

Novel, Large Scale, High Speed Text and Data Analysis Research



Presented by

Thomas E. Potok, Ph.D

Group Leader

Applied Software Engineering Research



Who we are



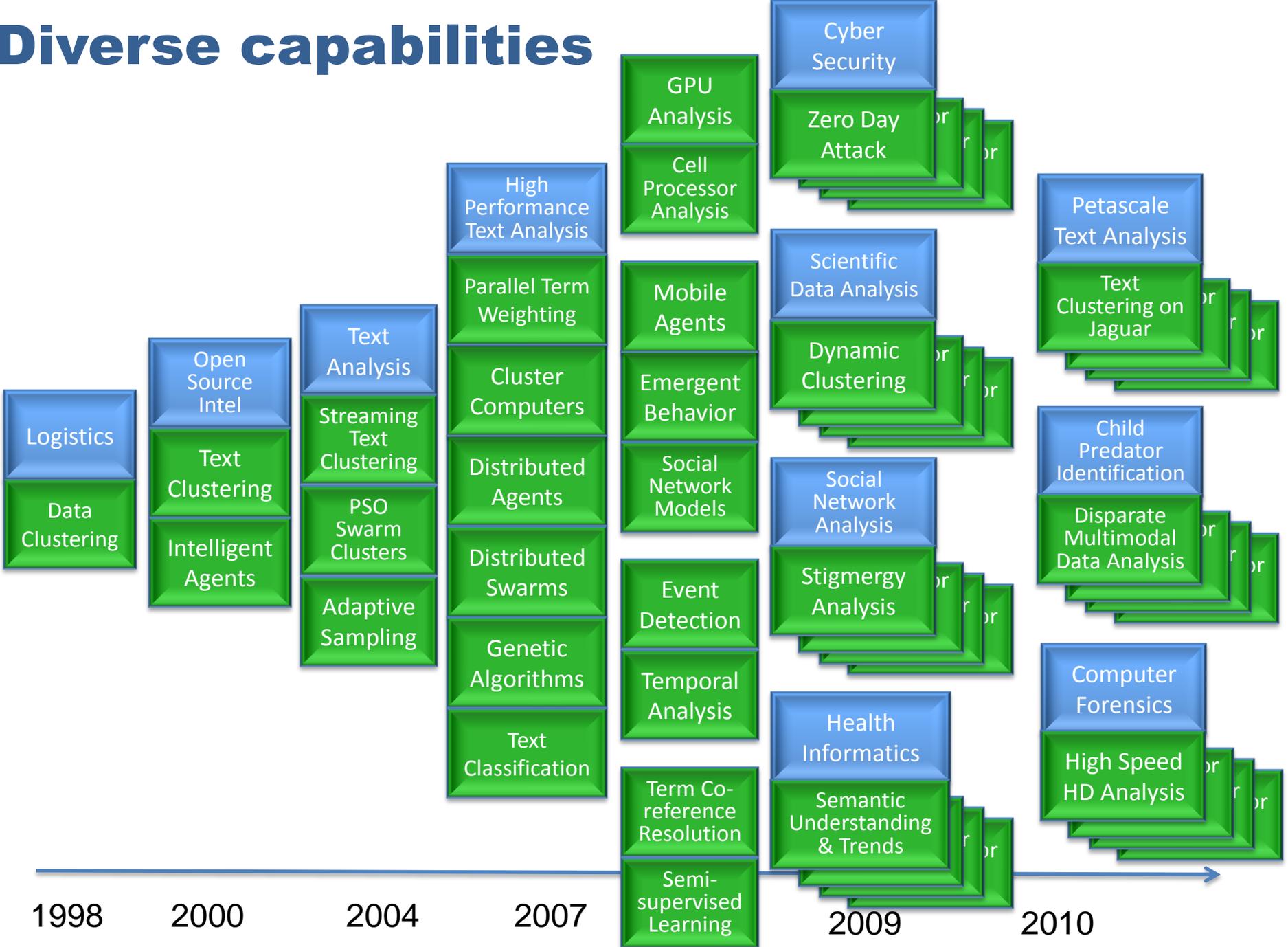
- **Group focus**

- 9 years of research in intelligent agents, emergent behavior, pervasive computing, machine learning, information retrieval, and knowledge discovery
- In 2009 we organized 7 research workshops and published 26 papers and book chapters
- Hands-on experience with DHS, Military, IC, and Industry

- **Group success**

- \$15M in research investment
- 14 staff members, 9 PhDs in computer science
- R&D 100 Award (Oscars of invention) in 2007

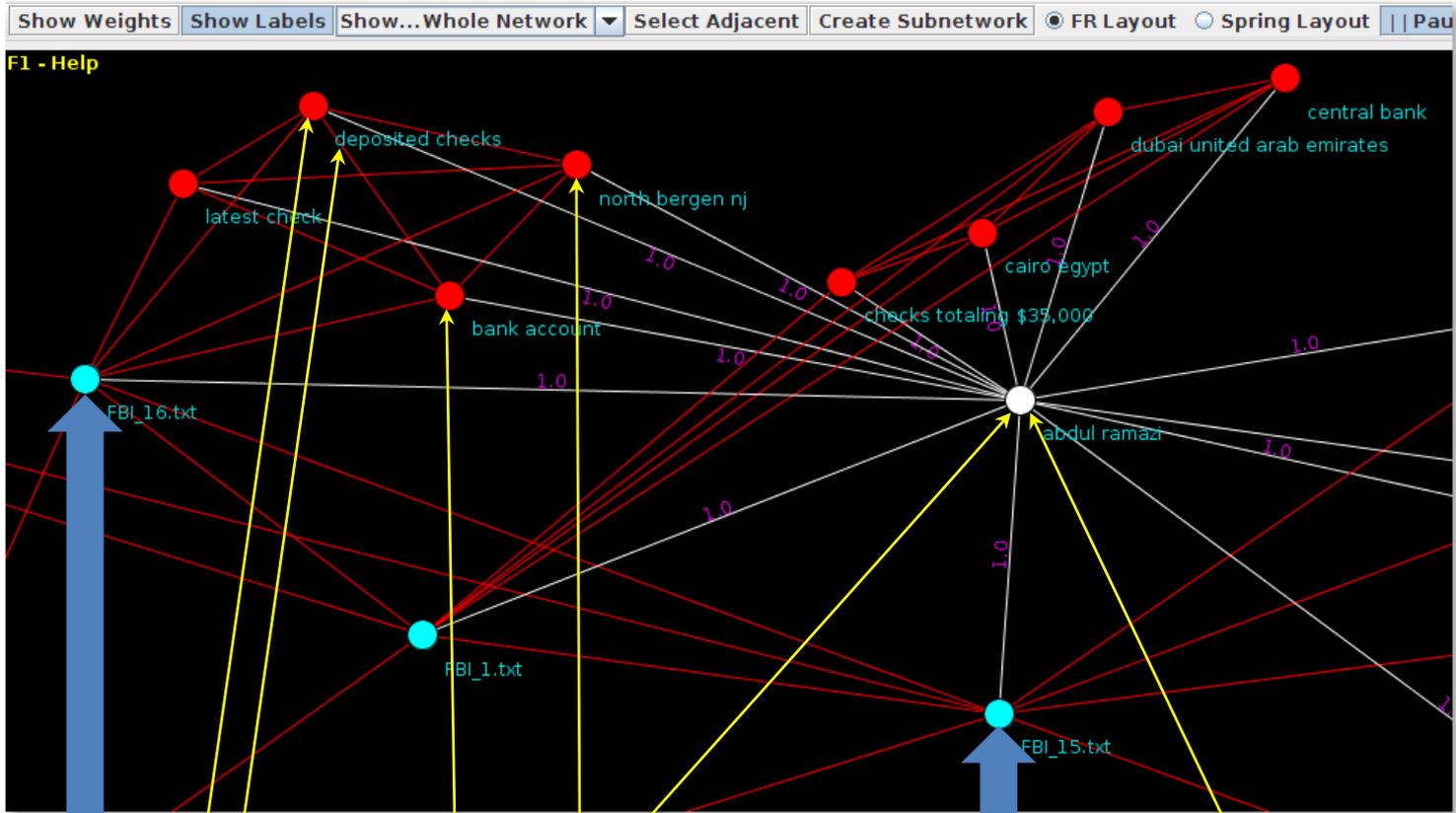
Diverse capabilities



Text analysis capability overview

Capability	Capacity in documents	Search Engines	Natural Language Tools	
Search/Query	100M+	Yes	No	Yes
Unsupervised classification	1M	Some	No	Yes
Supervised classification	1M	No	No	Yes
Clustering (Full document analysis)	100K	No	No	Yes
Term Frequency Analysis (Significant words)	100K	Yes, but not available to user	Yes	Yes
Semantic Extraction (People, places)	1000	No	Yes	Yes

Term network with document links



16) Report Date 22 April 2003. FBI: Hani al Hallak, of **North Bergen, NJ**, has **deposited checks** in his **bank account** that were drawn on First Union Bank account number 1070173749003 in Springfield, VA, in the name **Abdul Ramazi**. The latest check is dated 16 April 2003 and was in the amount of \$8500.

15) Report Date 20 April 2003. FBI: Mukhtar Galab has an account at the Virginia National Bank in Charlottesville, VA. Bank records say he has deposited several checks in the last three months, totaling \$13,000, drawn on account number 1070173749003 held by **Abdul Ramazi** at the First Union Bank in Springfield, VA

Petascale text analysis

- ORNL's Jaguar is the fastest computer in the world
 - 255,000 cores -10PB (13,400 1TB drives) of Storage -362TB of memory
- Google has indexed 1 Trillion unique URLs, but has not analyzed the content of the information
- We are currently developing petascale text analysis techniques to cluster (deep analysis) of 1 trillion documents using Jaguar



Exfiltrate the data from real targets

Hacker's Machines

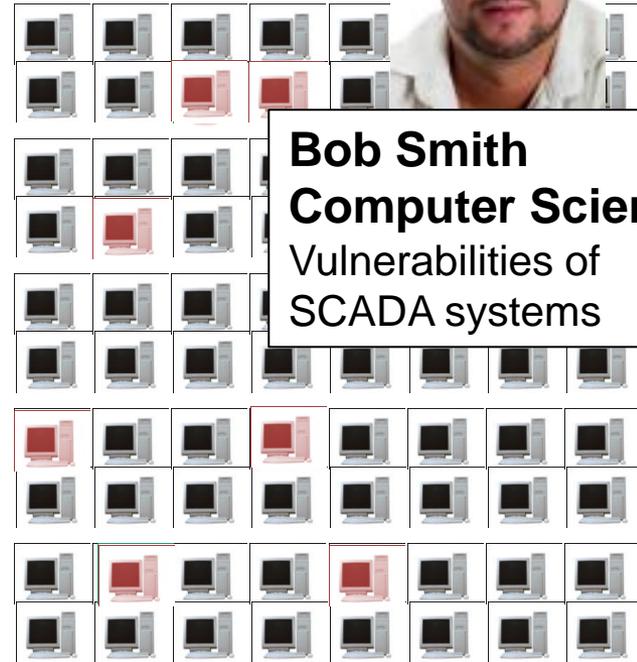


Seattle, WA
Starbucks



Jane Doe
Visitor Control
Names, SSN, DOB,
dates of visits,
clearances,
authorized buildings

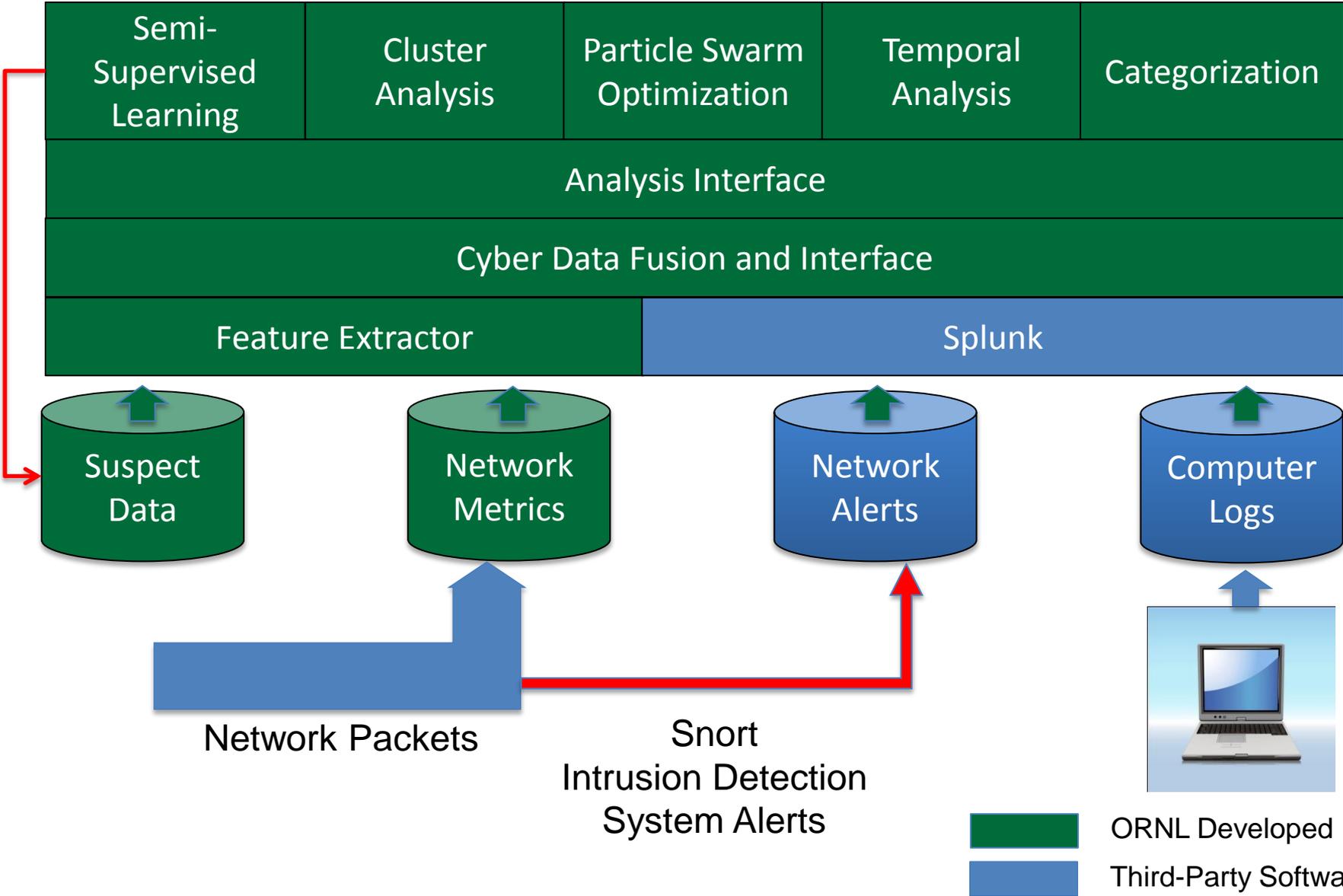
Target Network



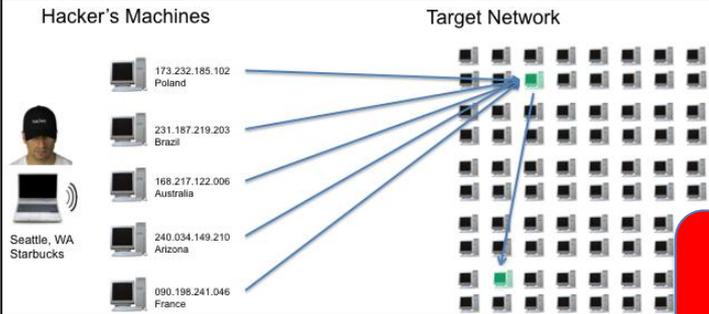
Bob Smith
Computer Scientist
Vulnerabilities of
SCADA systems

- Download all files from Smith and Doe's computers
- Sell Smith's data to highest bidder – Iran or North Korea
- Find open visitor requests for a visitor with a clearance
- Forge a driver license
- Pick up the badge at Visitor Control

New approach



Find attacks on vulnerable computers



The screenshot shows a web browser window displaying a security analysis tool. The browser address bar shows the URL: `http://192.168.56.101:8080/blazeds/CorrelationGUI/CorrelationGUI.html#`. The tool interface includes a navigation menu with 'Correlation Engine' and 'Analysis Engine'. A search section allows filtering by 'Time' (From: 04/06/2010, To: 04/06/2010) and 'Key' (Address, Words). Below the search, there are tabs for 'Listing', 'Cluster Analysis', 'Swarm Analysis', and 'Categorization'. A large circular radar chart visualizes data, with a tooltip for 'Document 348' showing details: Category: R2L Attempt-Success, Source IP: 160.91.238.170, Destination IP: 209.85.133.99, and Top Terms including 'calendar access' and 'web-cgi'. A red callout box with the text 'Alert Text Time Data' points to the radar chart. The status bar at the bottom indicates 'Transferring data from 192.168.56.101...'.

Find exploits

OAK RIDGE CYBER ANALYTICS

Correlation Engine Control
Cyber Analysis Configuration
Cyber Event Analysis

Hacker's Machines

- 173.232.185.102 Poland
- 231.187.219.203 Brazil
- 168.217.122.006 Australia
- 240.034.149.210 Arizona
- 090.198.241.046 France

Seattle, WA Starbucks

Target Network

Bob Smith
Computer Scientist
Vulnerabilities of SCADA systems

Jane Doe
Visitor Control
Names, SSN, DOB, Dates of visits, clearances, authorized buildings

20 mins Alerts Time Range From: Hour: 0 Min: 0 To: Hour: 0 Min: 0 Key IP Address: Search Words Keyword: << >>

Event Listing Cluster Analysis **Swarm Analysis** Categorization

Model: IP Address Historical
Iterations: [Slider]
Similarity: 35
Iteration: 130
Run Continue

Five external machines behaving suspiciously

Current node: 152
x: 134
y: 416
Src IP: 86.164.180.1
Src Port: Source_port N/A
Dest IP: Dest_ip N/A
Dest Port: Dest_port N/A
Time: time N/A
Behavior Code: 1
Behavior: 86.164.180.1 [Behavior unknown and alarm action is: (spp_frag3) Fragmentation overlap]

Third Party Proprietary

Current capabilities

- **TRL 7+ in 18 month**
 - Live Sprint and AT&T feeds
 - 3+ terabytes of network traffic daily
- **Extraction of network features – Alerts plus network traffic metrics**
- **Machine learning – Incorporate new insights into analysis**
- **Cluster analysis – Group similar alerts**
- **Particle swarm optimization – Group similar computer behavior**
- **Temporal analysis – View time vs. IP Addresses vs. alert type**
- **Categorization – Group alerts by defined category**

Child predators



Somer
Thompson
2002 - 2009



- This is Somer. She was born April 5th 2002 in Columbia, South Carolina.
- She enjoyed swimming and dancing. She often thought of herself as becoming a ballerina.
- On October 19th 2009 at 2:45pm Somer left Grove Park Elementary school to meet up with her brother and sister for their daily mile walk home.
- She ran ahead of her siblings and got separated from the group
- This was the very last time Somer Thompson was seen alive.
- Her body was found in a Georgia landfill 2 days later. At that time, there were no suspects.

Child predators (continued)



Jarred Harrell

He abducted Somer that day.
He was charged with raping
and murdering her.

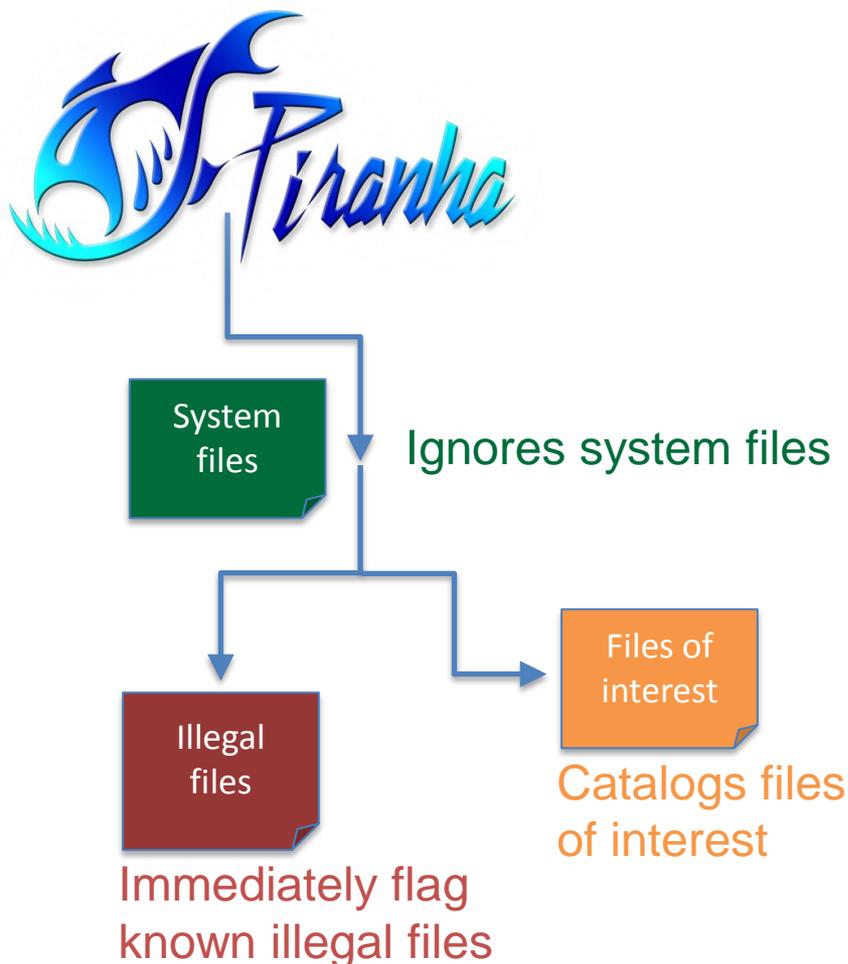
“...Police had possession of Harrell's computer and camera two months before Thompson disappeared, ...'they found Harrell's camera containing a video with horrifying images of Harrell molesting a three year-old child',”
reports ABC. - Saturday, 17 April 2010 10:54 ABC News' 20/20

Fact



If Jarred Harrell's computer and camera had been forensically analyzed in a timely manner, Somer would still be alive.

Sampling



- **Sample suspect's disk to find a predator immediately**
- **Assuming illegal files are grouped together becomes a search problem**
- **Full analysis can be completed later**

Artemis – 30 child pornography arrests in Knoxville, TN

- Working with the *Protect and Internet Crimes Against Children (ICAC) Task Force* in Knoxville, TN
- A prototype thumb drive system to rapidly analyze information on a hard drive and identify illegal material
- When booted loads Knoppix and Artemis to allows the officer to scan specific directories or the entire disk for images and documents.
- Image analysis looks for high percentage of skin tones, faces, or an illegal hash values
- Text analysis looks for documents that are similar to a provided set.

Scanner agent results for /home/v8q/Projects/ICAC Test Files

Status is **FINISHED** at Thu Feb 04 13:45:28 EST
 Started at **Thu Feb 04 13:36:50 EST 2010**

3 Word doc documents scanned.
 1076 Plain text documents scanned.
 1073 Image/Video documents scanned.
 2 PowerPoint Presentation documents scanned.
 24 HTML documents scanned.

1 documents categorized as **SHA - bad hash**.
 0 documents categorized as **CP Search**.
 1 documents categorized as **Bad Story 3**.
 23 documents categorized as **Skin Detected**.
 1 documents categorized as **Bad Story 2**.
 1 documents categorized as **Bad Story 1**.
 219 documents categorized as **Face Detected**.
 1 documents categorized as **Bad MS Word**.
 0 documents categorized as **Incest**.

Examiner's Name: TEP
 Owner's Name: TEP
 OS: Linux
 OS Architecture: amd64
 OS Version: 2.6.31-17-generic
 User's account: v8q

Comments:

Office

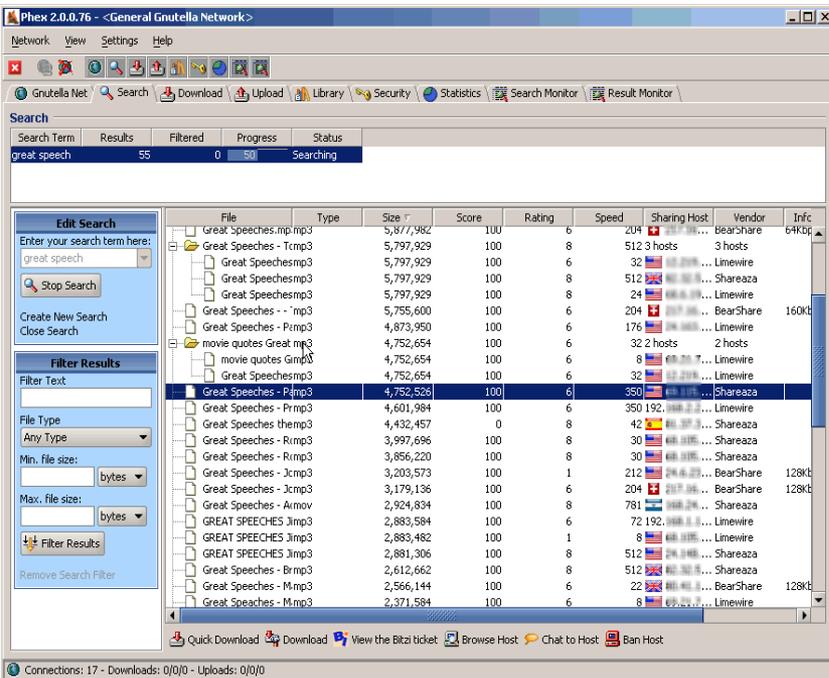
Officer Identified Image Count: 0

ID	Chk	Access File	Skin	Face detected	File Name	Type	Hash Value	Size	Category/Similarity
000		image07.doc	0	No	image07.doc	doc	5F796e4736a89b747676c37f6e814823e3c5819	154112	Category-Bad MS Word - (Score: 97%)
D			52	Yes	IMG_01811.JPG	JPG	328a6e8186ac4834834997184641458326278a	989859	N/A
E			63	Yes	IMG_02001.JPG	JPG	627094200475196279750af7481465120354627	288214	N/A
G			53	Yes	IMG_01761.JPG	JPG	85a6680c5380a429a42944812481458326278a	952054	N/A
I			11	Yes	100_0206.JPG	JPG	38a966517031a41424848955663096419305af	472832	N/A
J			10	Yes	100_0303.JPG	JPG	47132089305a610324845c754a4b4a6b499c	482998	N/A
005		0000000000.jpg	0	No	0000000000.jpg	jpg	323537461546128414409a2776350a90563a3	9021	Category-Bad Story 3 - (Score: 99%)
005		0000000000.jpg	0	No	0000000000.jpg	jpg	323537461546128414409a2776350a90563a3	9021	Category-Bad Story 2 - (Score: 99%)
007		0000000000.jpg	0	No	0000000000.jpg	jpg	132671036666666666666666666666666666	939	Category-Bad Story 1 - (Score: 99%)

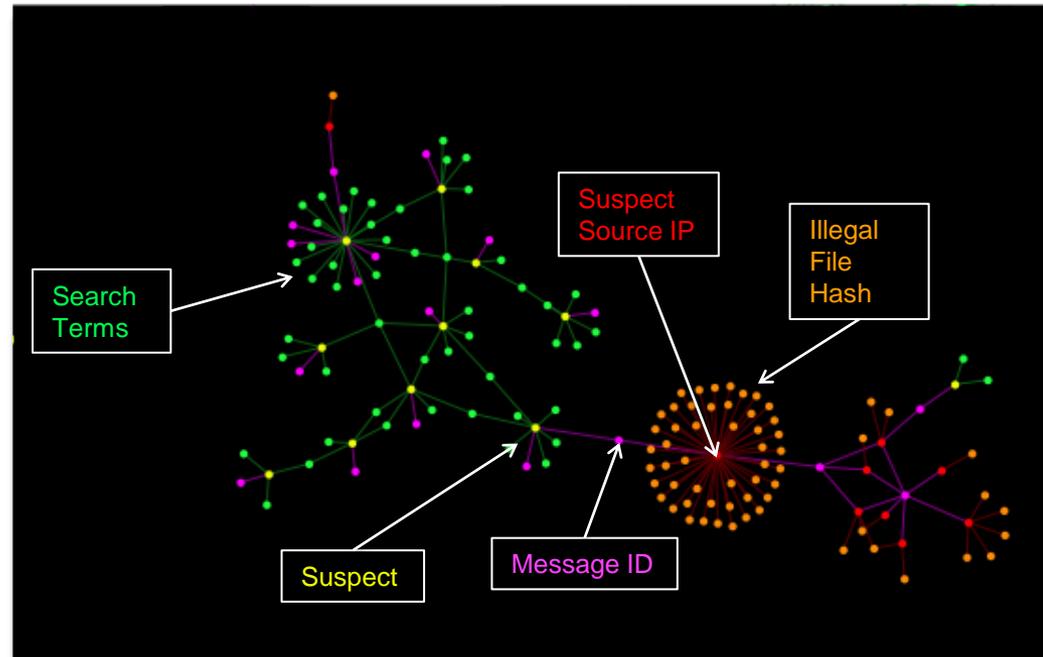
ASLAN – Graph Analysis of Large file sharing network

- You can make part of your computer public, so that others can download your files
- The software keeps track of the IP addresses of the computers that download the files
- Social networks can be built to see who is sharing from whom

File Sharing Software

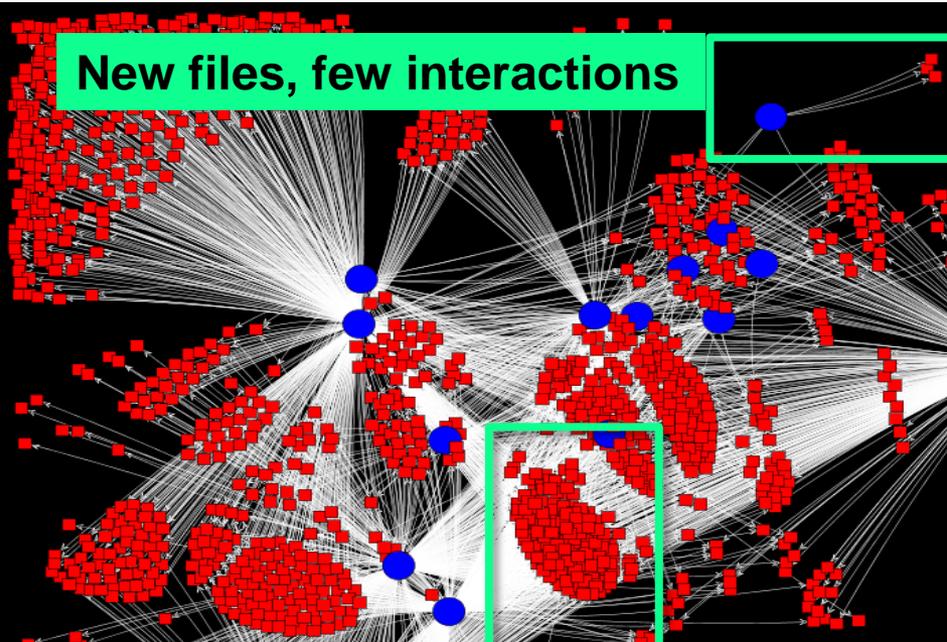


Social Network Graph

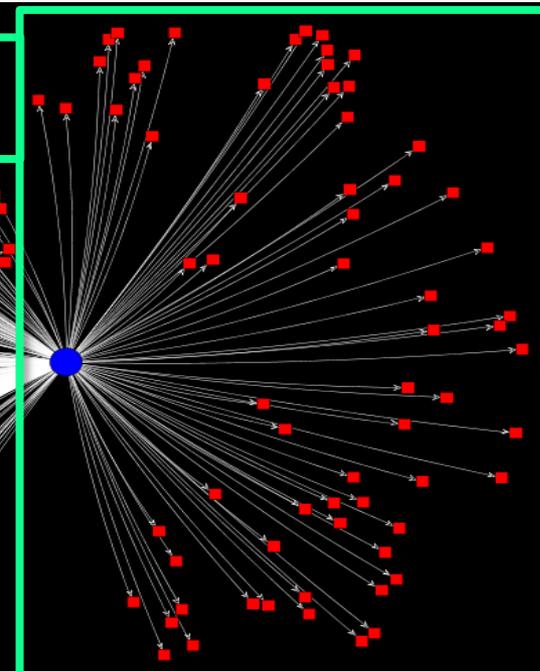
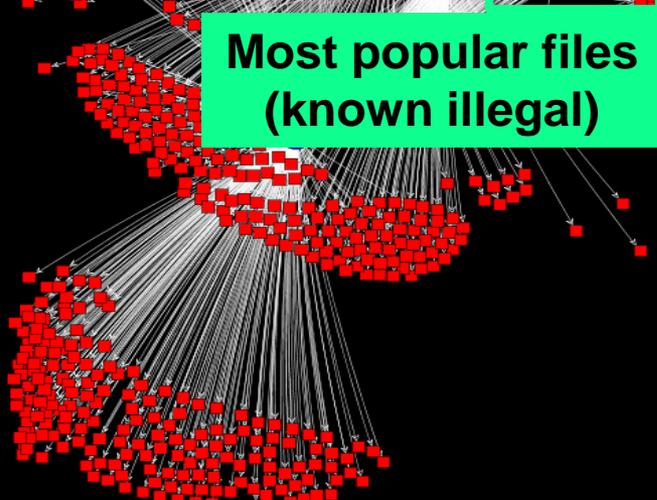


Use supercomputers to find the worst of the worst

New files, few interactions



Most popular files
(known illegal)



● Users
■ Files

Producing files
or getting them
from a hidden
network



Summary

- **Difficult national issues**
- **The challenge is in analyzing large volumes of unstructured data**
- **We are proven experts in analyzing large volumes of unstructured data**
- **Piranha and ORCA provides new ways of analyzing and correlating data**
- **This capability is applicable to any intelligence gathering and analyzing activity**

Contact

Thomas E. Potok, Ph.D

Group Leader

Applied Software Engineering Research

potokte@ornl.gov