Data Sources for Cyber Security Research

Melissa Turcotte

Advanced Research in Cyber Systems,
Los Alamos National Laboratory

25 September 2017

Acknowledgements

- Alex Kent, Los Alamos National Laboratory
- Curtis Hash, Ernst & Young
- Aaron McPhall, Los Alamos National Laboratory
- Neale Pickett, Los Alamos National Laboratory

Advanced Research in Cyber Systems, LANL

Cyber research group at large DOE national research laboratory.

- Identify gaps, emerging threats
- Data collection and sensors
- Modelling and simulation
- Tool/product development

Our perspective

Assume the initial compromise will happen; consequently largely focused on inside-the-perimeter detection.

Internal Cyber Data from ∼15K Computers (+ Other Devices)

- Network flow records (internal+external)
  - 600 million events/day, 13.9 GB compressed - since 2003
- Windows desktop event log records
  - 560 million events/day, 28.6 GB compressed - since 2008
- Internal DNS records
  - 150 million events/day, 5.0 GB compressed - since 2008
- Web proxy records
  - 62 million events/day, 5.4 GB - since 2005
- Email event metalog records
  - 320,000 events/day, 120 MB compressed - several years

∼ 1.4 Billion Events as ∼ 53 GB compressed per day collected

Cyber Data Challenges

- Where collected, not necessarily for cybersecurity purposes.
  - Debugging.
- Where collected, not intended for analytics.
  - Primary purpose is to support human operations.
- Not useful unless combined with other data sets.
- Collection is often incomplete.
- Collection, storage, and use has security, privacy, and human subject research implications.

Public Data Sets

- To motivate a larger research effort focused on operational cyber data LANL have released three publicly available datasets:
  - 9 month time-series user/computer bipartite, 2014
  - 58 day comprehensive, 2015
  - 90 day Netflow and Window Event Logs, 2017
- Internally collected from the LANL corporate network:
  - No outside (Internet) associations
  - Parsed and normalised (to some extent)

```
https://csr.lanl.gov/data/
https://csr.lanl.gov/data/2017.html
```

Unified Host and Network Dataset

- Netflow and Window Event Logs over 90 days.
  - Netflow Data Set $\rightarrow$ 145G
  - Windows Event Logs $\rightarrow$ 62G
- Identifying values de-identified (anonymized).
- De-identified values match across both data sets.
- Well-known network ports, system-level usernames, processes and core enterprise hosts not de-identified.
- Small set of hosts, users and processes were combined where they represented well-known redundant entities.

Network Data Set

Raw data consisted of NetFlow records exported from core network routers to a centralised collection server (limited to Protocols 1 (ICMP), 6 (TCP), and 17 (UDP)).

StartTime, EndTime, SrcIP, DstIP, Protocol, SrcPort, DstPort, Packets, Bytes

Modelling challenges

- Flows are uni-directional (uniflows) → no relationship between direction of a flow record and the initiator of a bi-directional connection.
- Duplication → flows can pass multiple routers, routers can be configured to track flows on ingress and egress.
- Lack of stable identifiers for network devices upon which to build models → IP addresses are transient (DHCP).

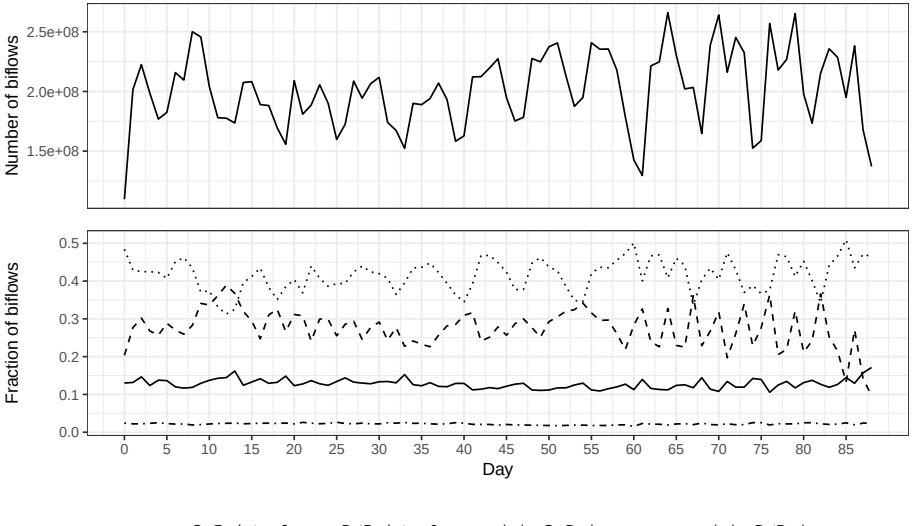
Data Transformation

- Bi-flowing
  - Aggregate duplicates
  - Marry opposing uniflows of bi-directional connections into a single, directed biflow record
- IP/HostName mapping
  - Daily snapshot of device inventory
  - DHCP logs
  - Failed mappings anonymized as IPx rather than Comp$x$

Time, Duration, SrcDevice, DstDevice, Protocol, SrcPort, DstPort, SrcPackets, DstPackets, SrcBytes, DstBytes

761,4434,Comp132598,Comp817788,6,Port12597,22,89159,85257,15495068,69768940
764,13161,Comp178973,Comp164069,17,137,137,325,0,30462,0
765,14369,Comp492856,Mail,6,Port30344,443,227,214,32300,9844
765,14431,Comp782574,Mail,6,Port28068,443,1637,3313,75302,1220077
765,17056,Comp378125,Mail,6,Port28068,443,3848,4096,177008,1441295
118785,14178,IP564116,Comp141988,17,5060,5060,1866,0,1477041,0

$\sim$ 58k named devices (Comp$x$), $\sim$ 873k failed name mappings (IPx)



Host Logs

- Network-only detection mechanisms are becoming less effective.
- Cyber defenders now rely heavily on endpoint agents and host event logs to detect and investigate security incidents.
- Host event logs capture nuanced details about what is happening at a machine level:
  - authentication, logons
  - processes
  - applications/services
- Relatively little attention has been given to analysing these data so far $\rightarrow$ not readily available.
- Associated challenges when using these data $\rightarrow$ parsing, and extracting relevant attributes an important first step (not intended for analytics).

Windows Host Log Data

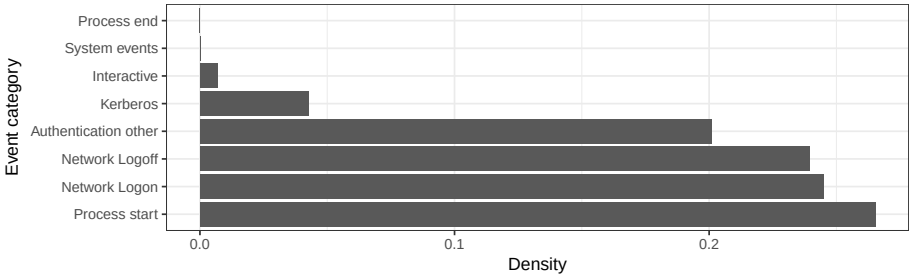
Subset of host event logs collected from computers running the Microsoft Windows operating system.

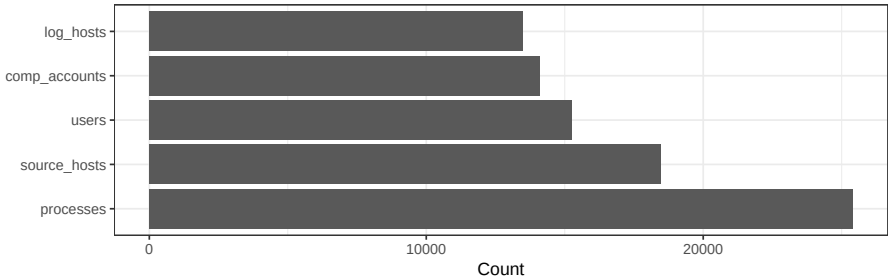
- Events only related to authentication and process activity on each machine.
- JSON format, one record per line → preserve structure of original events.
- Each record has an EventID which uniquely identifies the event.
- Not all events share the same set of attributes → event dependent.
- All records contain *EventID, LogHost* and *Time*.

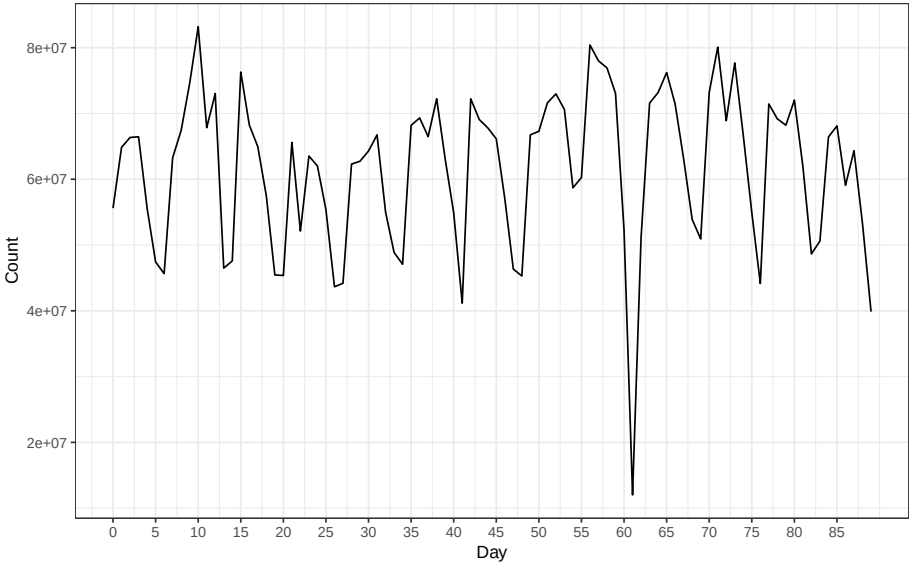
{"UserName": "Comp916004$", "EventID": 4624, "LogHost":"Comp916004", "LogonID": "0x1b5b5753", "DomainName": "Domain001", "LogonTypeDescription": "Network", "Source": "Comp916004", "AuthenticationPackage": "Kerberos", "Time": 109, "LogonType": 3}

{"UserName": "User186321", "EventID": 4648, "LogHost": "Comp916004", "LogonID": "0x3e4", "DomainName": "Domain001", "Destination": "Comp457365", "SubjectUserName": "Comp916004$", "ProcessName": "Proc423620.exe", "SubjectLogonID": "0x3e4", "Time": 108, "SubjectDomainName": "Domain001", "ProcessID": "0xd28"}

{"UserName": "User186321", "EventID": 4624, "LogHost": "Comp916004", "LogonID": "0x1abd30dd", "DomainName": "Domain001", "Source": "Comp916004", "LogonTypeDescription": "NetworkClearText", "ProcessName": "Proc423620.exe", "AuthenticationPackage": "Negotiate", "Time": 108, "LogonType": 8, "ProcessID": "0xd28"}

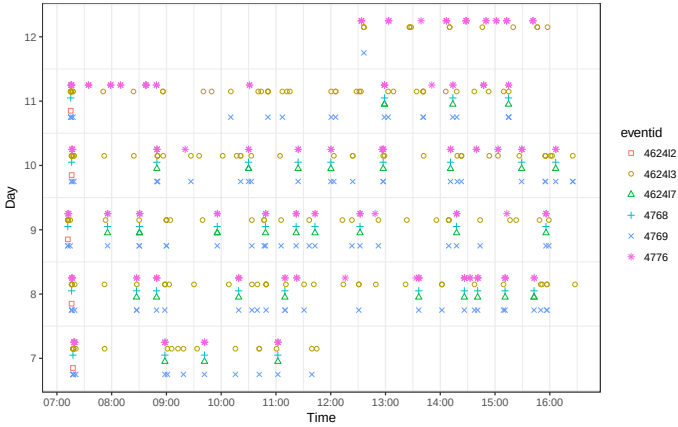

Process end
System events
Interactive
Kerberos
Authentication other
Network Logoff
Network Logon
Process start
Event category
0.0
0.1
0.2
Density


log_hosts
comp_accounts
users
source_hosts
processes
0
10000
20000
Count


8e+07
6e+07
4e+07
2e+07
Count
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
Day

Modelling Challenges

- Periodicity → computer regularly renewing credentials
- Correlated event types



Event times for a user in the LANL network

Research Opportunities

- For a given entity, extrapolate to higher-level more interpretable actions.
  - Remove or separate computer-driven events and correlations.
- Per computer, user and edge models that enable anomaly detection (utilise host logs).
  - Identify user and computer types and differences.
  - Community detection approaches, peer-based analyses.
- Explore process data in detail (little done here).
  - In particular, process trees.

{"UserName": "Comp646113$", "EventID": 4688, "LogHost": "Comp646113", "LogonID": "0x3e7", "DomainName": "Domain001", "ParentProcessName": "svchost", "ParentProcessID": "0x2178", "ProcessName": "wermgr.exe", "Time": 6905, "ProcessID": "0x6bc8"}

- Characterise risk within the network.
  - Host-user risk analysis.
  - Time-delay metrics on network penetration.
  - Network segmentation.
- Meaningful analyses that combine the two data sets.

How do we make anomaly detection more practical?

Data analytics and anomaly detection is very immature (and uncommon) in actual cybersecurity operations.

- Current cyber defenders generally discount these approaches in favour of signature detection and intuition.

Anomaly detection approaches suffer from:

- High false positives
- Lack of interpretability
- Inability to triage

Three common characteristics of good leads: relevant, detailed, and actionable.

"Incident Response & Computer Forensics", Luttgens et al.

Future Data Sets

Feedback can help us improve future data releases.

- What format do you find useful?
- What features and labels are most useful?
- What data sources?
- What time frames and lengths?

cyberdata@lanl.gov

`https://csr.lanl.gov/data/`
`https://csr.lanl.gov/data/2017.html`