

# INFORMATICS ENABLES DATA-DRIVEN DISCOVERY AND PREDICTION

Contact: Suzanne L. K. Rountree, Computer Science & Informatics, [slkroun@sandia.gov](mailto:slkroun@sandia.gov)

*Sandia is carving a unique niche and positioning itself as a leader in solving complex and large-scale informatics problems, especially those involving national security. ... Math and computer science researchers in informatics have teamed with mission-driven analysts and experts in uncertainty quantification and human factors to focus on decision-making in complex national security environments.*

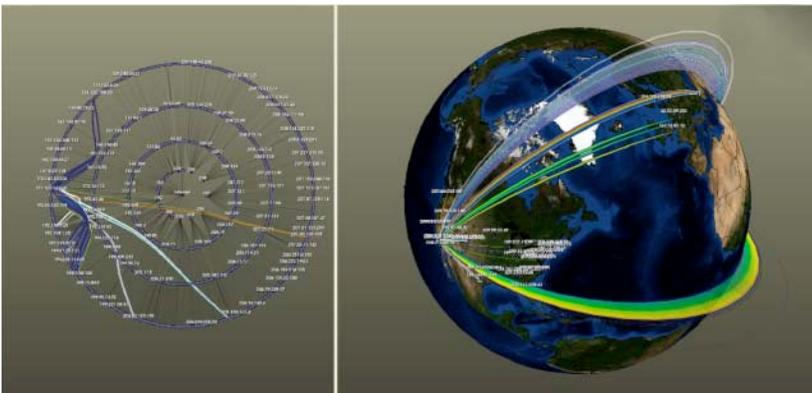
Informatics is the science that encompasses complex information and relationship-based analytic methods to support decision-making in uncertain and massive-data environments. Sandia is carving a unique niche and positioning itself as a leader in solving complex and large-scale informatics problems, especially those involving national security. To that end, Sandia is building a strong research program that utilizes many distinguishing laboratory strengths: discrete math, fast graph analytics, alternative data machines, linear algebra methods, information visualization, computer architectures, high-performance computing, and scalable algorithms.

Social networks, like terrorist networks, are a good example of complex relationships that challenge informatics capabilities. People can be described in terms of their human attributes, the characteristics of their activities, and the interrelationships between their activities that connect them to others. When people engage in suspicious activities that threaten national security — through networks of adversaries engaged in weapons proliferation, terrorism, cyber attacks, and other illicit activities — and when these adversarial networks in turn hide their activities among the complex interactions between legitimate and illegitimate secondary networks (e.g., for communication, computing, supply chain, and financial transactions), then transformational methodologies are needed to ferret out the threats buried within massive amounts of benign information [1].

Math and computer science researchers in informatics have teamed with mission-driven analysts and experts in uncertainty quantification and human factors to focus on decision-making in complex national security environments. One research project in particular, focuses on informatics for discovery and prediction. A sub-team on that project recently demonstrated a thin slice of end-to-end capabilities in a prototype targeting cyber analysis in computer networks. Using fabricated network traffic test data from MIT Lincoln Labs, the sub-team processed raw data and analyzed it to answer four postulated questions: (1) What computers were communicating with each other? (2) What network file transfers crossed country boundaries? (3) What information was contained in the transferred files? (4) What suspicious network activities occurred?

Figure 1 depicts several linked visualizations that answer question #2 [2]. The ring-view (left) shows successive rings of computer networks, sub-networks, and individual computers (white IP addresses) exchanging files across the Internet. The geospatial view (right) integrates an IP address lookup function to show the countries of origin and receipt where the computers reside. The example leverages scalable data queries with analytical graph searches and visualization for presentation to the analyst.

Figure 2 also shows how information visualization can quickly depict unusual or suspicious behaviors that need further investigation (question #4). All the “normal” and



**Figure 1.** Depiction of network file transfers between computer IP addresses (shown as arcs in the ring-view on the left), and network file transfers crossing country boundaries (shown as arcs in the geospatial view on the right).

expected traffic fades to the background visually (as dim blue arcs), while the suspicious traffic is highlighted in a rainbow of colors where each color represents a different form of network activity between two computers (like file transfers). This helps the analyst to focus on the small subset of larger network traffic to answer whether the suspicious activities are legitimate or illegitimate. The scenario applies conditional statistics to the results of a data query before visualizing the results of the graph search. Figure 3 shows the results from powerful graph searches used to cluster and aggregate related information [3]: finding and identifying communities (groups of colored squares) and finding connected sub-graphs that give multiple paths (black lines) between two specific nodes (upper left red and lower right yellow).

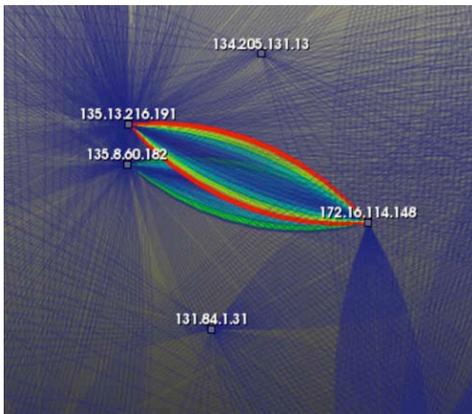
Sandia has begun to address the design of an overall distributed systems architecture, targeting large informatics applications, that leverages the capabilities of different computers to perform the tasks that each does best: data machines for database searches, multi-threaded machines for rapid graph queries, client-server architecture for delivery of visual outputs to user-analysts, and high-performance computing and distributed memory machines for math computations. In a recent success demonstrating part of a

distributed systems architecture, Sandia was able to link data accesses on a Netezza data machine with visual analytics running on our Red Storm high performance computer (Figure 4). Our early prototype and systems architecture design are positioning Sandia to analyze informatics problems of unprecedented scale and complexity.

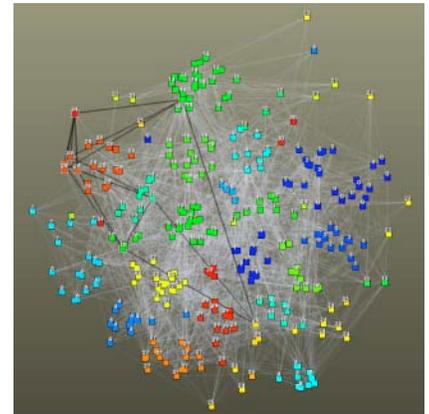
**References**

1. Colbaugh, R., Gosler, J., Glass, K. (2008). Some Intelligence Analysis Problems and Their Graph Formulations, invited paper, *Journal of Intelligence Community Research and Development*.
2. Wylie, B. (2008). GSpace: A Linear Time Graph Layout, Visualization and Data Analysis (VDA2008), San Jose, CA.
3. Hendrickson, B., Berry, J. (2008). Graph Analysis with High Performance Computing, *Computers in Science and Engineering*, 10 (2):14-19.

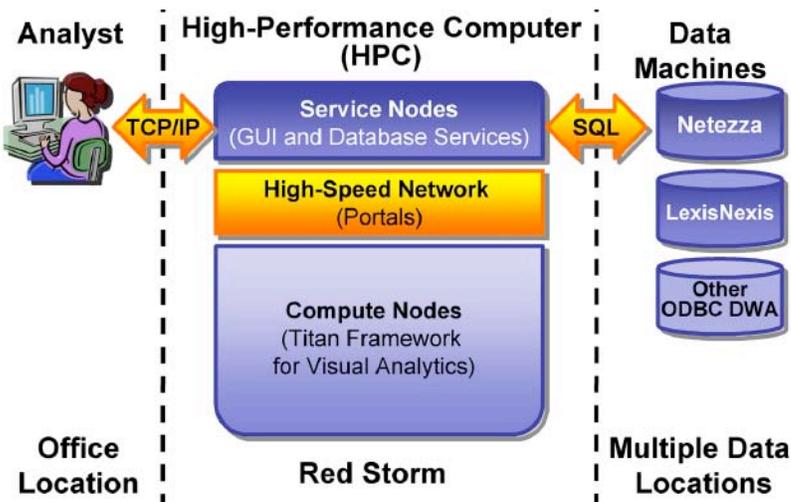
***This work has been supported by SNL's Laboratory Directed Research and Development (LDRD) Program.***



**Figure 2.** (Left) Example of information visualization with brightly colored arcs to highlight suspicious network activities. "Normal" network traffic fades visually when represented as dim blue arcs.



**Figure 3.** (Right) Powerful graph searches are used to cluster and aggregate related information (shown as commonly colored groupings). Connection sub-graphs depict multiple paths between two nodes in a graph (e.g., lines between upper left red node and lower right yellow node).



**Figure 4.** (Left) Distributed Systems Architecture: An analyst sitting at her workstation can run a high-performance computing (HPC) application on the Red Storm computer. The distributed systems architecture enables the HPC application to leverage remote data warehouse appliances (DWA) for efficient data access and management.