Behavioral Modeling to
Detect and Predict Insider Threat

Thermopylae Sciences + Technology, LLC

Administrative POC: Ms. Jeannine Feasel, jfeasel@t-sciences.com
Technical POC: George Romas, gromas@t-sciences.com

1400 14th Street North
Arlington, VA 22209
703-740-8768 (Main)
703-842-8599 (FAX)

Executive Summary

One of the most insidious threats to national and homeland security is the trusted insider. This individual has access to proprietary, sensitive, confidential, and classified information. Typically, there are minimal-to-no restrictions on access to information. In cases where protections are in place, such as Identity and Access Management (IAM) or Role Based Access Control (RBAC), the complexity of heterogeneous infrastructure and integrated software platforms create a variety of vulnerabilities and access paths. The current approach of defense-in-depth, using state-of-the-art technologies and solutions, has not significantly reduced the insider threat. Instead, improvements and reduced cost in consumer, wireless, and storage technologies has made it easier to access and exfiltrate restricted information.

Thermopylae Sciences + Technology, LLC (TST) and subcontractor Recorded Future (RF) propose to focus our research on aspects of the information – an individual's daily behavior on enterprise systems, his or her presence and perception on the Internet, and common data access patterns. Our team will use proven, existing technologies as a basis to build and refine behavioral and predictive models to help identify likely or ongoing insider threats and attacks.

The team will use both DHS-sponsored environments, Cyber Defense Technology Experimental Research (DETER) and the Protected Repository for the Defense of Infrastructure Against Cyber Threats (PREDICT), as well as message traffic and web sites, to build these models. These models will be validated using actual intrusion and attack data and web traffic. The goal is to develop the capability to correlate behavior, access patterns, and character to detect insider threats and predict likely attack and data exfiltration events.

The second part of our research will be to, again, build on existing technologies to develop 4D visualizations of the insider threat common operating picture. The goals are to clearly display vulnerabilities, as well as detected and predicted attacks. Leveraging our current solution platforms, this environment also allows analysts to collaborate.

TST and RF both have experience in managing IR&D projects. Our team is very familiar with leveraging virtualization to create and operate test environments. Our solutions are also heavily dependent on effective ingest, processing and analysis of large volumes of structured and unstructured data sources. We perform our research and develop our solutions based on current agile and iterative processes.

This proposed research has direct commercial applications not only for cybersecurity, but also for fraud detection, forensics and other areas.

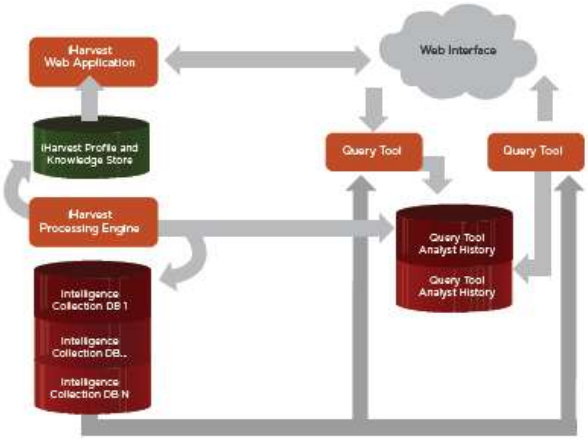
Major Gaps	• Behavior-based access control • More pervasive monitoring and profiling
Collect and Analyze	• Establish insider behavior and insider misuse data sets. Privacy concerns must be addressed. • Models are needed to represent both normal and abnormal insider activity. However, past experience with pitfalls of such models needs to be respected.
Detect	• Detection of insider abuse and suspected anomalies must be timely and reliable. • Data mining, modeling, and profiling techniques are needed to detect malicious insider activity. • Better techniques are needed to determine user intent from strict observation, as opposed to merely detecting deviations from expected policies. • Prediction and detection need to be effectively integrated.
Near Term	• Compile and compare existing studies relating to the insider threat. (Detect) • Develop data collection mechanisms and collect data. (Detect)
Medium Term	• Feature extraction & machine learning mechanisms to find outliers. (Detect) • Procedures to reliably and comparably evaluate insider threat protection methods. (Detect) • Develop behavior-based security. (Protect)

Technical Approach

TST and RF have current production technologies and solutions that can be used as a basis for this proposed insider threat research. TST licenses two products: iHarvest - models behavior of individuals or sensors, and those models can be modified according to specific requirements; iSpatial – 3D visualization platform. RF's Temporal Analytic Engine and dashboard processes large volumes of open source, unstructured text to create timelines of events that can be projected into the future. TST also leverages our Google Enterprise Partnership and subject matter experts to develop innovative solutions that satisfy our customer's mission requirements. Our core strengths in processing and integrating large volumes of data and developing 3D visual interfaces have been utilized in a variety of solutions, including geospatial applications, intelligence analysis, mobile development, cybersecurity, collaboration, and cloud computing. The team's combined capabilities will provide unique and effective research and development. TST devotes a upwards of 10% of its budget to R&D and has clearly demonstrated its R&D capabilities through rapid development of capabilities for the Army G2 and the SOUTHCOM.

For this research effort, TST will use our iHarvest ("Interest Harvest") product as a baseline to enhance and refine according to the requirements of this insider threat task area. iHarvest is an automated, simulated artificial intelligence system that monitors user activities unobtrusively, continually building and refining profiles based on TST's advanced, proprietary algorithms. User profiles can be custom-tailored and used to link analysts with common interests, forming collaborative alliances and information sharing opportunities at both the tactical and strategic levels. While current implementations utilize iHarvest to connect people, we have built in some cybersecurity capabilities that compare current activity against normal behavior. The graphic below (Figure 1 – iHarvest Architecture) depicts a typical deployment.

Figure 1 – iHarvest Architecture



Our teammate, Recorded Future, organizes the web, and all unstructured data, in a radically new and useful way. The world's 24x7 media flow is filled with temporal signals, including reports of what's transpired or statements of what's expected to come. Recorded Future's linguistic and statistical algorithms extract time-related information and through temporal reasoning we structure the unstructured. We help users understand relationships between entities and events over time. In doing so, we've formed the world's first temporal analytics engine.

The company was formed out of research conducted at the University of Maryland, MIT, and Chalmers University in Sweden. In 2009, In-Q-Tel invested in Recorded Future to help mature the technology for use in the intelligence community, and made a second investment in 2010. In addition, Recorded Future was awarded a research grant from VINNOVA, the Swedish Government Innovation Agency, in 2010.

The third component of our research is TST's iSpatial platform. iSpatial provides an environment that can process multiple data sources and formats, and display them in a single, 3D interface. Through implementations for our customers (e.g. SOUTHCOM, State Dept.), TST has developed the following iSpatial capabilities: visualize and track data on a 3D globe; manage real-time information and mobile messaging devices directly from a Web-based user interface; view customized data layers of 3D terrain, models and user graphics; plan and simulate missions; and collaborate and share data instantly in real-time.

Unique Approach

There are several areas of our approach that makes this research interesting and unique. First, the integration of TST's behavioral modeling and RF's predictive analytics provides a platform of proven technologies to build on. The individual strengths of our technologies are complementary. iHarvest's behavioral models can be improved and refined by RF's ability to discern an individual's character and actions from Internet content. RF's predictive analytics can be enhanced through iHarvest's models and its ability to connect analysts with similar interests. Additionally, we are data and analytics focused. We can ingest, process, and integrate large volumes of existing data with disparate formats. TST and RF can provide additional fidelity to this information by using iHarvest to build models of data characteristics and using the Temporal Analytic Engine to detect content patterns.

Since these technologies already exist, we can focus our research on improving the accuracy of our models in detecting and predicting insider threats. For example, we would like to extend our

behavioral models through the use of Bayesian Networks, utilizing some of the research concepts presented in the paper "Detecting Threatening Behavior Using Bayesian Networks", developed jointly by George Mason University and Information Extraction And Transport, Inc. under ARDA contract NBCHC030059. Our research efforts also involve outreach and partnerships with the following groups:

- MIT Computer Science and Artificial Intelligence Lab (CSAIL): http://www.csail.mit.edu/
- George Mason University Volgenau School of Engineering: http://www.cs.gmu.edu/
- Draper Laboratory: http://www.draper.com/

Test Environment

As stated above, we will use the DETER testbed and PREDICT repository, in conjunction with RF's ability to search Internet data. TST has already gained access to PREDICT and is currently evaluating the available datasets. TST and RF will also utilize our iHarvest and Temporal Analytic Engine technologies within this testbed. In addition, both companies will utilize our respective visualization solutions to develop more effective interfaces for detecting and predicting insider threats and attacks.

Personnel and Performer Qualifications and Experience

Both TST and RF have a strong focus on innovation and IR&D; most of our market-facing products and solutions have sprung from research efforts. Each of these efforts is planned as an actual project, with specific timelines, resources, and goals. If we lack any expertise to accomplish these goals, we have the ability to reach into organizations like Google or academic research laboratories. Below are listed the key personnel from TST and RF that have extensive experience in research and innovation, and will lead efforts related to this.

George Romas	George Romas is the Chief Technology Officer for Cyber at TST and has over 28 years' experience within the Intelligence Community, providing his strategic and tactical expertise across a wide range of technologies. George holds BS degrees in Computer Science and Economics from Union College and has done postgraduate work in the MS program for Computer Science at Virginia Tech. His experience includes modeling and simulation, systems programming, database development, systems engineering, cybersecurity, and enterprise architecture - with a strong focus on intelligence analysis solutions. Taking hiatus from the federal sector, Mr. Romas co-founded an Internet startup in July 1999 to develop a network-based multilevel security (MLS) appliance built on a trusted operating system. Prior to that, he worked at another startup that was developing control solutions for robotic manufacturing assembly lines. George holds an active TS/SCI security clearance.
Dr. William Ladd	Dr. Bill Ladd is Chief Analytic Officer at RF. Bill holds a PhD in Statistics from the University of Wisconsin-Madison and a BSE in Chemical Engineering from Princeton University. Before joining RF in 2010, Bill was SVP of Technology and Operations at Genstruct. There, Bill led the development of an extensive knowledgebase derived from scientific literature and the application of artificial intelligence methods to this knowledgebase to address pharmaceutical research challenges. Prior to Genstruct, Bill was responsible for analytic strategy for Spotfire (now TIBCO) and Gene Logic, and was a statistical analyst at Merck. Bill consults with several high tech companies and has published numerous peer-reviewed papers. At RF, Bill provides the cornerstone for our analytic approaches in converting unstructured text data to actionable findings. Bill holds an active TS/SCI security clearance with full-scope polygraph.

Commercialization Capabilities and Plan

TST's and RF's product solutions have gone from research projects to COTS and GOTS packages. TST has packaged its intellectual property into the iSpatial, iHarvest, and Ubiquity commercial platforms. The iSpatial/iHarvest integration is available to government, NGO, and other entities supporting Haiti relief efforts on the open web site http://www.3dudop.org. Our Ubiquity mobile platform (http://www.ubiquitymobile.com) allows users to create their own Android or iOS mobile apps. TST has used this platform to commercially deploy integrated, widget-based, GPS-enabled, mobile dashboards that can utilize augmented reality with camera phones. RF has also deployed its research on a freely available Internet site (http://www.recordedfuture.com). They have also commercialized their offerings with an innovative subscription model, with free, individual, group, enterprise, and Application Programming Interface (API) pricing levels.

Our close Google Enterprise Partnership affords us some unique advantages, including marketing opportunities, customer introductions, direct access to product teams and technical resources, and use of preview or soon-to-be-released Google products or components.

Costs, Work, and Schedule

As a Type II effort, this 24-month project will perform work according to the major milestones listed below. Teams of software developers, systems integrators and subject matter experts will develop capabilities using PREDICT and open source datasets. Those capabilities will be tested via defined experiments and test scenarios conducted on DETER. Performance metrics will be developed along with each capability, monitored during testing, and analyzed and reported to DHS. Costs will easily be contained within the Type II budget, since our team is starting with a baseline of mature products and capabilities.