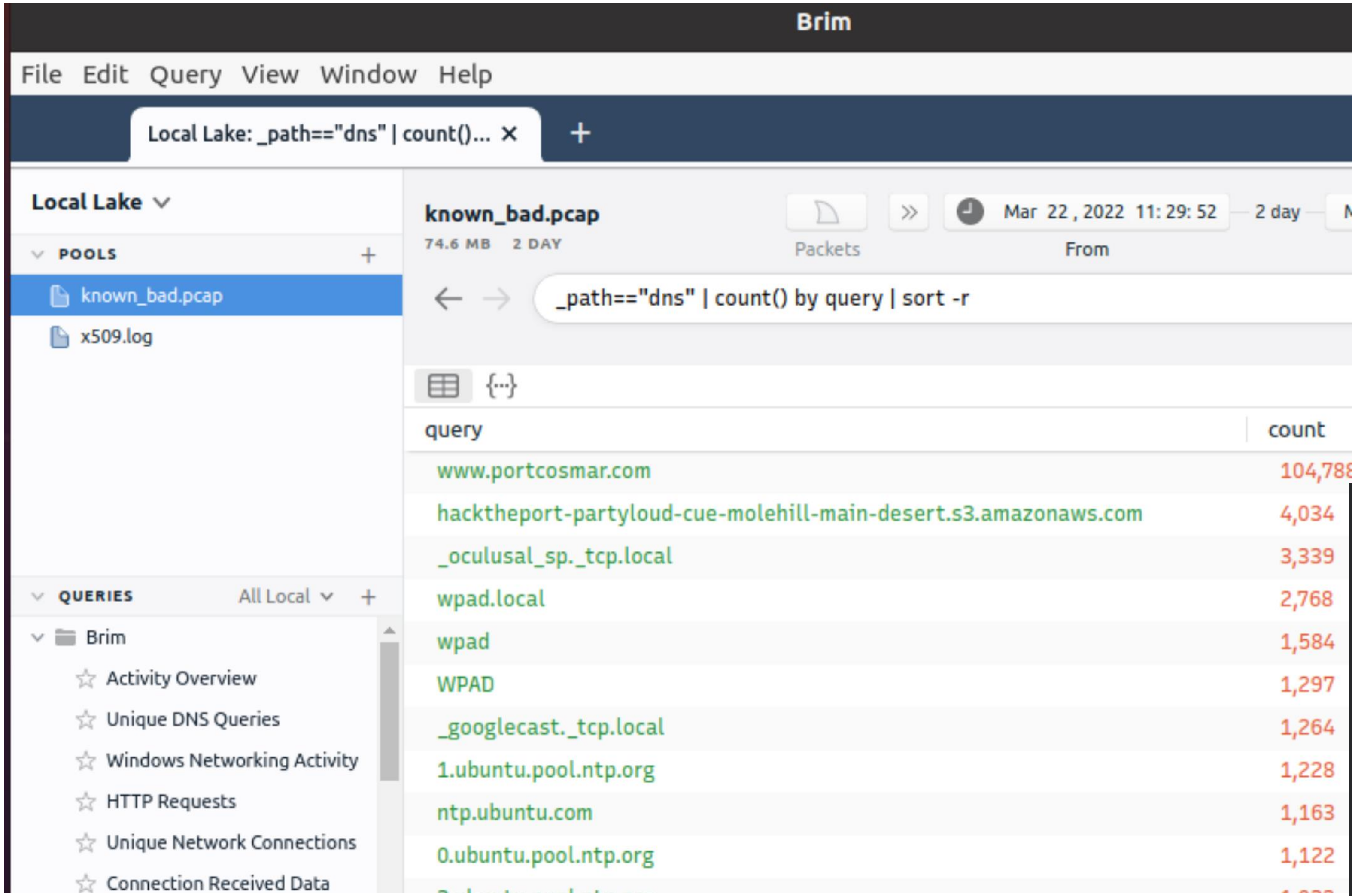
Displaying 3.7G of Dreamport's Known Bad PCAP



```
count() by _path | sort -r
```

_path	count
files	167,034
weird	160,120
dns	137,250
http	89,198
ssl	30,432
x509	29,748
syslog	13,847
ntp	1,524
modbus	498
ssh	285
dhcp	164

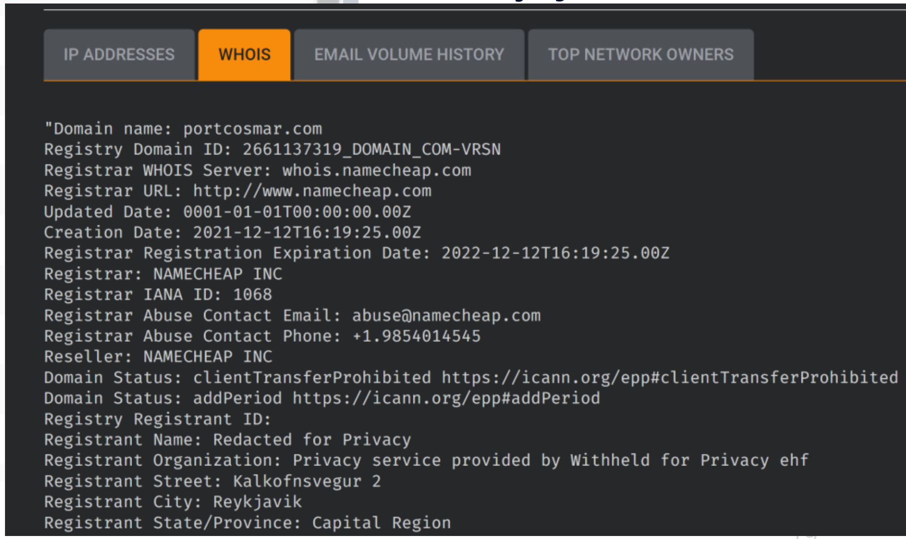
Brim – Ranked Unique DNS Queries



```
_path=="dns" | count() by query | sort -r
```

query	count
www.portcosmar.com	104,788
hacktheport-partyloud-cue-molehill-main-desert.s3.amazonaws.com	4,034
_oculusal_sp._tcp.local	3,339
wpad.local	2,768
wpad	1,584
WPAD	1,297
_googlecast._tcp.local	1,264
1.ubuntu.pool.ntp.org	1,228
ntp.ubuntu.com	1,163
0.ubuntu.pool.ntp.org	1,122

Portcosmar.com in Reykjavik, Iceland



```
"Domain name: portcosmar.com
Registry Domain ID: 2661137319_DOMAIN_COM-VRSN
Registrar WHOIS Server: whois.namecheap.com
Registrar URL: http://www.namecheap.com
Updated Date: 0001-01-01T00:00:00.00Z
Creation Date: 2021-12-12T16:19:25.00Z
Registrar Registration Expiration Date: 2022-12-12T16:19:25.00Z
Registrar: NAMECHEAP INC
Registrar IANA ID: 1068
Registrar Abuse Contact Email: abuse@namecheap.com
Registrar Abuse Contact Phone: +1.9854014545
Reseller: NAMECHEAP INC
Domain Status: clientTransferProhibited https://icann.org/epp#clientTransferProhibited
Domain Status: addPeriod https://icann.org/epp#addPeriod
Registry Registrant ID:
Registrant Name: Redacted for Privacy
Registrant Organization: Privacy service provided by Withheld for Privacy ehf
Registrant Street: Kalkofnsvegur 2
Registrant City: Reykjavik
Registrant State/Province: Capital Region
```

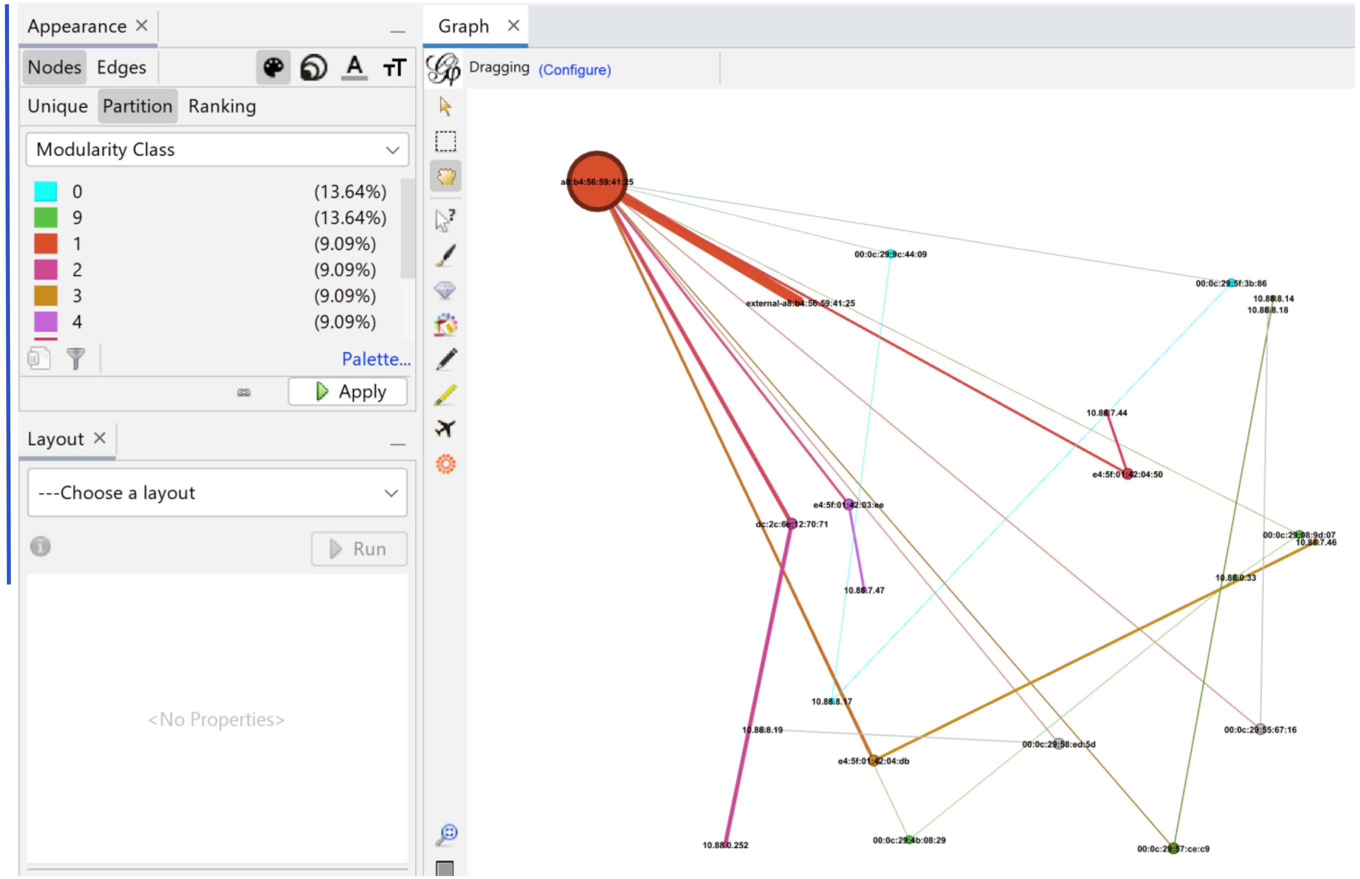
Gephi

Graphing analysis tool complete with statistical and graphing capabilities.

“Gephi is the leading visualization and exploration software for all kinds of graphs and networks. Gephi is open-source and free.”

https://gephi.org/

Displaying Known Bad top 100 hosts.



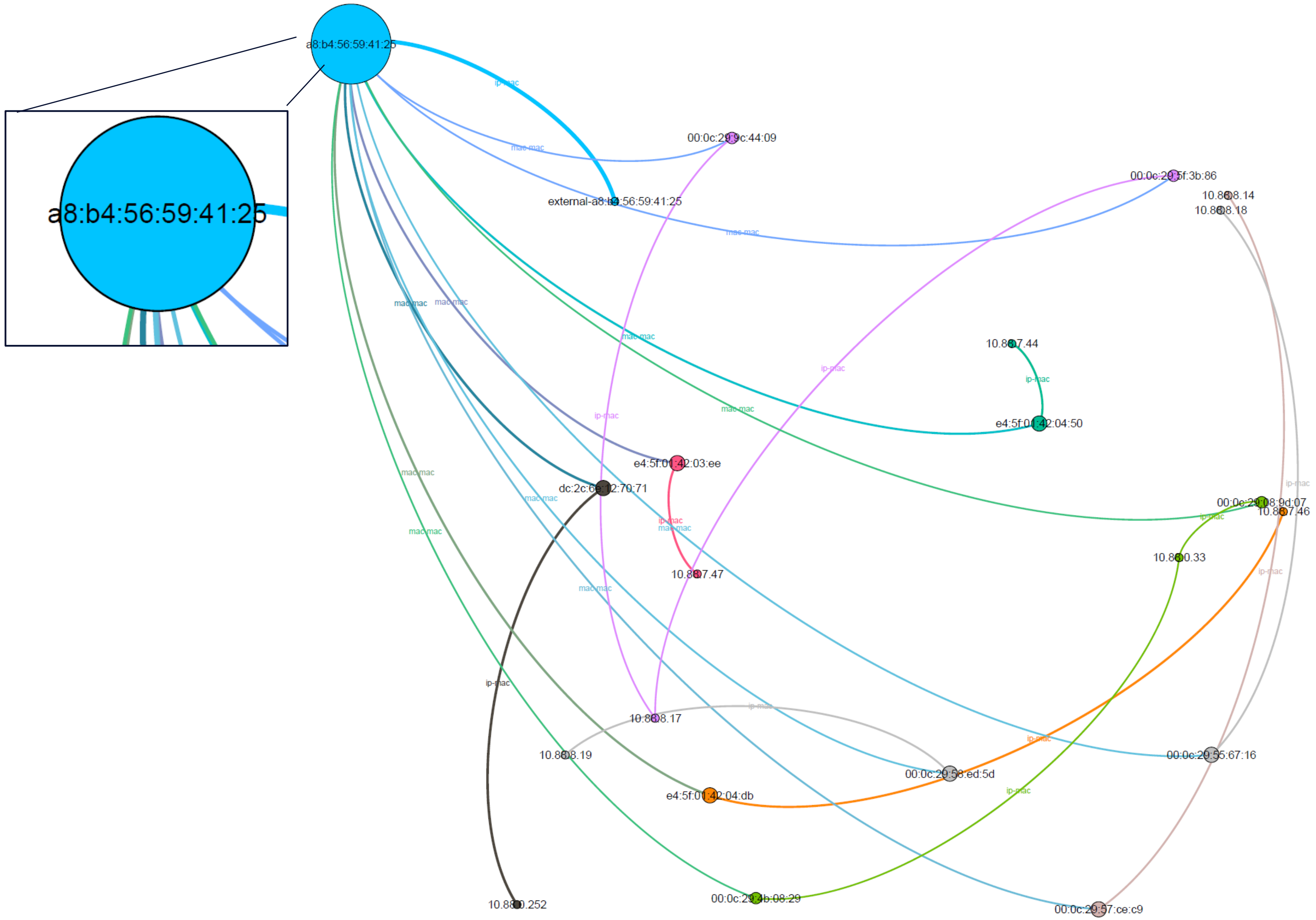
Modularity Class

🟦	0	(13.64%)
🟩	9	(13.64%)
🟥	1	(9.09%)
🟪	2	(9.09%)
🟧	3	(9.09%)
🟪	4	(9.09%)

---Choose a layout

<No Properties>

Gephi – Top 100 Known Bad Network Nodes



Source:

- Zeek conn.log

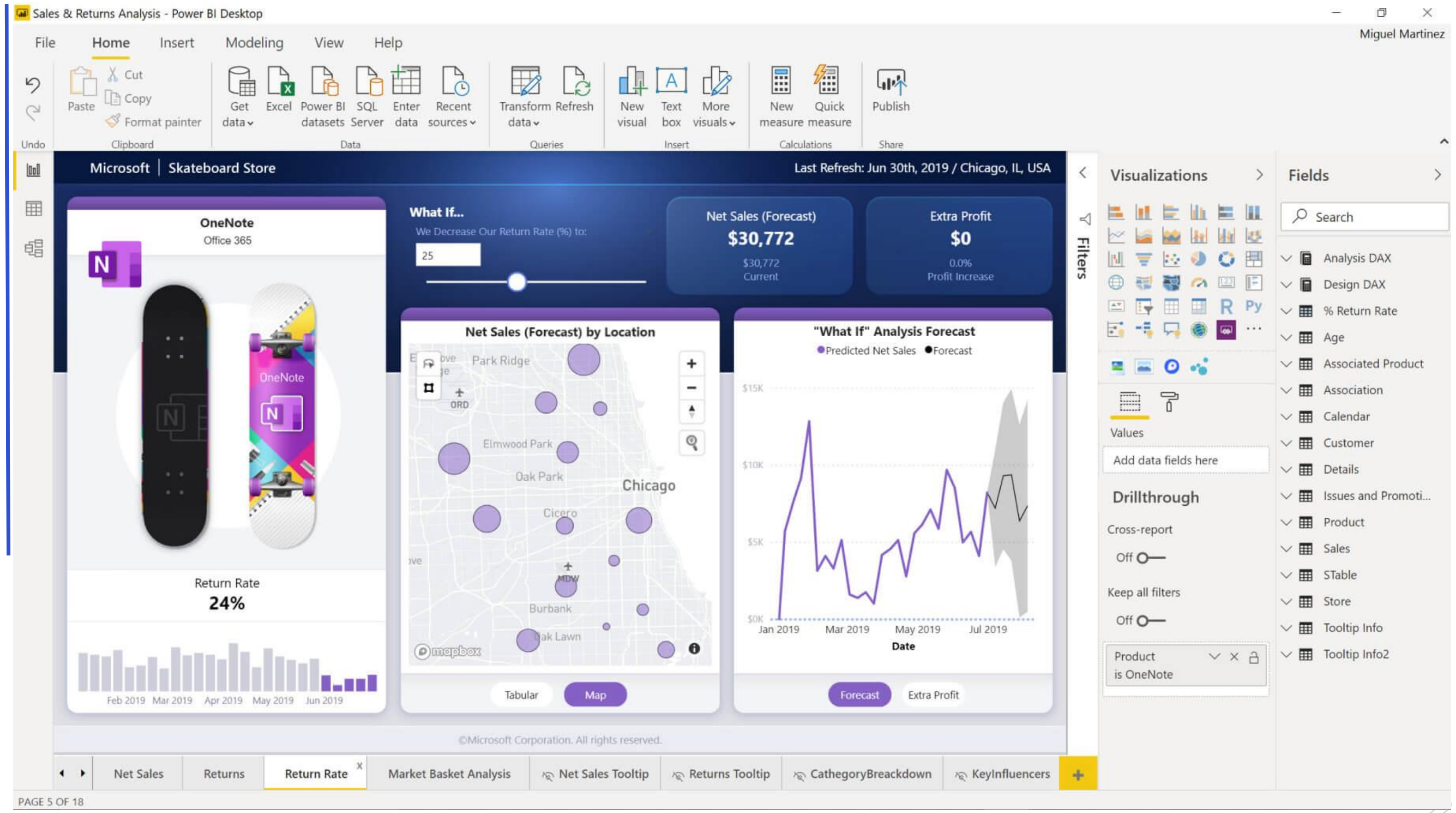
Shows:

- Modularity (Community) high modularity denotes many connections between nodes (as edges)
- Betweenness Centrality – denotes critical lines of communication

Power BI

Interactive dashboards with multiple charts, dashboards and data source adapters

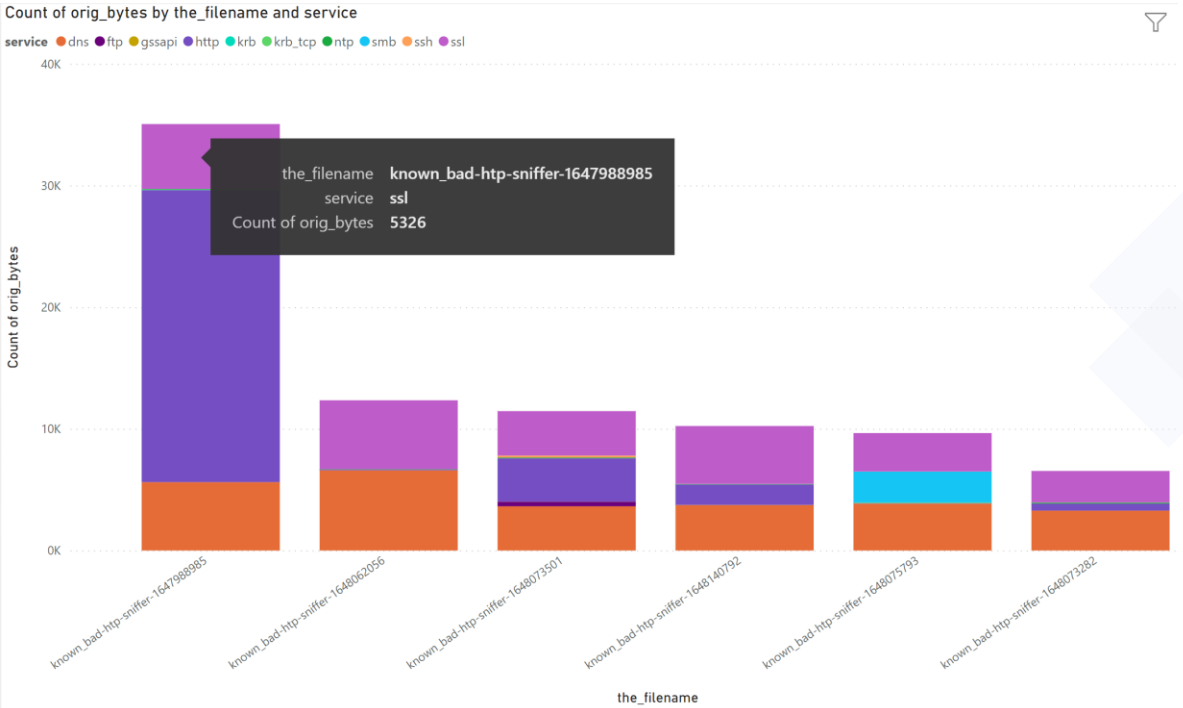
https://powerbi.microsoft.com/en-us/desktop/

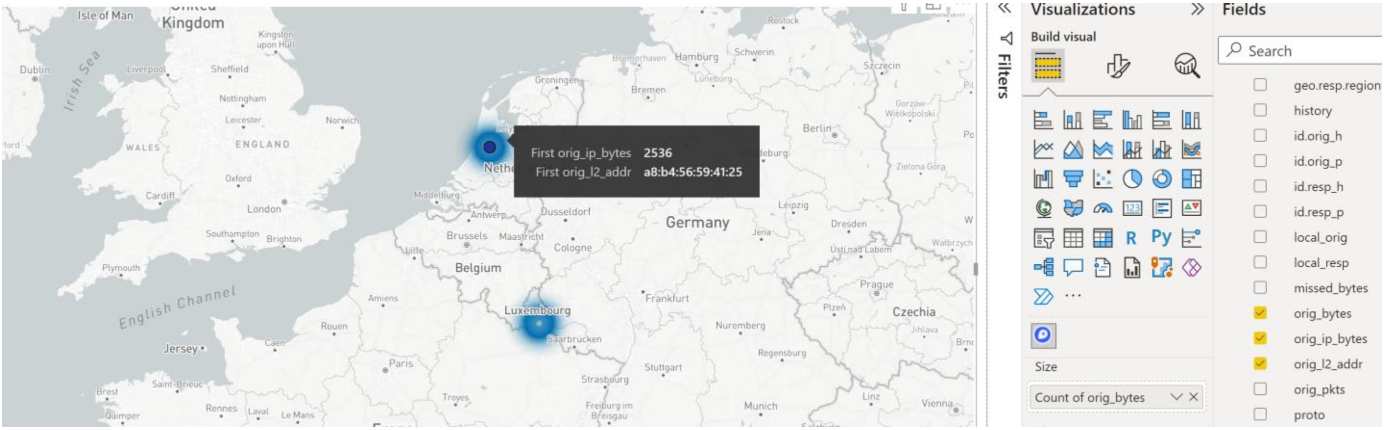


source: https://powerbi.microsoft.com/en-us/desktop

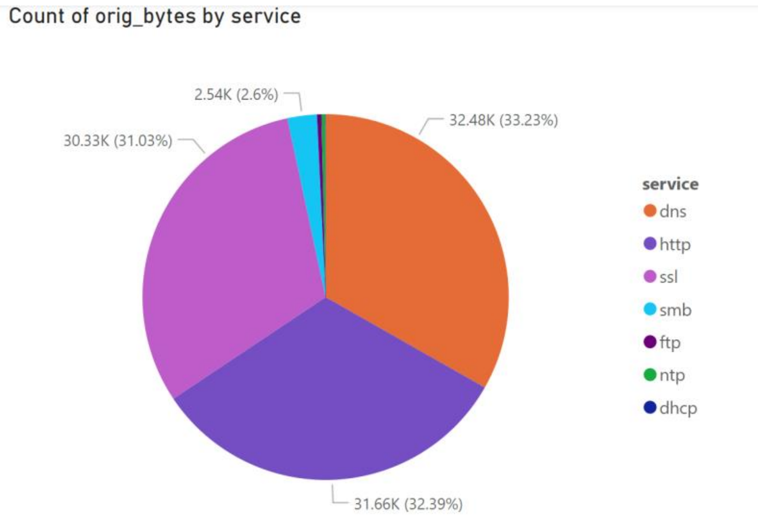
Power BI – Known Bad PCAP

- Source:
  - Zeek conn.log
- Shows:
  - Count of origin bytes by PCAP filename and service
  - Breakdown of orig_bytes by service type
  - Hostnames resolved to location





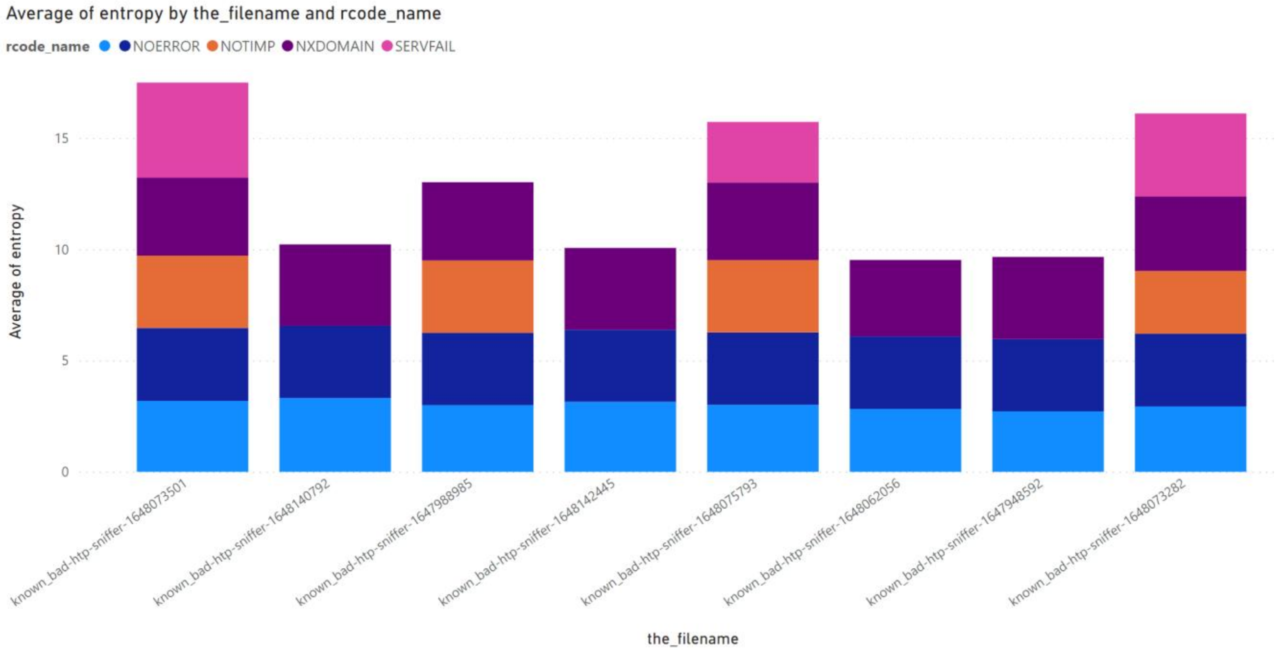
- ☐ geo.resp.region
- ☐ history
- ☐ id.orig_h
- ☐ id.orig_p
- ☐ id.resp_h
- ☐ id.resp_p
- ☐ local_orig
- ☐ local_resp
- ☐ missed_bytes
- ☑ orig_bytes
- ☑ orig_ip_bytes
- ☑ orig_l2_addr
- ☐ orig_pkts
- ☐ proto



Power BI – Known Bad PCAP

- Source:
  - Zeek dns.log
- Shows:
  - Entropy of DNS queries (names) w/ response code
  - This is a means to find Domain Generator Algorithms
  - https://en.wikipedia.org/wiki/Domain_generation_algorithm

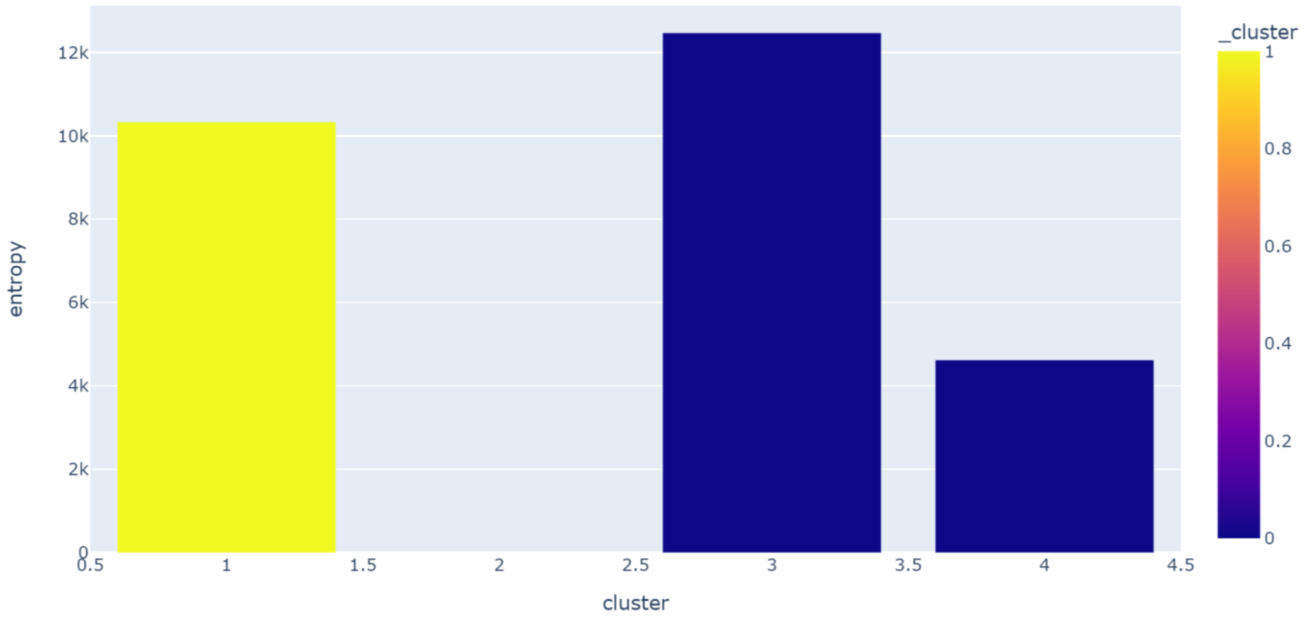
The entropy is approx the same for different RCodes



Average of entropy by the_filename and rcode_name

HDSCAN – Clustering with DNS Entropy

- Source:
  - Zeek dns.log
- Shows:
  - Entropy of DNS queries (names) clustered using HDBSCAN



Training

AI & ML Immersion Day

Thank you!

AI & ML Immersion Day