SAND2010-4194P

Detecting Anomalies and Patterns in Complex Data

CFTC Presentation
June 17, 2010

David G. Robinson, PhD
drobin@sandia.gov
Computer Science and Informatics
Sandia National Laboratories, NM

Outline

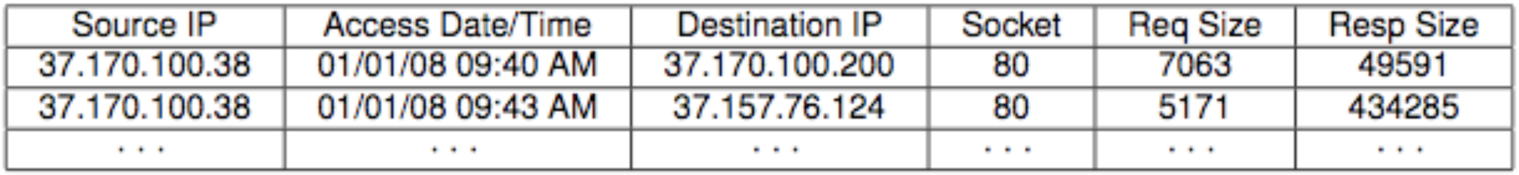
- Introduction
- Anomaly detection (example w/depth)
- Examples of other applicable technologies
- Questions

Cyber Insider Detection

We were presented with the problem of identifying an insider threat in a cyber environment. Header information from 115,414 network events over the course of 1 month were provided.

Analysis goal: Identify unusual patterns in network transactions along with the associated source IP and destination IP addresses

Raw Data …

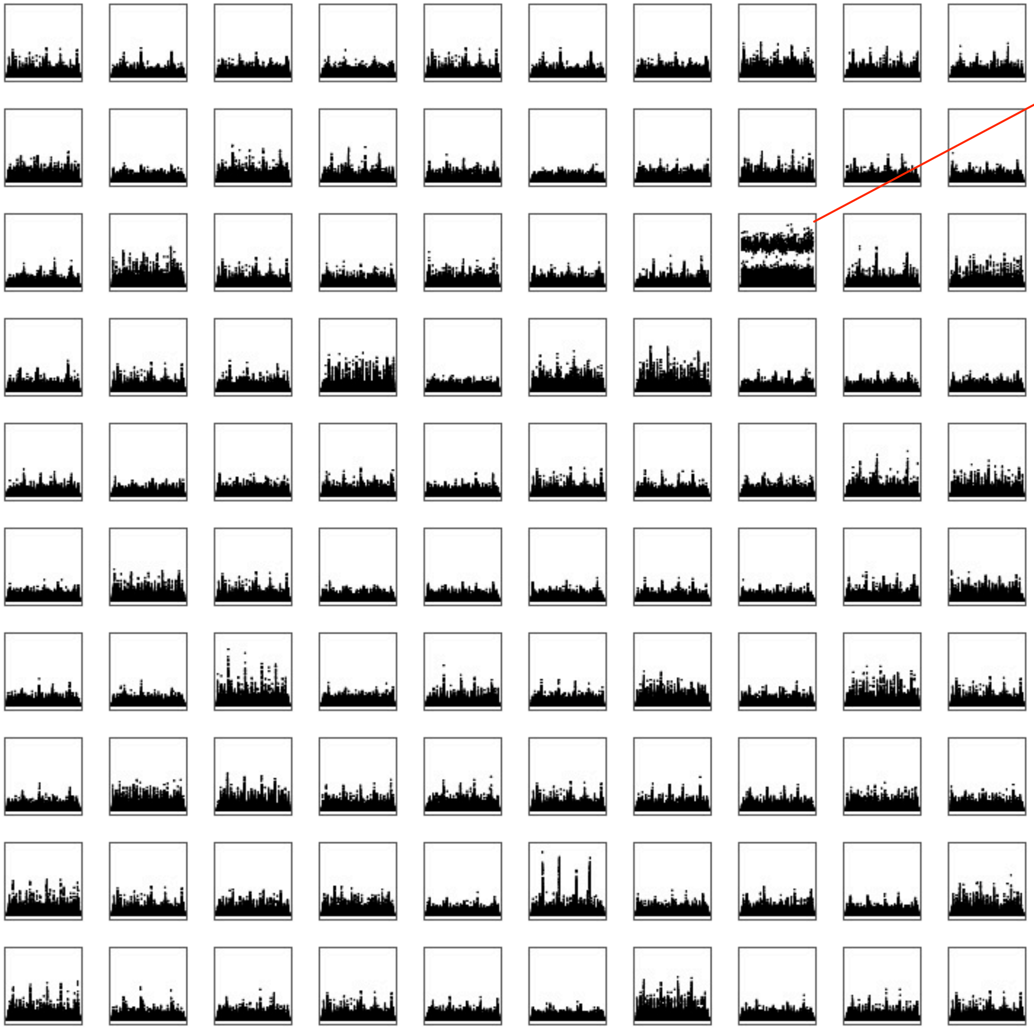


Source IP	Access Date/Time	Destination IP	Socket	Req Size	Resp Size
37.170.100.38	01/01/08 09:40 AM	37.170.100.200	80	7063	49591
37.170.100.38	01/01/08 09:43 AM	37.157.76.124	80	5171	434285
. . .	. . .	. . .	. . .	. . .	. . .

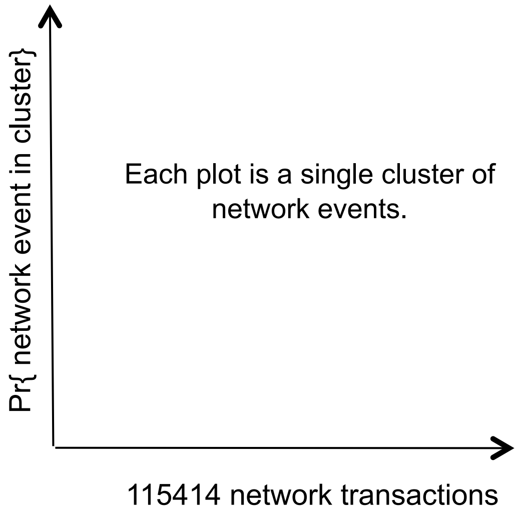
First step was the use of natural language methods to statistically cluster all 115k network transactions.

Second, two analysis methods were investigated to find specific anomalies within the network transactions: manual and automated.

Clustering of Network Events



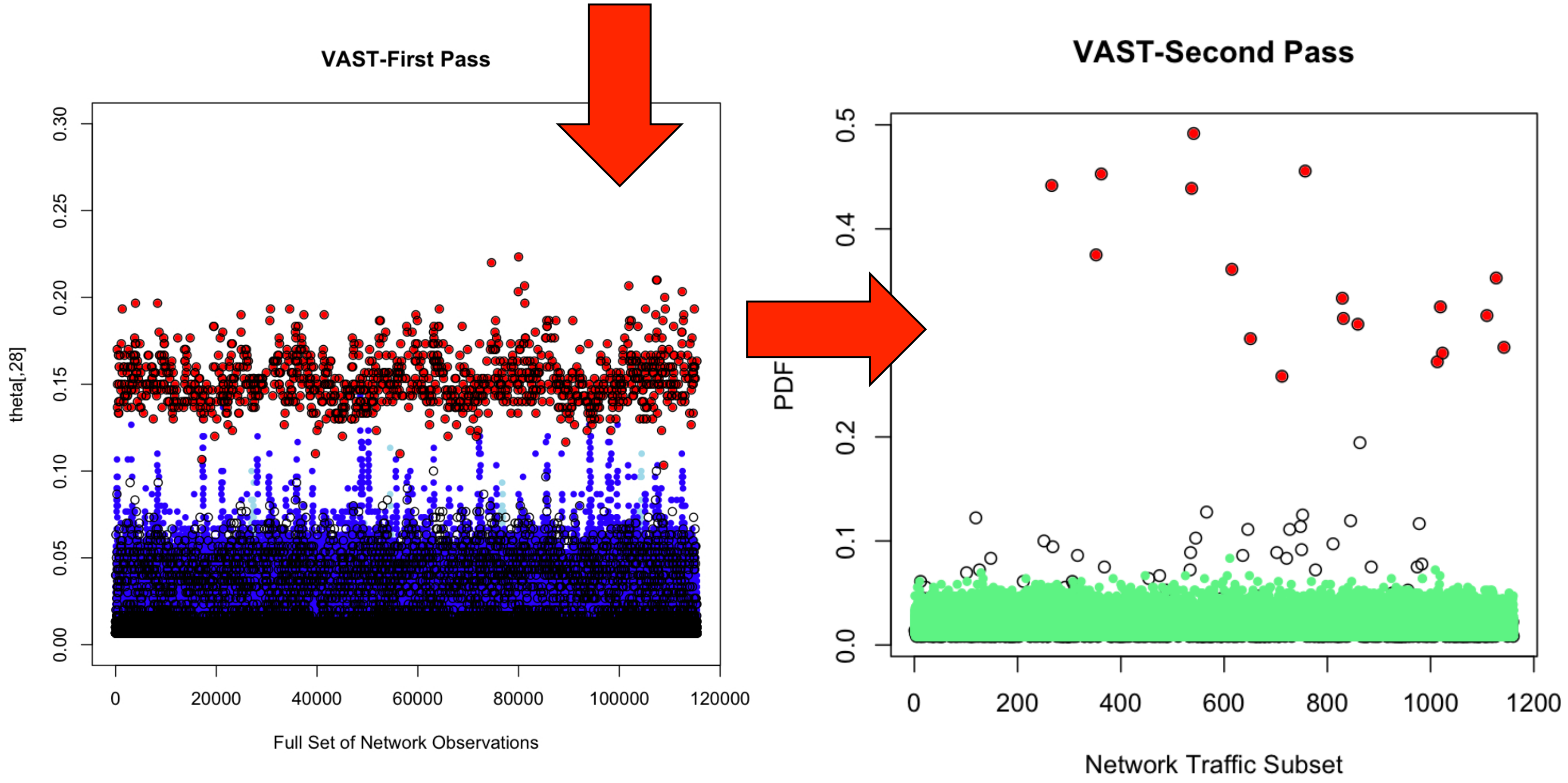
Note how this cluster is different than the others.



Each plot is a single cluster of network events.

Results from Manual Analysis

Cluster 28 selected from original 100 clusters for investigation.



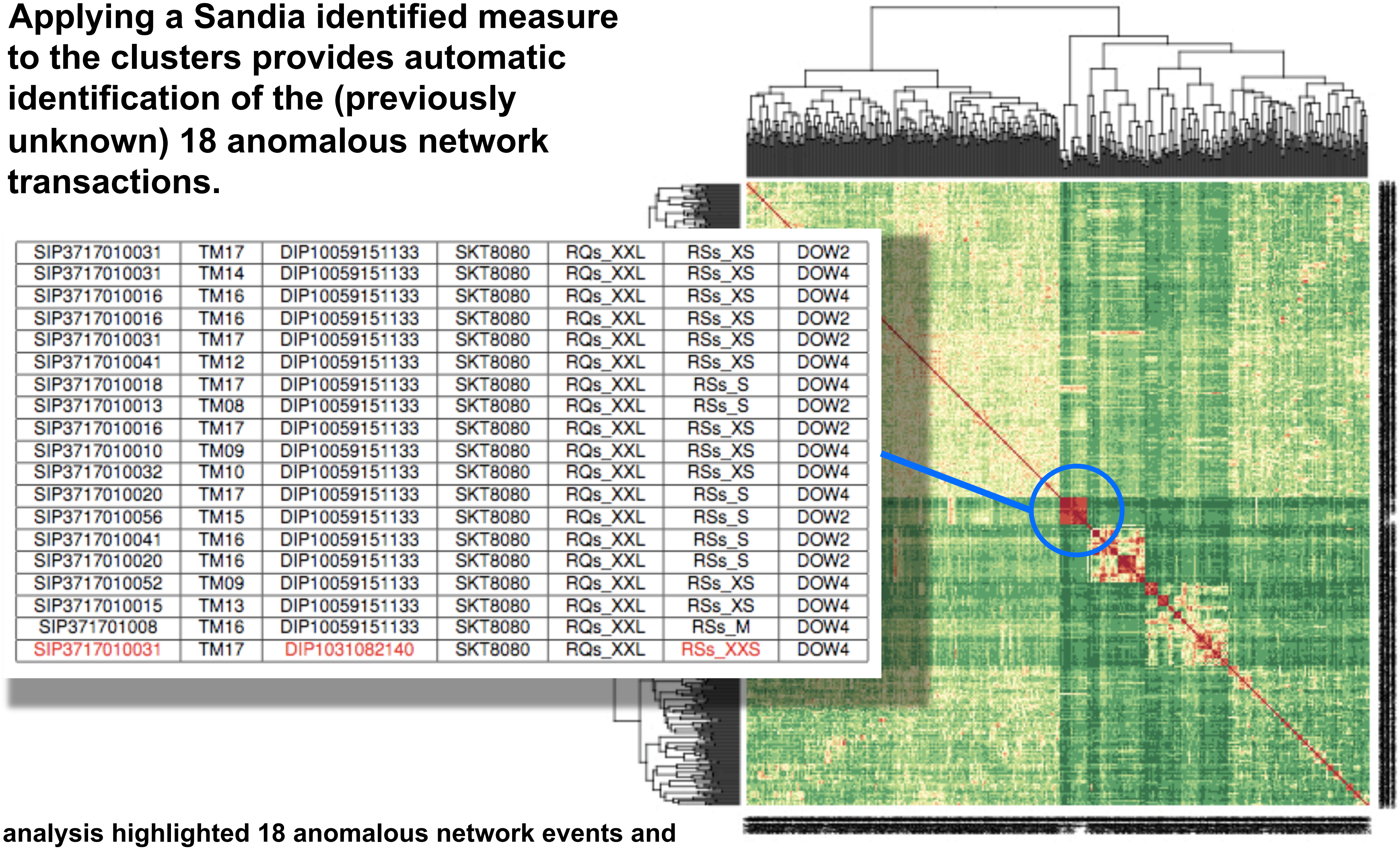
Phase 1: Red dots are individual traffic events down-selected from cluster 28 for additional analysis

Phase 2: Traffic selected from as anomalies for attention by cyber security.

Results from Automatic Anomaly Detection

Still working HPC implementation

- **Applying a Sandia identified measure to the clusters provides automatic identification of the (previously unknown) 18 anomalous network transactions.**

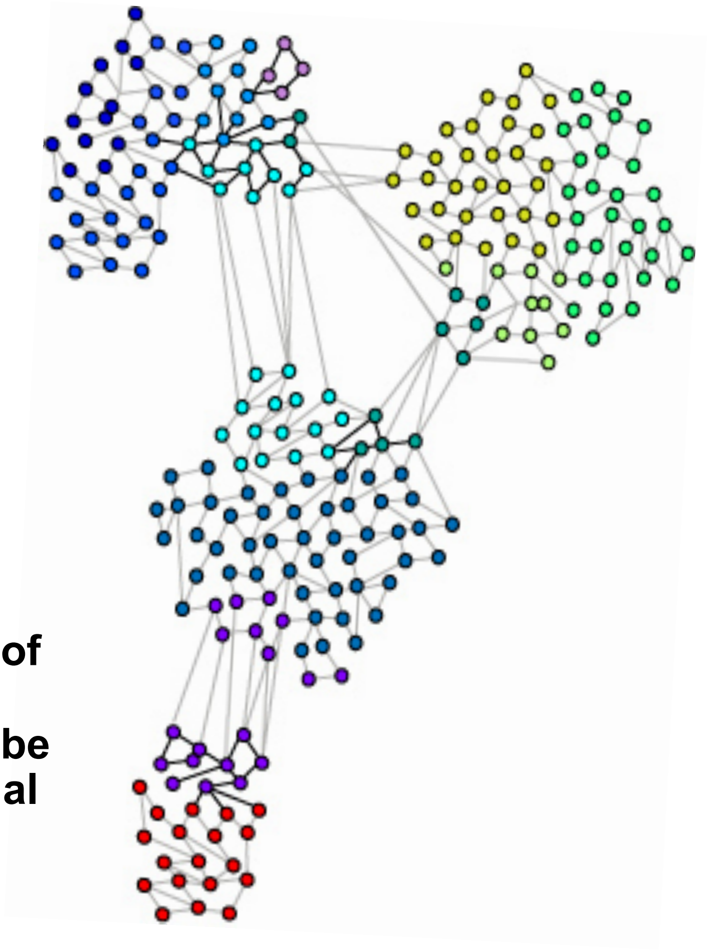


SIP3717010031	TM17	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW2
SIP3717010031	TM14	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW4
SIP3717010016	TM16	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW4
SIP3717010016	TM16	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW2
SIP3717010031	TM17	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW2
SIP3717010041	TM12	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW4
SIP3717010018	TM17	DIP10059151133	SKT8080	RQs_XXL	RSs_S	DOW4
SIP3717010013	TM08	DIP10059151133	SKT8080	RQs_XXL	RSs_S	DOW2
SIP3717010016	TM17	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW2
SIP3717010010	TM09	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW4
SIP3717010032	TM10	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW4
SIP3717010020	TM17	DIP10059151133	SKT8080	RQs_XXL	RSs_S	DOW4
SIP3717010056	TM15	DIP10059151133	SKT8080	RQs_XXL	RSs_S	DOW2
SIP3717010041	TM16	DIP10059151133	SKT8080	RQs_XXL	RSs_S	DOW2
SIP3717010020	TM16	DIP10059151133	SKT8080	RQs_XXL	RSs_S	DOW2
SIP3717010052	TM09	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW4
SIP3717010015	TM13	DIP10059151133	SKT8080	RQs_XXL	RSs_XS	DOW4
SIP371701008	TM16	DIP10059151133	SKT8080	RQs_XXL	RSs_M	DOW4
SIP3717010031	TM17	DIP1031082140	SKT8080	RQs_XXL	RSs_XXS	DOW4

The analysis highlighted 18 anomalous network events and these results match the IEEE VAST 2009 published 'truth'.

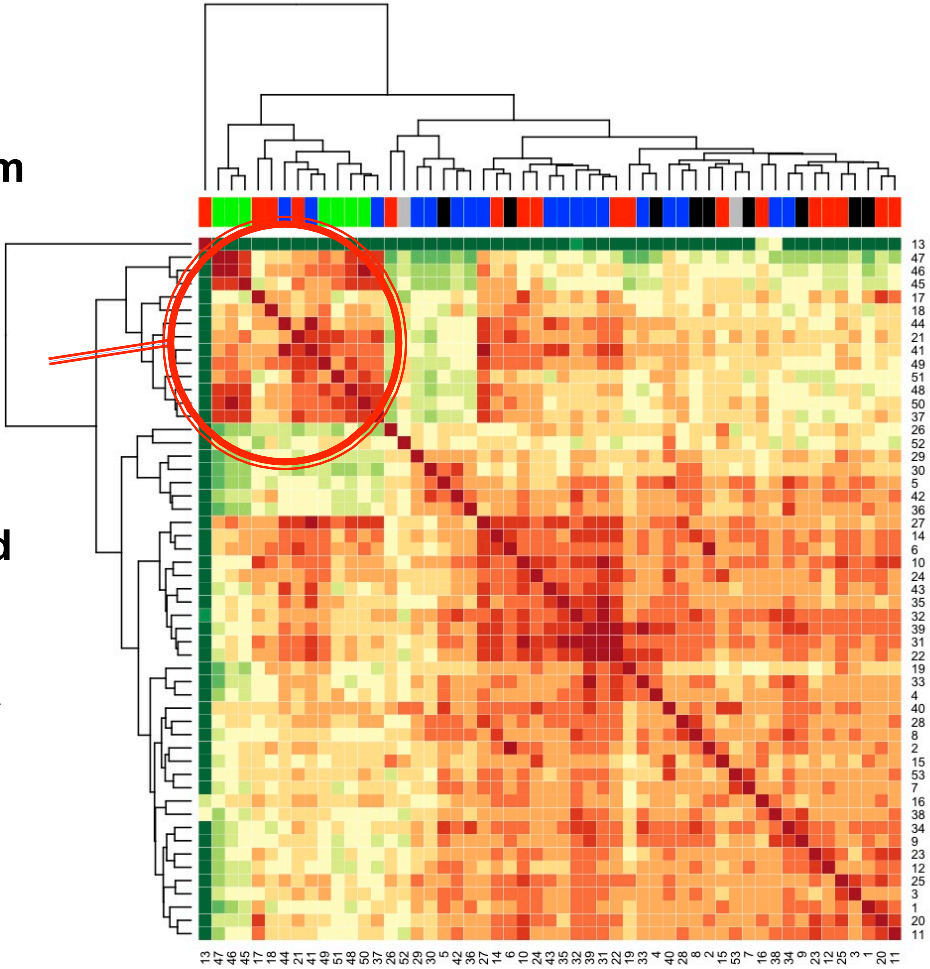
Unusual Information Communities

- Previous examples focused on discovering patterns in transaction data. Might also be interested in unusual and/or evolving patterns or communities in communication traffic.
- Every node is a possible sender or receiver and the edge contains the information exchanged.
- Applications for cyber security and intelligence community oriented toward identifying:
  - relationships that weren't known to exist
  - people who weren't known to be involved
  - unusual patterns in network structure both in/out, e.g. (spear) phishing / unusual information exchange
- In the financial community, traders and 'tippers' relationships must be discovered, their possession of insider information must be characterized, the relationship between the subjects and targets must be understood. In addition, the existence of any financial relationships must be established.
- Relationships are characterized using a risk-based approach: What is the probability that person X exchanged information Y with person Z?



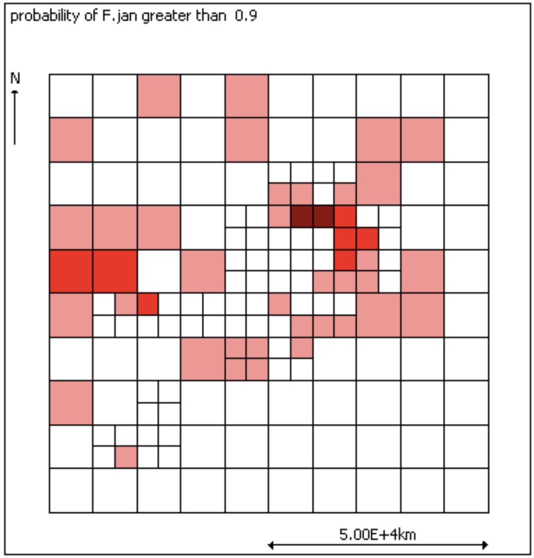
Coupled Event and Topic Analysis

- Couple previous structured network transaction data with unstructured text from reports, etc.
- Once a suspected anomaly has been identified then you will want to match unusual trading patterns with related data, e.g. merger announcements, to enhance confidence in anomaly.
- Goal: Identify critical topic discussions and relate those topics to network events in order to…
- Identify and support prosecution of insider trading
  - trading spikes and market surveillance are generally insufficient evidence. Formal data analysis of ‘who knew what’ can be used to identify ‘tipping’.



Other Active Pattern Detection Efforts

Sandia is currently using disease dynamics to model spatial and temporal patterns of Blue team and Red team interactions Afghanistan. Coupled with external information about weather patterns, tribal and religious, agricultural distribution, etc. Sandia can predict an embedded IED event with 80-90% accuracy. We also provide long and short term predictions of future risk of an event for specific dates and locations.



http://www

http://www.army.mil/-images/2009/07/09/44434/index.html

Instead of spatial correlations of interest might include unusual temporal correlations between/across commodity futures, e.g. weather as it impacts December oil versus January oil

Questions ?

CFTC Presentation

June 17, 2010

David G. Robinson, PhD
drobin@sandia.gov
Computer Science and Informatics
Sandia National Laboratories, NM

Backup Slides

Personal Background

- **Area of research: Predictive Analytics**
  - using statistics, game theory, data mining to find trends or patterns in data to characterize current events and predict future events.
    - Patterns in relationships between events
    - New/evolving patterns
    - Clustering
    - Forecasting
  - focus is on supporting decision making in the presence of uncertainty.

- **Current customers/projects**
  - NASA/JPL: predicting launch anomalies and subsequent health effects of high consequence launches
  - JIEDDO: predicting where/when of embedded IED events in AFG
  - Internal research (Networks Grand Challenge)
    - Anomaly detection
    - Sentiment analysis
    - Community detection

Critical Points to Take Away

- Contrary to popular opinion, you don’t have to fully understand what normal network behavior is to identify abnormal behavior. You just need to be able to identify ‘less normal’ behavior.

- To identify anomalous events, current security methods depend heavily on:
  - Long list of rules that is becoming longer and which can conflict
  - Years of experience to identify suspicious network events
  - Rules only apply to things you know have happened; new approach allows for discovery of unknowns.

- Professional cyber analysts took about 90 minutes to narrow the VAST data down to 80 events. The new methodology took 5 minutes and found the true 18 events (in complete ignorance).

- Methods can be implemented in a real-time monitoring fashion so suspect events can be identified early, with quantifiable risk associated with decision making.

- Algorithms are currently being tuned for implementation.

Natural Language Analysis of Events

Each network event is treated as a document and each characteristic of an event is treated as a word in the document.

No significant change detected over time – focus on static case

- **Investigated complete set of observations of network traffic**
  - **number of events: C=115414**
  - **number of words: N=20328**
  - **number of topics (clusters): T=100**

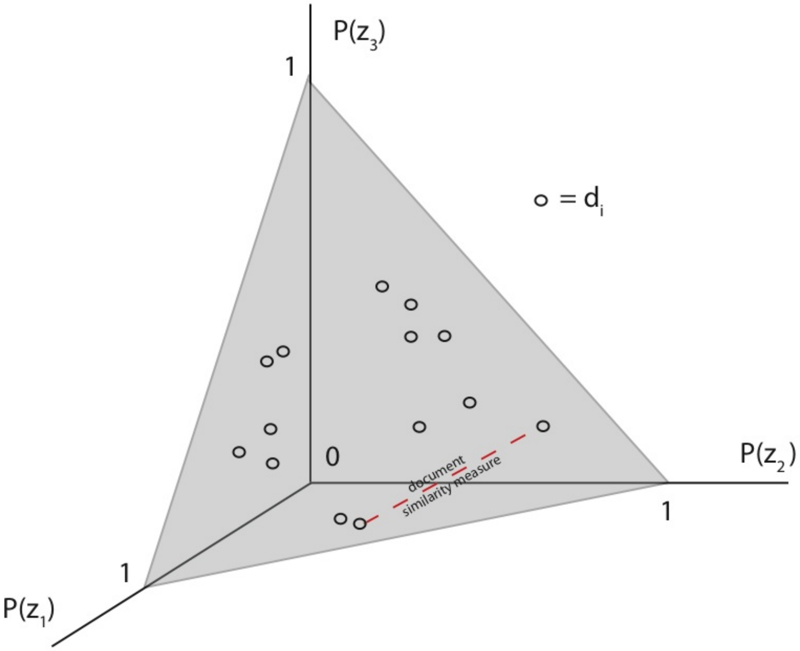
- **Latent Dirichlet Allocation used to build 'soft' clusters (topics) of network traffic characterized by:**
  - **Θ: CxT matrix where each row is conditional pdf of topics in document *i* :**
    - **θ_{ij} = Pr{topic z_j | document i}**
    - **Σ_j Pr{topic z_j | document i} = 1**

Two methods were explored to find anomalies within the patterns: manual and automated

Automated Approach

- Information based similarity measures used to automatically identify anomalous clusters of network events
- Investigated a number of measures to find the most discriminating.
- Hellinger distance is used to characterize the similarity between the mixture of topics in documents *i* and *j*.

$$D_{ij} = \sum_{k=1}^{T} \left(\sqrt{\theta_{ik}} - \sqrt{\theta_{jk}} \right)^2 / T$$

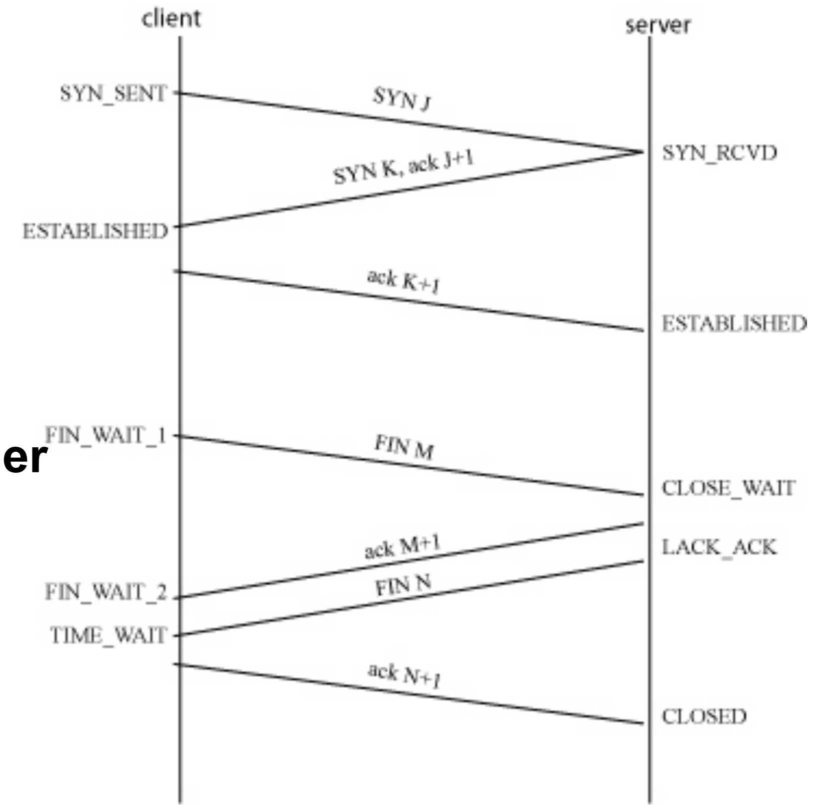


Cyber Security: Patterns in TCP/IP Traffic

- Looking for information leaks both intentional (malware) or unintentional (commercial software bug).

- Can we detect unusual client-server network TCP/IP traffic?

- Identify unusual traffic events for further investigation by cyber security.


client
server
SYN_SENT
SYN J
SYN_RCVD
SYN K, ack J+1
ESTABLISHED
ack K+1
ESTABLISHED
FIN_WAIT_1
FIN M
CLOSE_WAIT
ack M+1
LACK_ACK
FIN_WAIT_2
FIN N
TIME_WAIT
ack N+1
CLOSED