Network Discovery, Characterization, and Prediction

LDRD
LABORATORY DIRECTED RESEARCH & DEVELOPMENT

Sandia National Laboratories
Philip Kegelmeyer

Problem

Interactive Informatics on Massive Data

Networks engaged in weapons proliferation, terrorism, cyber attacks, clandestine resale of dual-use imports, arms and drug smuggling, and other illicit activities are major threats to national security. These adversarial networks, in turn, rely on legitimate and illegitimate secondary networks for financial, supply chain, communication, recruiting, and fund-raising activities. Complexity, dynamism, resilience and adaptability make adversarial networks extremely difficult to identify and disrupt. Often the only way an adversary may be detected is through the networks they use. In short, our real adversaries are networks.

The **Discovery** of adversarial networks is immensely difficult, as a network may only reveal itself by the union of its parts. Relevant data comes from many sources and is geographically and temporally dispersed. Thus, very large and heterogeneous data collections must be analyzed collectively to detect networks.

The **Characterization** of networks requires methods for identifying hidden properties and relationships, and for analyzing the structure of a network to learn about its purpose and the roles of its components.

Structure also suggests likely evolution and intent, allowing **Prediction** of the future shapes of the network.

In this project, we rigorously elicit the needs of the analyst community intent on defending our critical infrastructures, do basic research on uncertainty, research and evaluate novel network analysis algorithms, and implement that research to address those needs to create a flexible, interactive capability for interactive analysis of large datasets. The project team includes research mathematicians, developers, experts in user elicitation, and end-users, and so has all the needed talent to span the full LDRD spectrum from Discover through Create to Prove.

Approach

Analyst Focused, Prototype Tested, Interactive Graph Analysis

A Human Factors team (led by a cultural anthropologist) to ensure that all R&D addresses real analysis needs.

Research Directions:

- Scalable attributed relational graph (ARG) algorithms: find communities, connections, trends
- Linear and multi-linear algebra methods, because ARGs are naturally manipulated as tensors.
- Statistics and uncertainty: find and quantify anomalies, model and communicate uncertainty
- Advanced visualization of complex data and its uncertainty.
- Prediction: model graph evolution, assess predictability, make predictions when possible, quantify their reliability.
- Integration: enable usability and interactivity at scale through optimal architecture use and parallel algorithm implementation.

Prototypes: yearly, integrate capabilities into a single tool, apply to real analyst needs, and evaluate.

Prediction

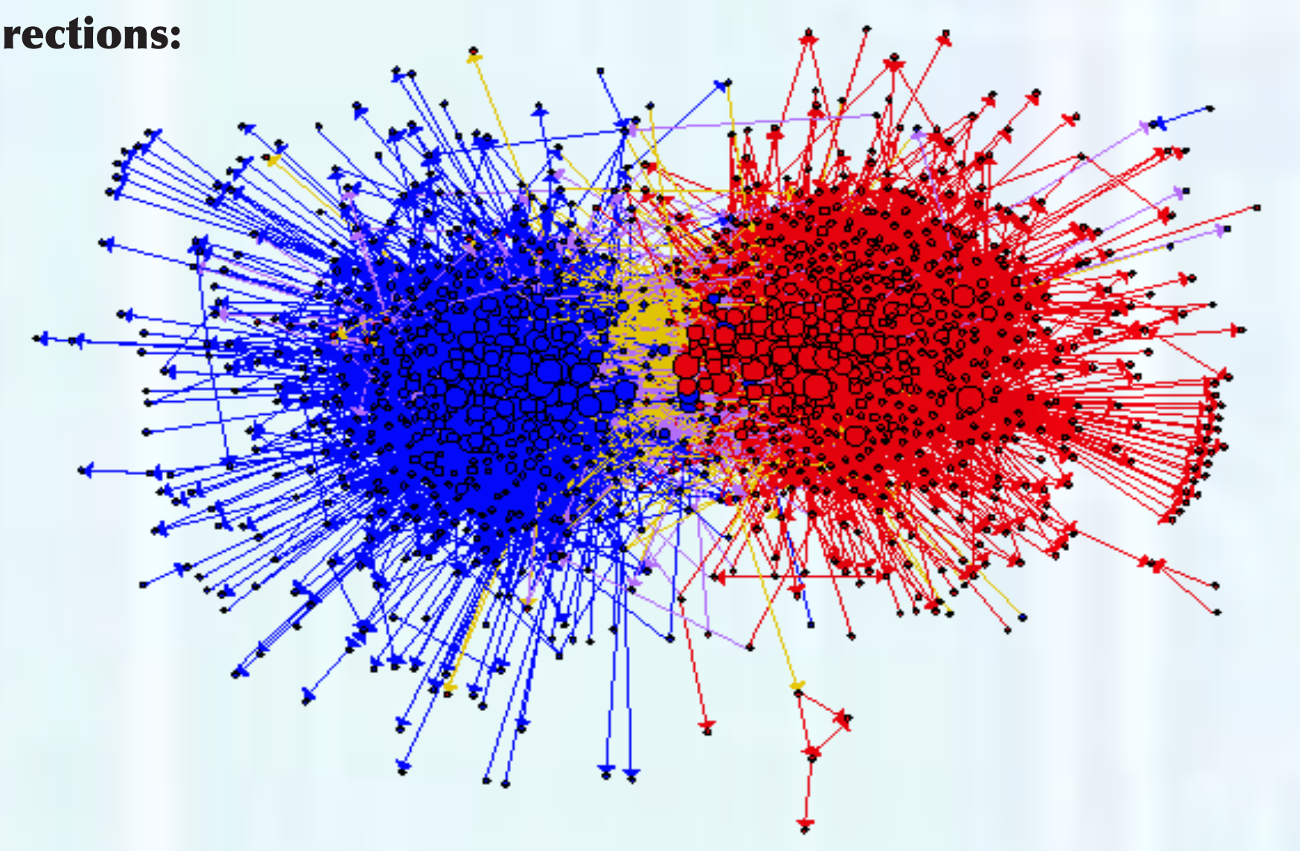
One example, out of many, of our research directions:
Discover network structure which can be characterized and exploited for prediction or to show futility of prediction.

Illustrative examples

- Diffusion on social networks.
- Adversaries on coevolving networks.

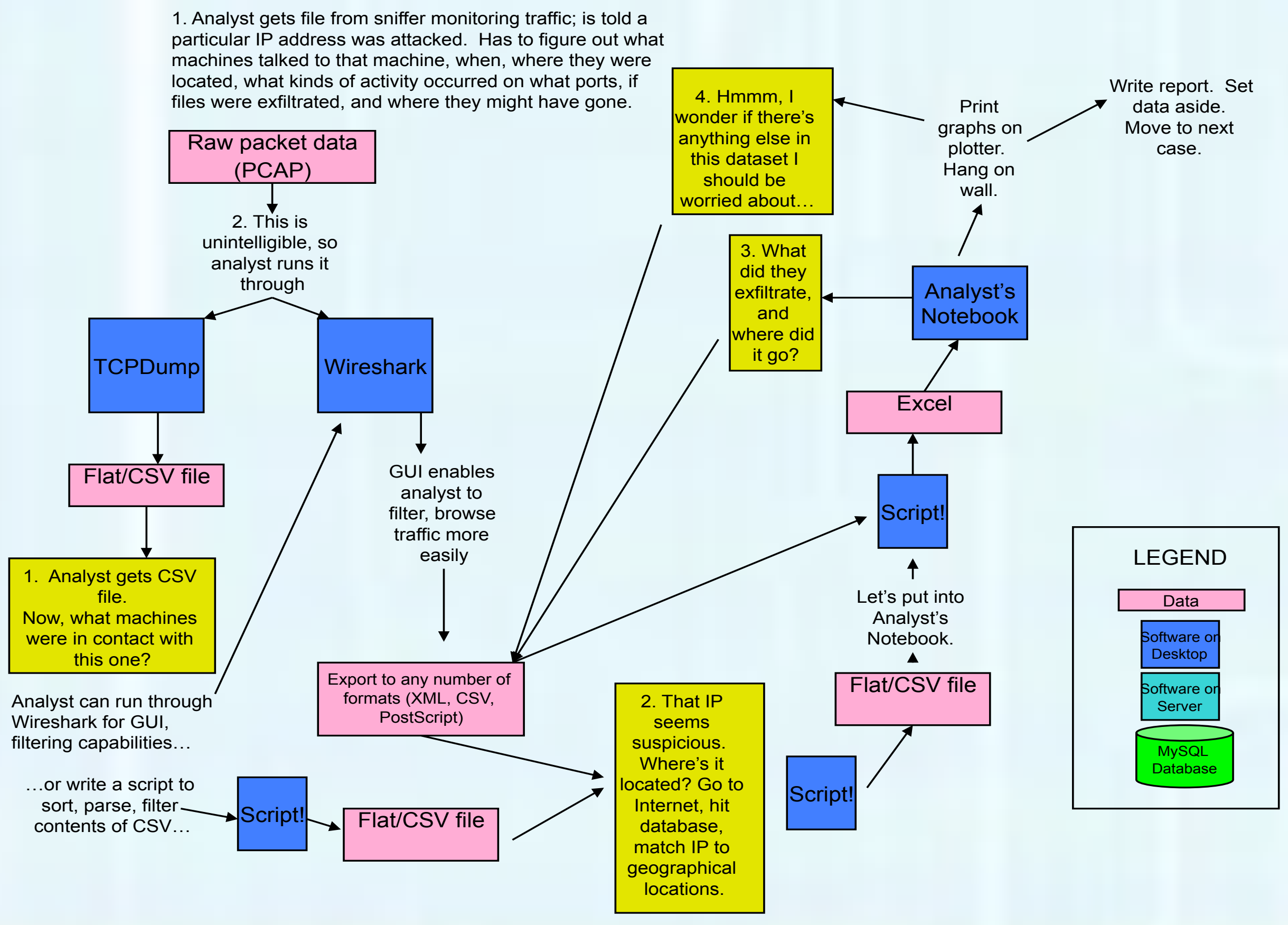
Some key elements

- Predictability before prediction.
- Scalable analytics (e.g., assessment without simulation).
- Uncertainty quantification via "robust yet fragile" framework.



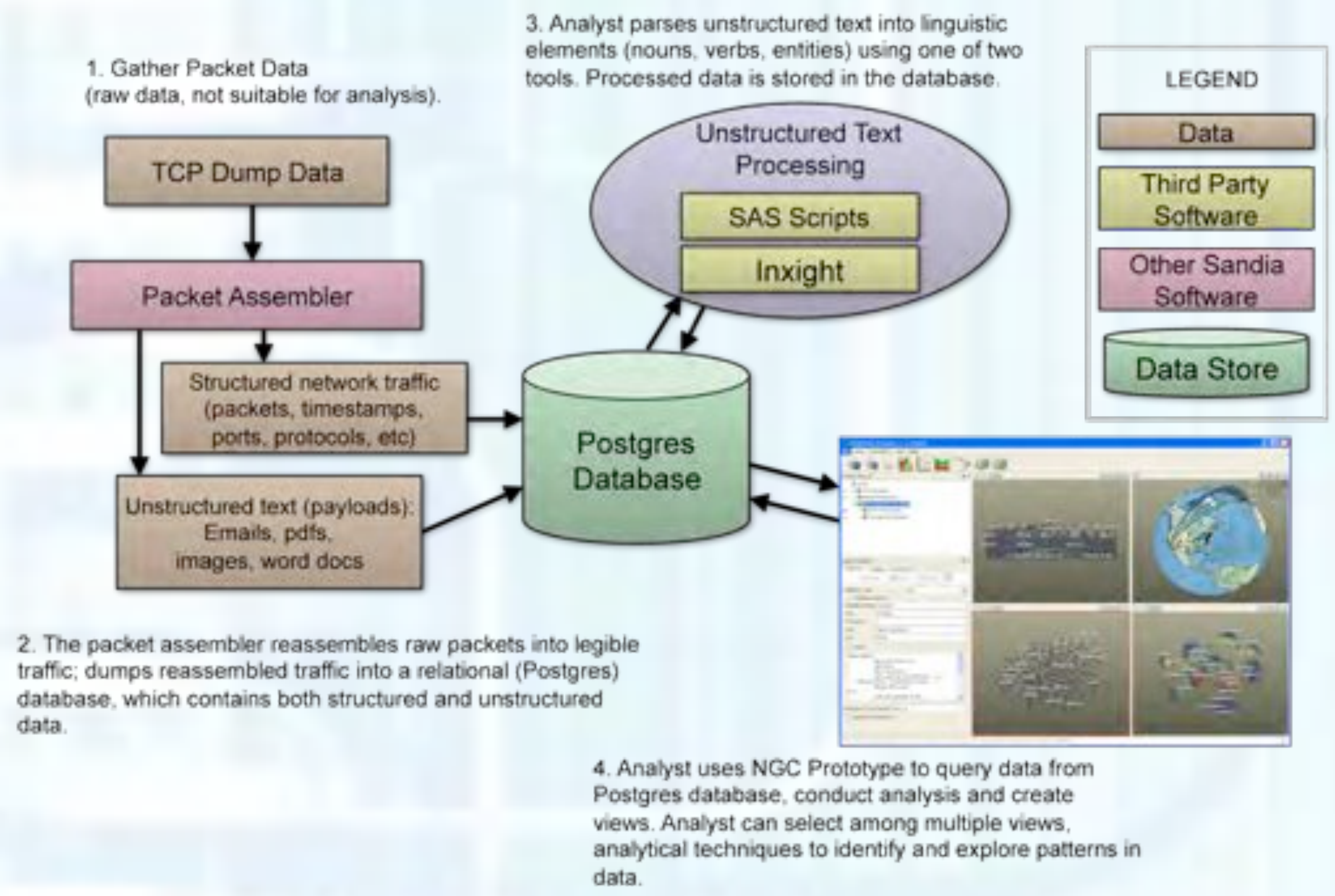
Results

The Way It Was ...


1. Analyst gets file from sniffer monitoring traffic; is told a particular IP address was attacked. Has to figure out what machines talked to that machine, when, where they were located, what kinds of activity occurred on what ports, if files were exfiltrated, and where they might have gone.
Raw packet data (PCAP)
2. This is unintelligible, so analyst runs it through
TCPDump
Wireshark
Flat/CSV file
1. Analyst gets CSV file. Now, what machines were in contact with this one?
Analyst can run through Wireshark for GUI, filtering capabilities…
…or write a script to sort, parse, filter contents of CSV…
Script!
Flat/CSV file
GUI enables analyst to filter, browse traffic more easily
Export to any number of formats (XML, CSV, PostScript)
2. That IP seems suspicious. Where's it located? Go to Internet, hit database, match IP to geographical locations.
Script!
Flat/CSV file
Let's put into Analyst's Notebook.
Script!
Excel
Analyst's Notebook
3. What did they exfiltrate, and where did it go?
4. Hmmm, I wonder if there's anything else in this dataset I should be worried about…
Print graphs on plotter. Hang on wall.
Write report. Set data aside. Move to next case.
LEGEND
Data
Software on Desktop
Software on Server
MySQL Database

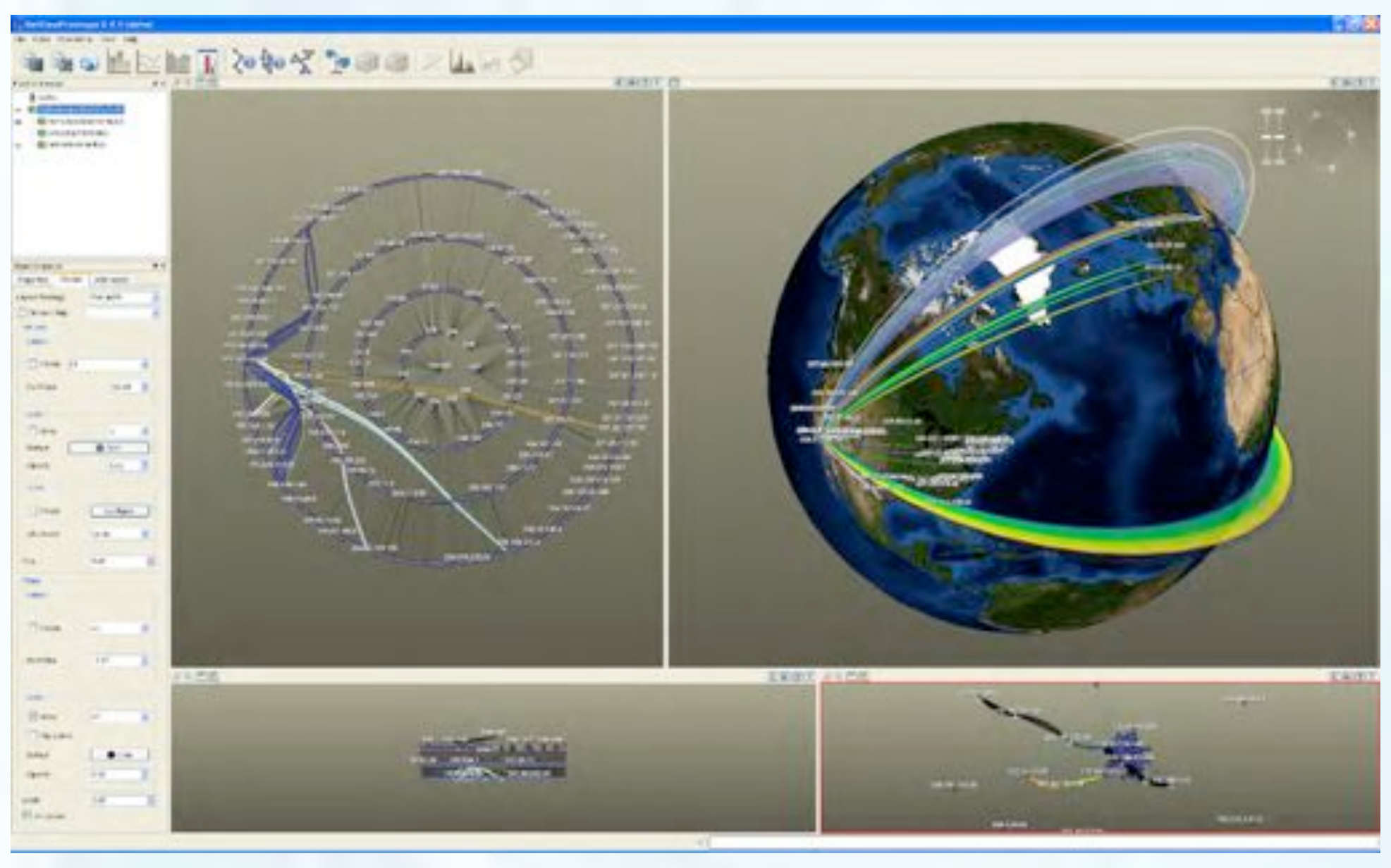
Results (cont.)

Simplified Data Ingestion and Processing Model


1. Gather Packet Data (raw data, not suitable for analysis).
TCP Dump Data
Packet Assembler
Structured network traffic (packets, timestamps, ports, protocols, etc)
Unstructured text (payloads): Emails, pdfs, images, word docs
Postgres Database
3. Analyst parses unstructured text into linguistic elements (nouns, verbs, entities) using one of two tools. Processed data is stored in the database.
Unstructured Text Processing
SAS Scripts
Inxight
LEGEND
Data
Third Party Software
Other Sandia Software
Data Store
2. The packet assembler reassembles raw packets into legible traffic; dumps reassembled traffic into a relational (Postgres) database, which contains both structured and unstructured data.
4. Analyst uses NGC Prototype to query data from Postgres database, conduct analysis and create views. Analyst can select among multiple views, analytical techniques to identify and explore patterns in data.

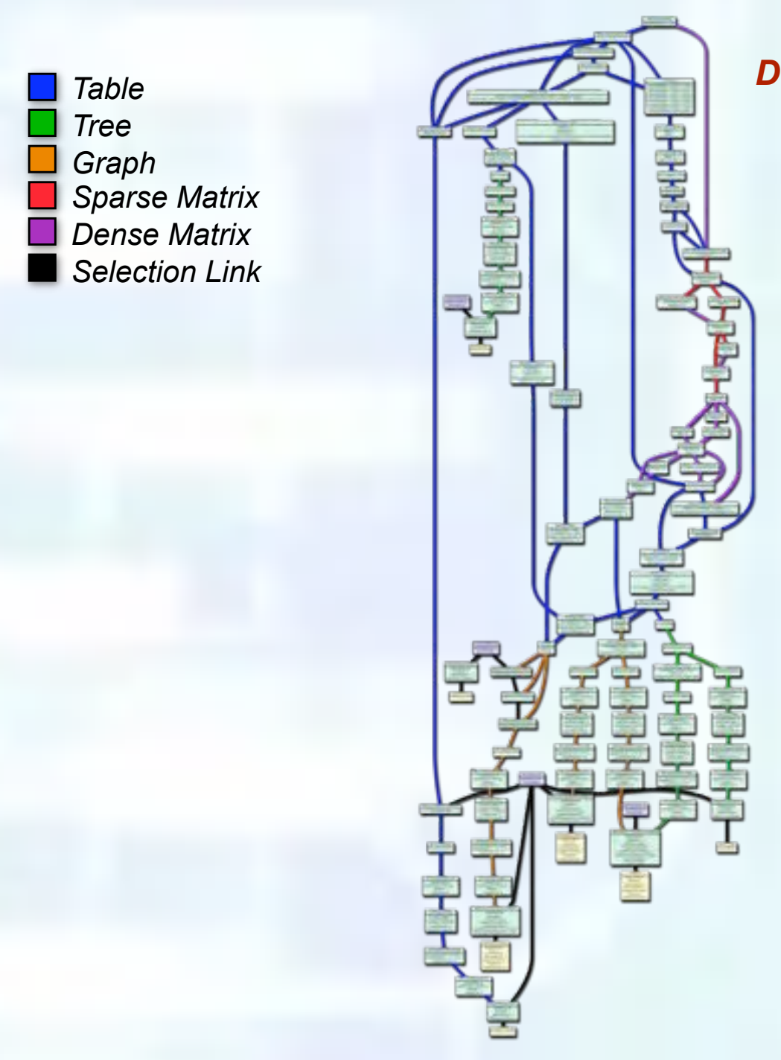
Use Cases Addressed by Cybersecurity Prototype

- Which Machines are Talking to Each Other?
- Which network transfers crossed political boundaries?
- What 'payloads' were contained in the network transfers?
- Analyst needs suspicious (out of the 'norm') behaviors "flagged" for further exploration



Significance

Titan Toolkit Provides End to End Functionality



Disk/Database to Eyeballs

- *Disk or Database Sources*
- *Data Ingestion (Doc Tables)*
    - *Tokenization/N-Grams*
    - *Term Dictionary/Document Dictionary*
- *Named Entity Recognition Extraction (Doc/Entity Table)*
    - *Entity Ontology Processing (Tree)*
- *Term Document Matrix (Sparse Matrix)*
    - *Power Weighting*
- *Trilinos SVD (Sparse Matrix)*
- *"Concept" Matrices (Dense Matrices)*
    - *Document/Concept Dense Matrix (RSV)*
    - *Singular Values (SV)*
    - *Term/Concept Dense Matrix (LSV)*
    - *Log Weighting*
- *Document Cosine Similarity (Edge Table)*
- *G-Mean Clustering (Doc Table/LSV/MIP-RSV)*
- *Document Cluster Processing (Tree)*
- *TableToGraph (Tables/Graph) ← Data Fusion!*
    - *Documents/Entities*
    - *Document Similarities (SVD/Cosine)*
    - *Entity/Document co-occurrence*
- *Hierarchical Document View (Tree/Graph) ← Data Fusion!*
- *Multi-ST Graph Algorithm (Graph)*

- *All data sent to "Views" (Table/Tree/Graph)*
- *All Views are "Annotation" Linked (Link)*

Analytic Capability Already Deployed, Already Shared, and Already Attracting External Funds

The first prototype was a prototype, yet it has been adopted and applied against real cyber data:

- "The work developed in this LDRD has a lot of potential. NetView has already been valuable in addressing several cyber analysis tasks that were presented to Sandia. I have much confidence that through continued collaboration this work will benefit the Labs and the nation in the area of Threat Analysis and Awareness."
- At it's half-way mark, capabilities developed by the Networks Grand Challenge have brought in $924K in WFO research and development funds.
- Capabilities broadly shared with informatics community:
    - 70 publications so far
    - The Titan toolkit is open source (www.kitware.com/InfovisWiki)
    - Published and presented Titan tutorials

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department
of Energy's National Nuclear Security Administration under contract
DE-AC04-94AL85000. SAND2009-4906P

NNSA National Nuclear Security Administration

Sandia National Laboratories