Direct to Phase II SBIR Topic SB133-002

Defense Against National Vulnerabilities in Public Data

FINAL REPORT

Defense Against National Vulnerabilities in Public Data

Open Data Threat

Contract: W911NF-14-C-0031

Period of Performance: March 13, 2014 – January 31, 2017

Prepared for: Mr. Wade Shen (wade.shen@darpa.mil), DARPA

Brian Sandberg (brian.sandberg.ctr@darpa.mil), DARPA

Joseph Traugott (joseph.f.traugott2.civ@mail.mil)

Debra Leuschel (debra.l.leuschel.civ@mail.mil)

Prepared by: Rich Salaway (rich.salaway@istresearch.com IST Research

Submitted: 28 February 2017

Direct to Phase II SBIR Topic SB133-002

Defense Against National Vulnerabilities in Public Data

Program History

IST Research identified an emerging threat to national security and a growing problem in the availability of organized computational and methodological approaches to confront it:

"We live in a world of unbounded data that is perpetually bombarding us with new information. The deluge of data also brings with it a host of unrevealed vulnerabilities that could have significant impacts on national security. Trying to make sense of this morass of unstructured and structured data has become increasingly challenging. Not only do we have an explosion both in the number of social services, but also the amount of data any one emits. In addition, the most interesting data is often not available through API's but requires complex automated scraping routines. Lastly, we still have the legacy of growing stores of static data on critical assets and various methodologies for assessing their vulnerabilities."

The challenge created by the evolving diversity of new data sources is the numerous techniques needed to exploit the data to discover vulnerabilities. Algorithmic approaches ranging from Bayesian theory to statistical mechanics and many more are used to exploit data. Discovering these vulnerabilities requires having the ability to implement a wide range of algorithms in a rapid and lightweight manner. Computation across dynamically configuring topologies for a wide range of algorithmic approaches is exceedingly difficult.

Objective

The initial goal of the ODT project was to examine how these emerging data sources posed risks to exposing domestic national security interests through unintentional disclosure of sensitive information through open data like social media. In the process of developing methods to quantify and track the domestic open data threat, DARPA redirected the focus of the research to using open data to track and analyze the threat of ISIS.

Under this directive, the team focused on tracking the activities of ISIS on social media and related information sharing resources to examine issues such as:

- Dissemination of propaganda
- Radicalization and recruitment of Western assets
- Network structure of ISIS operatives on social media

A variety of algorithmic approaches were tested to provide substantive insights and answers to these questions. These methods were consolidated into an ISIS daily report, which aggregated a daily collection of data on ISIS social media activities to create a set of indicators for tracking and analyzing propaganda, recruitment, and networking activities. The daily reports were fueled by a live pipeline that identified potential ISIS activity on social media and created a dossier for

human validation through an automated reporting system. Once validated, new ISIS nodes were added to the pipeline for content collection and aggregation.

The experience with the ISIS daily reports drove the creation of a platform that allows arbitrary new data sources to be added to the application and re-purposed by users for a variety of tasks. Discovered data sources can be used to fuel a pipeline in the application that further transforms or enriches the data collected. These transformations and enrichments can also be published and reused by other users of the system.

Once the pipeline collects the data, it can be transformed into a variety of data formats for static data analysis. To help facilitate further analysis, a Jupyter notebook extension was created that allows users to stream their snapshots into a notebook for further analysis. The extension also allows users to publish their snippets of analysis code as well as discover other user's snippets for their own use. Snippet discovery is keyword based and includes a compatibility index that provides user awareness of other methods working with their specific data.

The initial prototype focused on the analysis of streaming data and facilitating collaborative analysis of it. The team then focused on integrating the analysis of static data sets with the streaming system that was built in phase one. One of the challenges the team found in this task is that many of the most interesting data sets are too large to make local compute feasible. For example, the year of geo-located Tweets obtained from Twitter for the project had 2.1 billion messages.

To assist in facilitating analysis of these comprehensive static data sets, the team developed Juno. The Juno project created an infrastructure for Jupyter notebooks to remotely execute code against kernels hosted where large data sets are natively stored. This effort includes the creation of a Jupyter routing engine that allows a local or hosted notebook to execute tasks against remote kernels deployed close to large hosted data sets. This means that users do not have to download data to their analytical environment, but can leave data where it is and send their methods/code to the data and have the results displayed in their notebook.

Going a step further, the team integrated Juno with their pipelines through a library called Machine, which allows streaming pipeline to operate natively on a remote Jupyter kernel. This advancement allows users to perform both streaming and batch operations in a Jupyter notebook, by using Juno to run both classes of computation where it is most appropriate for the operation.

Technical Objective and Significant Events by Task

Objective: Develop a proof-of-concept system that can provide automatic feedback on the measurable risk inherent with various collections of data.

IST Research architected and prototyped a solution by working collaboratively with government and industry partners. The designed platform provides algorithmic computation in real time, not only generating results that are machine-readable but are structured to be understandable by humans. This back-end provides the stream of potential vulnerabilities based on algorithmic orchestration geared towards specific subject matter domains. The platform is a lightweight distributed platform that allows for the quick orchestration and training of algorithms from a variety of domains and programming languages.

The philosophy used provides a simple and efficient platform for identifying, sequencing and training a custom set of algorithms and data geared to a specific domain and problem set. Specifically, an approach and platform designed from its inception to ingest multi-dimensional real-time data and provide a dynamic metadata service to link the most appropriate algorithms with all incoming data sources.

Significant Events by Task:

Task 1: Privacy Policy Submission

Objective: Develop a privacy policy for the program that allows for aggressive research while guaranteeing that individual privacy is protected.

- IST Research developed and implemented a comprehensive privacy policy for the program that covers the acquisition, processing, storage, and presentation of the data collected through the life of the program.

Task 2: Data Acquisition Architecture

Objective: Develop a data acquisition architecture that can successfully ingest 1,000,000 records per hour from up to 100 different open data sources.

- Developed and operate a data acquisition architecture comprised of the four following major components:
  - Robust website scraping architecture capable of simultaneously scraping records from 100 different websites. This component was designed to operate even in the presence of counter-scraping technologies.
  - Ingestion of GNIP social data platform feeds which provide full fidelity access to over 20 social data providers.

- Investigation and ingestion of subscription based precision data sources (Business Intelligence Databases, Monster, others).
- Flexible data architecture that allows for ingestion of heterogeneous data flows and prepares data for transformation, streaming into a sensor system, and loading for visualization.

Task 3: Human-in-the-Loop (HITL) Data Attribute Enhancement Capability

Objective: Develop the ability for a human micro-tasking workforce to rapidly add cleansing and structure to unstructured or purposefully obfuscated data sources.

- Building from microtasking interfaces that IST Research developed for its crowd sourcing work around the world, the team developed a fast, human-in-the-loop data attribute enhancement interface. This interface is configurable to serve very small tasks or a large group of users, allowing them to add structure or specific attributes to data items for rapid image comparison and/or data disambiguation. The capability was designed so the operational team can quickly launch new tasks or a series of tasks to find, identify, translate, and categorize information of interest. This system was also designed to help train algorithmic sensors in the next tasks.

Task 4: Develop Vulnerability and Mitigation Algorithm Development

Objective: Develop algorithms to provide vulnerability measures for real time streams.

- Algorithms were developed under this task that provide real time vulnerability measures for dynamic data flows. New data sources are emerging at increasingly rapid rates. These sources range from social data to data emanating from emerging ubiquitous computing environments and every imaginable data source in between. A flexible approach was required to provide vulnerability measures for such dynamic data. Since the content and topical specificity of data sources varied so widely, a structural approach was used to design measures that would scale effectively.

Task 5: Real Time Orchestration System

Objective: Develop an orchestration framework that combines the family of algorithms from Task 4 into an effective alerting system.

- IST and its partners developed a platform specifically engineered to provide rapid algorithmic orchestration across dynamic heterogeneous data streams; streams as prolific

as Twitter's 27,000 events per second or dynamic updates from automated web scrapers like Roxy, operating at 1,000's per day.

Task 6: Risk Visualization Interface

Objective: Develop a demonstration interface for illustration of risk.

- Delivered a set of tools that organizations and governments can use to understand and make decisions based on the vulnerability of their open data footprint. An intuitive human machine interface was developed to allow for organizations to launch discovery projects, query real time feeds, understand risks and the developing risk posture of their organizations.

Task 7: Design Mitigation Methodologies

Objective: Design a mitigation framework that couples prototype recommender sensors with a human interface that allows for organizations to train the system to proactively search for known vulnerabilities.

- Using human assisted learning, a series of deployable sensors / classifiers were created and deployed through a mitigation framework to assist in proactive notification of previously identified vulnerabilities. Alerting mechanisms were developed to notify organizations when similar vulnerabilities are seen again.

Task 8: Commercialization Activities

Objective: Create robust real time demonstration capabilities that are deployable to potential corporate and government customers.

- IST Research developed several high impact demonstrations regarding the need for technologies to measure and inform organizations on the level of risk present in their open data footprint.

Task 9: Program Design and Management

Objective: Provide oversight across all components of program execution from technical design to business and program management, to ensure timely and complete execution of contract objectives.

- IST Research provided the overall technical and operational management of the program, leading design efforts. Design efforts were driven by a process that IST Research calls Double Helix Development. Our Double Helix Development refers to the tight coupling

of operators and technologists as they team together to explore the possibility of radical changes in the way operators perform their jobs while being equipped with revolutionary technologies. True innovation and viral acceptance of new concepts comes most often from environments where technology and operations can coevolve.

Option 1

Task 1: Design and Develop Proof-of-Concept Integration with Static Data Systems

Objective: Design and develop a proof-of-concept architecture that combines real time approaches with static data analysis.

- Provided identification, visualization and recommendations based on the risk of open data flowing through real time or near real time environments; families of large scale, static, open data stores that contain vulnerabilities in their content. Although the scope of this work was *not* to develop robust tools and techniques to address static data stores, a proof-of-concept architecture *was* designed and developed. IST Research framed possible solutions for incorporating the fully developed real time capture, sensing, and orchestration system into a larger system that leverages the static data side of the equation.

Option 2

Task 1: Develop and Deploy Threat Monitoring and Detection System for Open Data

Objective: Combine proven components of previous tasks into a fully operational system and conduct real time operational deployments of the system in support of government and commercial customers.

- Combining the real-time orchestration components with static data stores and visualization components allowed for operational trials of the system to be completed. IST Research fully integrated all components into a prototype system supporting at least three operational trials supporting mutually agreeable end customers.

Technical Challenges

The ability to conduct large scale collection, aggregation, and viewing of dissimilar data sources coupled with real-time algorithmic analysis on those data streams to analyze both emerging and existing threats was a challenge with many complex variables and dependencies, however significant technical issues were not encountered.

Important Events

IST Research fully integrated all components into a prototype system supporting at least three operational trials supporting mutually agreeable end customers.

Significant Hardware/Software Development

2reation of a highly iterative and flexible framework providing decision makers with a tool with which they can rapidly view their organizations open data footprint and discover existing vulnerabilities and identify future threats of the same type. #

Satellite Imagery Adapters	Planet and Carto Adapter
Juno Services	Vulnerability Assessment Framework Computer Model
Facebook Graph API Connector	Digital Globe GBDX Kernel Launcher
Nervana Infrastructure Adapter	Jupyter React Library
Timbr Machine	Threat Monitoring and Detection System
Risk Visualization Interface	Human in the Loop Attribute Enhancement Interface
Linear SVM Classifier for Tweet Detection	Intuitive Human Machine Interface
Deployable Sensors / Classifiers	Robust Website Scraping Architecture
GNIP Social Data Platform Feed	Real time Algorithmic Computation
Data Analysis - Jupyter Notebook	Visualization Tool Integration
Algorithm Orchestration Framework	Data Acquisition Architecture
Real Time Alert Web Interface	Analytics Development
Analysis Dashboard	Data Collection and Management Web Interface

Software Deliverable

Actual program software will be delivered to the Contract Office Representative under separate cover in DVD format.

Special Comments

This product is the creation of a highly iterative and flexible framework that provides decision makers with a set of lenses with which they can rapidly view their organizations open data footprint and discover existing vulnerabilities.

These lenses are comprised of advanced extraction tools, real time sensor algorithms, and lightweight visualizations which enable a human team to look at the data space, absorb findings quickly, make decisions or connections, refocus the lens and iterate as quickly and as often as they desire.

As vulnerabilities emerge, the system allows for sensors to be placed into the same lens to identify future threats of the same kind, and allow organizations to stay on top of their public data footprint.