Lessons Learned from 10 Years of Network Analysis R&D for Defense and Intel Customers

Thayne Coffman
FloCon 2012
Austin, TX
The Speaker’s Perspective

- **21CT**
  - 12 years old, 90 ppl., Austin/SA/DC
  - Broad-spectrum R&D for DoD & IC
  - Now focused on applying LYNXeon™ graph analytics to flow data for USG & commercial

- **Me**
  - CS, AI, signal processing, pattern classification
  - 10 years @ 21CT: research, mgmt, strategy
  - Work marries graphs, signals, cyber, SNA, classification

- “Network” analysis == social or cyber
- Nobody is omniscient
Executive Summary

1. Analysts need tools that enable flexible workflows
2. Analysts need tools that run mid-complexity analytics
3. Anomaly detection is worth continued investment, but it will never be the whole answer
1. Analysts need tools that enable flexible workflows

2. Analysts need tools that run mid-complexity analytics

3. Anomaly detection is worth continued investment, but it will never be the whole answer
Network Analytics for Intel. & Cyber

Net analytics concept (intel)

1st gen proto. (intel)

Net analytics concept (cyber)

1st gen proto. (cyber)

2nd gen operational POC (cyber)

LYNXeon analyzes 1B flows

LYNXeon operational use (intel)

1998: CERT established

Book: Small Worlds

1988: DARPA graph analytics programs

9/11 Attacks

US-CERT established

LYNXeon operational use (intel)

1st FloCon

NetFlow v5 broad support

US-CERT established

Saddam Hussein capture via SNA

Book: Understanding Terror Networks

CYBERCOM established

LYNXeon GA release & operational use (cyber)

Death of Usama bin Laden

SNA is now a staple in intel analysis

Cyber network analysis is now mainstream

21CT has matured capabilities in both areas
Lesson 1: The Problem

- Too much data to search & understand unaided
  (Severe challenges in even automated processing)
- Too many attacks to run to ground
- Urgent need for deeply buried answers
Lesson 1: Doing it Wrong

- Try to take the analyst out of the loop
- Massive, inflexible, automated, integrated data mining “solutions”
- Fixed workflows built around standing queries

\[ P(F+) = 0.001\% \times 10^9 \text{ flows} = 10^4 \text{ false positives. Now what?} \]
Lesson 1: Doing it Right

Analysts need tools that enable flexible workflows.

- Embrace an analyst-centric iterative process
  - Avoid hardcoded analytics & workflows
  - Sandbox tools – i.e., platforms
  - Minimize timespan of: ideas/workflows → prototype analytics → reusable tools
  - Distill, mature, scale, apply, integrate, catalog, and share analytics
1. Analysts need tools that enable flexible workflows

2. Analysts need tools that run mid-complexity analytics

3. Anomaly detection is worth continued investment, but it will never be the whole answer
Cat and Mouse in a Changing World

The environment keeps changing
Attacks & attackers keep changing
Tools are constantly changing to keep up

Snort

ArcSight v1.0
21CT 1st gen tool released
SiLK v0.1

1998
2000
2002
2004
2006
2008
2010
2012

Facebook
Twitter
NetFlix free streaming
Anonymous: NGO political attacks
Stuxnex: SCADA

Titan Rain: state sponsored?
Caribe: mobile devices
Zus: financial theft

SiLK v1.0
21CT 2nd gen POC operational
SiLK v2.4.5
LYNXeon GA release & operational use

The environment keeps changing
Attacks & attackers keep changing
Tools are constantly changing to keep up

Facebook
Twitter
NetFlix free streaming
Anonymous: NGO political attacks
Stuxnex: SCADA

Titan Rain: state sponsored?
Caribe: mobile devices
Zus: financial theft

SiLK v1.0
21CT 2nd gen POC operational
SiLK v2.4.5
LYNXeon GA release & operational use
Lesson 2: The Problem

- Unexpected changes in environment and attacks
- Signatures only catch what they’re looking for
- Anomaly detection doesn’t fill all the gaps “yet”
Lesson 2: Doing it Wrong

- Try to make your signatures flexible

- Contract murders example
  - $10^4$-$10^5$ elements to search
  - Multi-level complex patterns
  - Matches 1.3M variations
  - …and inexact matching

- That’s flexible enough, right?
The Intelligence Analysis Bathtub

- Massive systems = accept the bathtub (but don’t say that)
- “Flexible patterns” = accept the bathtub (but don’t say that)
- How do we really invert the bathtub?
Lesson 2: Doing it Right

Analysts need tools that run mid-complexity analytics.

- Too small = return to overload
- Just right = simple correlations
- Too big = never flexible enough
- Combine with flexible workflows
  - Bite-sized fast & scalable analytics
  - Analyst builds ad hoc analysis chains based on task, attack, & data exploration
  - Run, see results, augment/pivot, repeat
- Embrace and enable the analyst in the loop
1. Analysts need tools that enable flexible workflows
2. Analysts need tools that run mid-complexity analytics
3. Anomaly detection is worth continued investment, but it will never be the whole answer
A Brief History of Time Anomaly Detection

1998
1994+: Host AD w/ histograms & profiling

2000
1986+: Network AD w/ histograms & profiling

2002
w/ neural networks

2004
w/ parametric statistics

2006
w/ SOMs and clustering
w/ using spectral & dim. reduction techniques

2008
w/ SNA metric features (patented)
w/ using context
w/ using human heuristics

2010
2012

AD has been a goal for over 25 years.

Still lots of room to grow.

21CT has contributed novel approaches to AD.
Lesson 3: The Problem

- Can anomaly detection fill the detection gap?
- Changing environments, tactics, attacks, and data
- Too much data, and too little
- The smart adversaries try to look normal

A.D. HAPPY!  A.D. SAD!
Lesson 3: Doing it Wrong

- Rely on AD as an auto-magic detector that finds (only) bad people
  - \( P(F+) \) will never be zero
  - Many technical challenges remain: training data, generality, flexibility

- Accepts the bathtub, once again

- True generalized AD == a human, strong AI, or oracle
Lesson 3: Doing it Right

Anomaly detection is worth continued investment, but it will never be the whole answer.

- Inherent gaps point back to analyst-centric model
- Use for analyst cueing like other detectors
- Still lots of room to grow
- Consider these 4 ideas...
Lesson 3.1: Look for Better Features

- Traditional features == communication quantity
- Social network analysis metrics == communication structure
Lesson 3.2: Leverage Context

- Flexibly pull in external context data (hard)
- Condition training data
- Then cluster & group
Lesson 3.3: Leverage Domain Expertise

- Leverage analyst expertise to locally modify sensitivity
- Makes anomaly detection more adaptive

21CT prototype built under AFRL anomaly detection research effort
Lesson 3.4: Manage Dimensions and Data

- Submanifold learning & dimensionality reduction
- Sparse representations, sparse matrix completion
Conclusions

1. Analysts need tools that enable flexible workflows
   – Human must be inside the loop, and needs help
   – One workflow will never fit all

2. Analysts need tools that run mid-complexity analytics
   – Hand-in-hand with flexible workflows
   – Truly inverts the bathtub

3. Anomaly detection is worth continued investment, but it will never be the whole answer
   – Lots of room to grow and value to add
   – But full AD means a human or strong AI
Questions & Discussion

For future questions, contact:
Dr. Thayne Coffman
Chief Technology Officer
21CT
tcoffman@21technologies.com