BNL-105470-2014-IR

BROOKHAVEN NATIONAL LABORATORY'S CAPABILITIES FOR ADVANCED ANALYSES OF CYBER THREATS

Michael DePhillips

January 2014

Brookhaven National Laboratory

U.S. Department of Energy

DISCLAIMER

TABLE OF CONTENTS

TABLES

FIGURES

O.O INTRODUCTION

This paper attempts to speak to two distinct, and in the case of Brookhaven National Laboratory (BNL), separate audiences. An attempt is made herein to introduce the domain of National Security, often represented by the Intelligence Community (IC) to the academic culture of open science and vise-versa. Acquaintance brings no risk; only understanding through which opportunity may arise.

The paper is divided into the following sections:

Section 1: Executive Summary: A concise treatment of the major topics and conclusions in this paper.

Section 2: Overview; A general introduction to the ideas and issues that led to the writing of this document.

Section 3: Networked Computer-inspired Intelligence Topics: A brief overview of some major issues facing workers and researchers engaged in various aspects of national security, often represented by the IC.

Section 4: Brookhaven as a National Laboratory: A discussion of the unique programs and scientific expertise that can be targeted at Brookhaven National Laboratory, to solve issues detailed in Section 3.

Section 5: Poised For Solutions; A description of the overlap between Sections 3 and 4 that might well lead to a fruitful collaboration between BNL scientists and those involved in National Security programs.

Section 6: Sample Projects: A number of rough sketches of potential projects, intended to spark discussions and thought.

Section 7: Summary and Conclusions

The ultimate goal is to build a bridge of understanding between the culture of science at BNL and the needs of the IC. From this understanding, decision makers will be alerted to rare opportunities to apply, in a parallel trajectory, those unique, advanced techniques already operational at BNL that can solve complex problems in national security facing the nation's networked computers.

1.0 EXECUTIVE SUMMARY

BNL has several ongoing, mature, and successful programs and areas of core scientific expertise that readily could be modified to address problems facing national security and efforts by the IC related to securing our nation's computer networks. In supporting these programs, BNL houses an expansive, scalable infrastructure built exclusively for transporting, storing, and analyzing large disparate data-sets. Our ongoing research projects on various infrastructural issues in computer science undoubtedly would be relevant to national security. Furthermore, BNL

frequently partners with researchers in academia and industry worldwide to foster unique and innovative ideas for expanding research opportunities and extending our insights. Because the basic science conducted at BNL is unique, such projects have led to advanced techniques, unlike any others, to support our mission of discovery. Many of them are modular techniques, thus making them ideal for abstraction and retrofitting to other uses including those facing national security, specifically the safety of the nation's cyber space.

The significance of protecting our computer networks is readily understood. The proclamation of President Obama that the future of our nation depends upon it provides a rallying point toward which to mobilize our resources and efforts. In particular, his comment that research and security must work together in finding solutions for protecting our computer networks should resonate deeply within the National Laboratories. Each laboratory has separate areas of expertise partly based on their researchers chosen experiments; therefore, each offers singular knowledge, expertise, and associated unique technologies. Inclusive in this expertise is the extremely valuable know-how that went into developing the processes needed to create and support the science.

These experiments, and equally interestingly, the processes formulated to sustain them, can be modified, retrofitted, and applied to networked computer security and its analysis. This approach already was adopted quite successfully in the financial sector. However, the lack of its widespread use is not hard to understand. Keeping in mind the basic open research mission of the National Laboratories, researchers there, may not have looked to apply their discoveries to anything other than the endpoints that they were funded to attain. Nonetheless, at a base level, the novel advanced techniques discovered at the National Laboratories and used successfully to move, store, and analyze petabytes of data from physics experiments undoubtedly will prove be of extreme interest to security professionals. For example, novel approaches to moving and storing large, disparate data-sets, which are necessary for finding, for example, Perfect Liquid, encompasses major advances. These advancements were successfully documented, and might well be used for National Security and the IC. There also are pockets of important researchers and collaborations being conducted by scientists in industry and at BNL's neighboring Universities (SUNY Stony Brook, Columbia NY) on subjects that can be directly related to national security cyber initiatives.

BNL has significant potential for research to address problems related to national security, cyber security, and the IC. Section 3 discusses the problems and open-ended questions that could be addressed by ongoing projects at BNL. Table 1 lists techniques that were developed by a particular project that could be considered applicable to an issue being worked on by a National Security compartment. This table is the first instance of "crossing" these disciplines, therefore some terminology may not be immediately clear. These terms are defined and described in later, more appropriate sections.

Table 1. National Security Cyber Issues and BNL Projects

Advanced Technique	Issue	BNL Project
Pattern Recognition	Information "Gap Analysis"	Relativistic Heavy Ion Collider (RHIC)
Disparate Data Analysis	Unclassified Collection Requirements	RHIC / Center for Scientific Computing (CSC)
Bench-marking Analysis	Unauthorized CPU usage	CSC
	Targeting	RHIC / CSC
	Intelligence Collection Requirements	RHIC / CSC
	Operations of Non-state Actors	RHIC
	Cultural Social Structures	CSC
Probabilistic Risk Assessment and Modeling	Prevention of Sabotage/Terrorism/ Insider Threat	Probabilistic Risk Assessment for Accidents at Nuclear Reactors

Table 2 is more specific about such techniques. It lists ones that can be abstracted from existing work underway at BNL alongside the existing project, and the issue that will be considered. This listing is by no means exhaustive; it is simply illustrative of useful activities that can be undertaken with minimal start-up overhead expenses, and be completed with a high probability for success. The projects are detailed in Section 5.

Table 2. Sample Projects

Sample Project	Existing Program	Alternate Trajectory
Data Encapsulation	RHIC	Big-Data and High-Volume Acquisition. Transfer and Storage
Framework	RHIC	Disparate Data Analysis
Unstructured Data Analysis	NY Blue	Disparate Data Analysis
Probabilistic Risk Assessment for Human Systems using Networks	PRA for NRC	Insider Threat Mitigation
Data Diversion/Filtering	RHIC / Triggering	Network Protection and IC Analysis
Cultural Clustering	CSC	IC Hierarchies
Super Computing - Usage Analysis	CSC	Stolen CPU cycles Unauthorized Usage

The scale of these listed projects vary. Often projects are generated to answer larger questions, issues, or problems such as pattern recognition or anomaly detection. These are valuable, needed approaches; however, many components are required to complete these tasks. Many subtleties in techniques, necessary infrastructure, and nuances can be gleaned from a solid knowledge base that comes from mature experiments. When considering modularization, each major step of a large experiment potentially might be modified for research in national security applications. These smaller projects, while producing usable products, can be then combined into a larger-scale accomplishment.

2.0 GENERAL OVERVIEW

It would be difficult to overstate the detrimental current and future impacts that would result from networked computers aggressively being used for deceitful or destructive purposes. The effects would impact almost every aspect of the nation's economy and security. In less than a decade, we have seen an important shift in cyber security practices, from stopping nuisances, albeit sometimes destructive, to being essential considerations when connecting a device to the Internet. Even with the over-simplified examples that are given, it is not hard to accept the enormous impact of networked computers on the various facets of our society. Commercial enterprises that protect networks and the data stored therein have grown, creating a thriving industry protecting the assets of established communities. The academic landscape of computers in science has expanded wherein security has become an established tract of study and research in many programs. Law-enforcement entities now have cyber squads, and our defense groups expresses interest in both the offensive- and defensive-aspects of cyber warfare. On the larger scale of national security, the curbing of cyber-based intrusions for the purpose of illicitly obtaining data is a major concern of government agencies. Finally, the prevention of the adverse effects of hacking undertaken by nation states or by groups bound by a common ideology is regularly discussed and stressed in the training of new employees.

There have been many assertions of the high priority of protecting our networks. Often implicit in them is the notion that there is a need to understand the techniques, psychology, and science that is used to engage in this exploitation. However, cyber-security best practices are based on reactive defensive-postures to previously identified threats that perhaps are dictated by available technologies. Preventative techniques typically are sequestered to end-user training and awareness, or passively identifying incoming-traffic based on indicators. It can be argued that these preventative techniques also are reactive because they are based on past events. These techniques can be effective but only in a limited capacity since they do not anticipate change nor predict patterns; they are only reactive. Consequently, there is an abundance of cyber events to address. These events, resulting in an intrusion, are exploitations of unknown or unaddressed vulnerabilities. If the intrusion is a successful disruption, the purpose could be the establishment of persistence on a network for the purposes of extracting data, or to prepare the battlefield for a future cyber-attack. The goal of these malicious attacks could be to shut down vital infrastructure, or be aimed at the destruction of the network or the machinery controlled by nodes on the network. Every minute of each day a barrage of intruders probe systems and continually attack networks. This is a daunting fact in terms of the effort needed to defend against an intrusion. However, looking at these phenomena from the perspective of data acquisition, a tremendous amount of information, albeit masked by noise, is being delivered to all who have an

inclination and the skills to analyze it. It is within these advanced techniques needed to provide this type of analysis where the untapped skill sets of BNL could be employed.

How to analyze unknown threats in "Big Data" effectively and efficiently is the challenge facing network security-professionals. Also, those who wish to gain insights into humans through network data (i.e., the IC) have a keen interest in achieving this goal. The intent of this paper is to discuss the many detailed facets needed to answer questions that already have been developed within the disciples of the science conducted at BNL.

Large, seemingly disparate data streams are the hallmark of large physics experiments. Scientific success lies in the ability to harness increasingly larger flow-rates of "Big Data". Success typically is achieved by developing advanced and novel techniques, including methods that ultimately enable predictive models, machine learning, and real-time decision-making. Such techniques often are used to understand rare events and predict probable outcomes for the sake of discovery, and to prevent or control possible calamities. They are in widespread use in the scientific community, but, to date, and to our knowledge, much of the protection of networks relies on reactive techniques based on known incidents in contrast to discovery from the big data available to us. Therefore, the weight of his office notwithstanding, one of the most provocative and astute proclamations made by President Obama in his remarks on Securing Our Nations Cyber Infrastructure (May 29, 2009) was his acknowledgement of the need for research. His talk accented the seriousness with which the cyber frontier must be protected; indeed, he stated our Nation's future depends on it. The president also stated that both security entities must work in conjunction with researchers to harness from the technical marvels this frontier has yet to reveal, as well as protecting ourselves. It is with this charge in mind that BNL brings its world-class scientific expertise to the frontier of networked computers.

3.0 NETWORKED-COMPUTER-INSPIRED INTELLIGENCE TOPICS

Risking oversimplification, networked computer-security can be divided into three categories: 1) Those conducting the day-to-day business of the users on the network; 2) those that are programmed to be used offensively to attack, misuse, or break a trust relation within the network; and, 3) those used to defend the network, which for the discussion would include the nodes used to monitor the network. Due to the considerable rapidity in advancements and the ever-increasing supply of data both transferred and stored, it is difficult to quantify the effect of misuse on society. This then makes it difficult to prioritize which issues or problems are worth exploring in detail, and, from a research standpoint, the myriad issues and variations of problems continually will fill volumes of proceedings and journals. BNL has explored in great detail many issues that can be applied to networked computers with scientific programs that are mature, ongoing, and continuously evolving to adapt to increasing data rates. This reduces and focuses the areas of discussion in this paper to a subset bounded by established expertise at BNL.

The government agencies dependent on 'intelligence" generally are subject to a never-ending intelligence cycle (Figure 1).

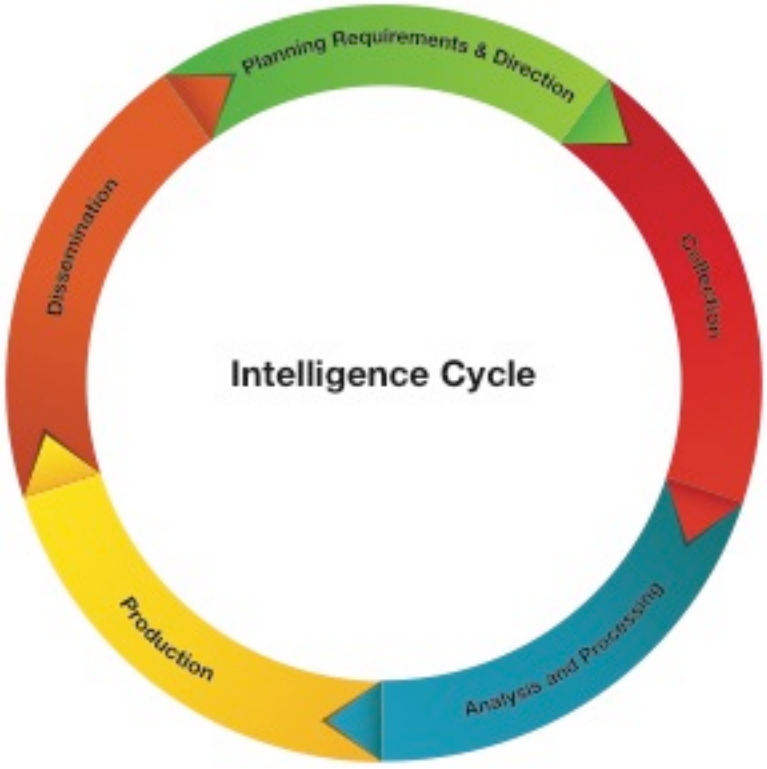

Figure 1. Intelligence Cycle

Networked computers have impacted the phases of this cycle in three ways:

1. They have enhanced the capabilities underlying the completion of each phase.
2. They have expanded the capabilities and understanding resulting from each phase.
3. They can be used to analyze data being used to upgrade or counter the efforts of each phase.

Accordingly, the following subsections were divided into major topics, specific to areas dependent on intelligence, in which networked computers can be used both offensively and defensively. Each section's discussion contains some issues that apply to the section's name, and is constrained by work and discoveries resulting from ongoing programs and achievements at BNL.

3.1 Espionage

> **Issue 1:** *What are the collection requirements and strategies of nations using networked computers to conduct espionage?*

Although there is an abundance of literature, case studies and definitions of espionage, the public forum contains few formal presentations of how espionage is committed in the 21st century. One major difference between our modern environment and espionage of the past is the networked computer. There is no shortage of admissions that it has irreversibly changed civilizations' second oldest game, but other than obvious observations, such as Facebook affording a spy an

abundance of targeted information, little is known as to who or what is being targeted, and what is being collected. In that regard, everything is assumed to be a target, and for good reasons; that is, there are myriad attempts at breaking into the nation's systems. Publically available reports have even gone as far as identifying the nation state of China, as opposed to bands of rouge hackers that happen to come from China, as an aggressive and egregious violator of rules against breaking into networked systems. This report validates a long-term suspicion that there are national plans to gather US information illicitly from our networks. However, it offered little insight into whether there is a prioritization of the collection or if it is truly opportunistic and random. It would be interesting to see if other nations have a specific agenda and hide in the noise of China's obvious onslaught. Also, if their attempts are more specific; specifically what is being targeted is a very interesting question.

Issue 2: *What unclassified information, individually or in combination, is interesting to other nations, or, at worst becomes classified if placed in some special grouping?*

Traditionally a smoking gun in an espionage case would be finding classified information in a place it should not be. This definition is challenged by repeated attempts at stealing and the theft of unclassified information. Whether this falls under the heading of espionage is an interesting point. Colloquially speaking, (i.e., not using the legal definition), a boiled-down definition of espionage could simply be the stealing of secrets. Not everything secret is classified (e.g., PII, proprietary process or design information). Theft implies some breach of trust. So, we could define espionage one way as the illicit collection of information that is not intended for public dissemination, via means that include a breach of trust. With this logic, the breaking into systems by way of a networked computer to steal secret information can also be considered an act of espionage. Again, probably many legal quandaries need to be considered here, but for the purposes of categorizing technical hostilities, this definition will suffice.

There is a trade-off in determining what to steal; spending time analyzing potential targets for a high value yield of information, or grabbing everything possible and then determining if what was stolen is worth analyzing. This is a "pay now or pay later" scenario, with the payment coming in the form of the evaluation of material. Pay now evaluates a target, be it a person or a machine, and assessing if it is worth expending resources to steal its secrets (or persuade it to steal secrets). Paying later is a philosophy of collecting everything possible and the sorting through it to determine if there is anything valuable, or if any combination of the information can be assembled in such a way to reveal something interesting. Both are legitimate collection techniques and both are used by intelligence agencies. Networked computers aid in every aspect of both techniques, including the evaluation as to whether information was targeted or randomly gathered.

Issue 3: *Detection of the Stealing of CPU Cycles*

Stealing CPU cycles also fits the given definition of espionage, especially if the results are used to supply a nation or an organized, non-nation state group, with information that would not normally be accessible to them. An example of such a breach of trust is a user using a cluster or super computer for calculations not stated in the original usage agreement. The end goal of this breach would be to calculate data, sort through data, or to see how it is processed through the

machine, thereby learning about the computing process. This is valuable information, which is not available to nor is it intended for access by the general public.

> **Other Issues:** *The following sub sections add some detail to the general issues stated above when considering networked computers.*

3.1.1 Targeting

> **Issue 4:** *Detection and Influence of the Spotting and Assessing of Targets*

Targeting is another cyclical operation dependent on informational feedback loops (Figure 2).

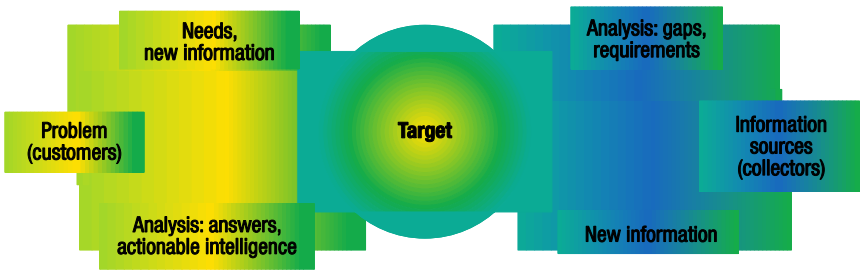

Figure 2. Targeting Cycle

Generally, choosing a useful target falls into two sub categories; spotting and assessing. This is true when targeting a person or using the Internet to target a machine. Each step is made easier by using networked computers compared to when this type of activity was accomplished primarily though human subterfuge. Advancements have been made in efforts to detect, deter, or influence these activities. The means of detection is simple; identify the action when it is being done. To deter or influence the events requires a more sophisticated approach because both presupposes some understanding of what an intelligence service is looking for. This understanding could be simple; for example, identifying a nuclear scientist with access to weapons data, but it could quickly become very fuzzy as the amount of data increases when considering secondary- and tertiary-targets. That is, using one group of people as a means to access a second group of people.

3.1.1.1 Spotting

Intelligence services recruit people to spy for them. They also are looking for repositories of data that may be useful in providing strategic- or tactical-insight into an adversary. The availability of potential candidates greatly increased with the Internet, especially with the immense popularity of social networks. Naturally, the availability of useful information also has risen substantially. Social networks, especially the smaller, subject-specific ones, provide pools of people who have been pre-filtered, categorized, described, and binned for only the cost of signing on to a website. The possibilities of approaching potential assets are virtually endless, and the cost of researching a particular target is low. Subject-specific areas also are helpful

when pooling together particular areas of interest and the people associated with them. Once potential candidates are found the assessing stage begins.

3.1.1.2 Assessing

The ability to assess a potential target used to be constrained by the danger of potential exposure in doing so. Human interaction also may limit the chance of a successful assessment. The anonymity of the Internet provides adequate cover for one person to research, and to repeatedly contact a subject. A false persona can be altered many times without the risk of exposure. Each contact will yield additional information, thereby increasing the fidelity of the assessment. Gathering secondary information about a target also has eased with the staggering amount of open source data available to the public.

3.1.2 Collection

Issue 5: *Uncovering Patterns that Reveal the Collection Requirement*

Collection is the purposeful gathering of information. As mentioned earlier, one method of collection with and from networked computers is the attempt to exfiltrate as much data as possible from an infiltrated system. This process, nicknamed "slurping", also comes with massive attempts at infiltration. Constant phishing attempts are a tangible reminder that this paradigm is real and used in full force.

The concept of gathering as much information as possible and sorting through it later is not new. Wire-tapping operations or capturing electronic emanations from power- and phone-lines (i.e., tempest) during the cold-war provided intelligence services with an abundance of data. There is speculation that so much data was collected that it was never fully analyzed. Now, with Big Data, massive numbers of data sets require advanced techniques to effectively handle, and ultimately, to analyze. Networked computers not only have to provide a way to collect data but also must effectively transport, store, and then analyze what was exfiltrated. This is not easily achieved; therefore, it becomes interesting to guess whether a collection item might be hidden in massive "slurps". Issue 5 stated above implies a desire to understanding a collector's priorities. Also and equally subtle is the discovery of a threshold that triggers the attacking group's decision to escalate their efforts, for example, by crossing the barrier from open source collection to social engineering or hacking. It also is best not to trivialize the sophistication needed to efficiently transport large amounts of data undetected. Understanding the various techniques in which this is done is a vital component of collection and would lead to insights into the issue stated above, as well as the capabilities.

3.1.3 The non-nation-states

Issue 6: *Detection of the Organization of non-nation-state actors*

Traditionally, espionage is carried out by a nation to learn the secrets or intentions of another nation. However, there have always been groups of individuals bound not by political boundaries, but rather by a devotion to an ideology. Networked computers have advanced their capabilities to operate independent of geography, allowing dispersed groups to act as a cohesive

unit. In fact, the Internet has created its own non-nation-state group, "The Hacktavist". These groups are likely involved in activities that fit most definitions of espionage. Even more curious is the looseness of the group affiliation, reacting to events using methods without well-defined command and control, as opposed by a stated goal, known directives, or strategy. Further, finding structure and patterns in behavior that is intrinsically not patterned, or worse, whose patterns are intentionally altered, offers some special challenges. Studying the many dimensions to patterns may bring insight into an organization's priorities or the social structure of these groups.

3.2 Counter-espionage

Much of what was discussed in the previous sub-section also applies to the defense against espionage or sabotage. However, even with much overlap, counter-espionage efforts differ in that they often focus on attribution, and the understanding of an adversary's requirement as opposed to collection. There also is the detection and prevention of acts of espionage; it has its own set of complexities; but these, however, overlap more succinctly with security entities, such as law enforcement. In protecting networks with scientific capabilities three key issues are having the following expertise:

1. To sort, bin, and collate large amounts of unstructured, sometime asynchronous data in an effort to establish attribution;
2. To sort, bin and collate the massive collection of unclassified, open source information to mitigate further attempts, or at least witness and understand, what is being taken.
3. To detect the potentiality and establish a probability that an insider is capable and motived to commit treason.

Implicit in these capabilities is the more subtle and challenging aspect of finding a small signal in a highly noisy environment. This will help in discovering what is valuable to an adversary by understanding which unclassified information is being targeted. It also may lead to the discovery of the technical techniques that an adversary uses to exfiltrate the data from a system.

3.2.1 Asymmetric approach

An asymmetric approach to collection makes it difficult to establish a pattern, so that it is easier for an adversary to hide in the noise of Internet traffic. In this case, finding a signal in the noise must be based on some criteria that will add structure to the nonlinear approach. If the collection approach is indeed random, a different criteria (e.g., targeted information) may be used find a pattern. As mentioned earlier there are many dimensions to patterns, therefore there are many ways to look at seemingly random events.

3.2.2 Collection of unclassified material

Remembering our definition of espionage that asserts that there must be a breach of trust and the capture or theft of something secret, we have a very small cross-section with the unclassified world. Section 3.1 addresses intelligence gathering, which would be collection without both stated criteria met, which certainly is a more prevalent activity. It is important to keep in mind that a correct combination of unclassified material, especially with regard to Science and

Technology that is proprietary or not ready for publication, and, therefore, must be stolen, can reveal information that is meant to be kept secret or may become classified in the proper context. The issue here is the protection of areas that do not explicitly require protection. Protection not only means preventing theft but also understanding the intent to steal it, which would also help to protect the people who are producing the information from potential targeting.

3.2.3 Insider Threat

With regard to counter-espionage, insiders becomes a potential threat when they have access to classified material as well as to a foreign intelligence service. Espionage is committed with an unauthorized disclosure of the material to that service.

The prevention of espionage is preferable to discovering the committed act. Detecting subtle clues that would indicate an inclination of an insider to carry out this crime involves a large effort by nation states and industry. Typically, this effort focuses on the social- and behavioral-traits of the insider. In addition to certain such underlying traits, other dimensions must be in place to consummate the crime, (i.e., access to classified- or sensitive-material, as well as someone to give it to). In keeping with the topic of cyber-intrusion, access via networked pathways to sensitive information is a potential vector for an insider to accumulate data. Such access, for example, could be mapped and identified showing the ease in which someone can get at the information. This is a concept similar to the "traveling salesman" problem so familiar to computer-science students, which tries to find the shortest distance between two points. In this case, increasing the virtual distance between a potential insider threat and valuable sensitive material may suffice to prevent access while the human investigation progresses.

3.2.4 Network Intrusion

It is a common public belief that our unclassified networks are so susceptible to compromise that we assume, unless we encrypt our data, that either criminal or nation-state adversaries will have access to it. Perhaps this is hyperbole with regard to the more closed- and firewalled-systems of an organization, but it begs the question as to whom and what is on our systems. This can be a tricky question, for example patching, scanning, blocking, and using a quarantine space will not detect malicious code that already has gained persistence on a network. Considering the exfiltration of data, most intrusions attempt to create a connection with a command-and-control machine that may deliver more malware or attempt to upload data. The connection of the hacked machine to the command- and control-node is done by a beaconing protocol, outward from the hacked machine within the network to a machine outside the organization's firewall. Checking for beacons to known command- and control-machines is effective, but learning of the existence of such a machine is more challenging. Beacons usually follow some pattern; however, that pattern may initially appear to be random. Also, randomness can be faked; therefore, finding an existing pattern hidden within this obfuscation provides additional challenges. As with nature, complex patterns are not often visible to casual observation, and advanced techniques need to be employed for discovery.

3.3 Intelligence and Counterintelligence

> **Issue 7:** *Uncovering and understanding what intelligence services are collecting, including open source targets.*

This is a super-set of espionage targeting in that it includes information that is not secret. In fact, a keener sense of precision would state that part of the intelligence cycle usually associated with espionage (i.e., spotting and assessing) should be placed in this section due to the amount of open source (i.e., not secret) information used in the process. There are some staggering statistics stating the disproportional amount unclassified information being collected compared to that which is secret, sometimes known as economic espionage. An interesting point with regard to protecting assets is that if information is available for all to see, consume or collect, and an intelligence service is a subset of, and included in "all", then it should be perfectly acceptable for an intelligence service to gather this data. The nuance an acclimated reader will raise is that, at least traditionally, intelligence services are purposeful in their collection. Therefore, what they care about is, in itself, useful information. Unless this information is misinformation used for deception, or is random, then it has value in providing insight into what information a nation is clandestinely seeking. The word clandestine can be challenged here in that we are discussing open source information. However, the word is used because proper channels (e.g., requests for collaboration) are not being followed to gain access to the data. It is clandestine when there is at least an appearance of intent to shield the purpose of collection from observation and discovery.

3.4 Cultural Structures

> **Issue 8:** *Established intelligence social structure hidden within the social structure of a culture.*

All organizations, including the intelligence services, have an organizational structure designed to effectively carry out its mission. Within this structure are defined roles that establish the roles and responsibilities of the employee. Within intelligence communities, some of these positions have colorful names such as handlers, spotters, and collectors. Also, as with most organizations, these positions often have an associated hierarchical structure. Within this structure, these roles have defined relationships between each employee, be it a one to many (e.g., one handler, many collectors) or conversely, many to one. These relationships are joined by communication and data flow. These flows can be mapped forming a direction graph, the people forming the vertices and the information pathway the edges. This mapping is normally and easily done when the vertices are known; therefore, the information flow is defined. The mapping also can be completed, although it is a much more difficult task, by gaining a thorough understanding of the information flow, thereby revealing the vertices and, therefore, the personnel of an organization. Clusters and sub-clusters of the organization can be discovered using established scientific cluster-finding techniques.

3.5 Counter Terrorism

> **Issue 9:** *Predicting the radicalization of an employee, guest, or user or discovering a radicalized person before that person causes harm.*

Network traffic can be used as an indicator to help discover individuals who are being radicalized. However, when this discovery is based solely on an alert triggered by some predefined indicators, the preponderance of identified individuals are bound to be false positives. This is not an indictment of indicators: however, just using key words or site visits by themselves are not adequate to avoid a barrage of false positives. To increase the probability that an indicator is positive, patterns need to be identified and motivations should be able to be established. To that end, statistics need to be gathered and analyzed to increase the probability of detection. Predictive probabilistic risk models not uncommon in the scientific community would be useful in this effort.

4.0 BROOKHAVEN AS A NATIONAL LABORATORY

Understanding the need for, and developing advanced techniques and processes are detailed, rigorous endeavors. Many different scientific- and academic-disciplines may be called upon to firmly capture a working understanding of phenomena. The DOE National Laboratories, in particular Brookhaven, are enclaves in which world-class examples of this synthesis of knowledge is developed and used on a daily basis. The previous section discussed some open issues faced by the IC. The premise of this paper suggests that these advanced techniques could be used to assist in working on these topics. Fortunately, many of these advanced techniques already have been developed at BNL, and are currently being used for its scientific endeavors.

The National Laboratory structure was established to inspire collaboration and competition, viz., the motivating forces toward basic science discoveries. The uniqueness of each of the laboratories, at least initially, revolves around the large expensive machinery that is the basis for experimentation. There could be some overlap between the laboratories in various areas of research, so inspiring competition for funding; however, the machinery involved will not be duplicated, thereby ensuring a unique brand of technology at each individual laboratory to run, extract, and store data from the machines. These data also must be accessed and analyzed. The processes used to achieve this are unique to each experiment and, ergo, each laboratory. Best practices are shared, open source papers are written, but there will always be unique and novel avenues of understanding, research and rigor at different venues based on the machines they run. Brookhaven is illustrious with its avenues of expertise unique to the types of science and experiments being conducted. It is this rigor, and the discoveries used in research being conducted at BNL that can be adapted for use to work on the topics discussed in Section 3.

4.1 High Energy-, Nuclear-, and Particle-Physics

The large physics experiments conducted at BNL are spectacular, and publishing, in premier international science journals the world-class discoveries made that indelibly will impact humanity is “business as usual” for BNL. These experiments study very subtle components of nature and the associated phenomena. The experiments themselves are complex interweaving of machinery electronics and software designed to detect and then convert minute physical phenomena into electronic data. The data, in turn, must be reconstructed to represent the physical event such that a scientist can analyze, study, and understand the original phenomena. This process often requires the processing of very large amounts of data, so large in fact, and moving so rapidly, that novel techniques must be used to capture, store, and transport it.

Multistep processes are created to sort through files of raw data and create a usable, comprehensible data structure. Custom code is written to understand large files of raw data, and transforms them into structures ready for analysis. The analysis undertaken to reveal physical phenomena; however, these techniques can be applied to discover other types of phenomena, including the aggressive and nefarious use of networked computers.

4.1.1 RHIC

Charged with the discovery of Quark Gluon Plasma and succeeding with the discovery of a new form of matter referred to as the perfect-liquid, the operations of the Relativistic Heavy Ion Collider explore the state of the universe immediately following the Big Bang. The research involves the acceleration and collision of different types of particles. Two beams of particles each traveling roughly the speed of light in opposite directions intersect, and collide. At the points of intersection, detectors collect data resulting from these collisions. Two stages of this process are relevant here:

1. Those conducted online (i.e., when the experiment is running); and,
2. Those that occur offline, using stored data, at some future time.

4.1.1.1 Online

Millions of particles, resulting from millions of collisions contact the detectors each second creating an electronic impulse at the point of impact. The particle now is represented by a bit of data (i.e., a charge that is denoted by 0 or a 1). Accumulating and filtering these bits require both a triggering system and a data-acquisition system. The triggering system is a configurable combination of hardware; firmware and software that act as a filter governing which bits are captured by the data-acquisition system. This system then forms meaningful files from the data stream, viz., groupings of raw data with certain consistent attributions (i.e., headers, footers, and length) that contain such information. Environmental conditions data that give contextual meaning to the captured data stream is collected in a different stream. Both file streams are related by a timestamp, which is essential when reconstructing the data later. The files then are transported to a tape array for long-term storage. The conditions data containing atmospheric data will be combined with the raw data files for calibration, and to add physical context during the offline reconstruction phase.

4.1.1.2 Offline

Taking petabytes of data and turning them into useful data-structures requires having a knowledge of segments of the data stream contained in the raw-data file, such as what conditions and calibrations that put the data into meaningful context matches the raw- data file, along with an understanding of how the data will be used. This process is undertaken with experiment-specific software, physics-specific framework software developed and maintained by physicists, and a very large clustered Linux farm. The key to discovering anything interesting from these data is this processing, from raw segments of data streams to manageable, consistent data structures, that will facilitate sophisticated analysis. Specialized analytic codes then are applied to reconstructed data structures. This process includes track finding, which is the discovering of vertices, or events, that caused a trail of secondary data- points. The analysis also is used to

uncover rare events, patterns, and clusters all of which give insights into what was once a large array of unstructured, disparate data-points.

4.1.2 ATLAS

ATLAS is a detector at the Large Hadron Collider (LHC). In supporting the ATLAS experiment, for many years BNL has generated the largest data flows of the DOE Energy Science Network (ESnet), the fastest science network in the world. The 80 Gbps total bandwidth connectivity to ESnet provides the means to receive, process, and disseminate data transfers at an unprecedented scale. In addition, ESnet has large bandwidth connectivity between BNL and New York City, providing virtually unlimited wavelengths on demand to the ESnet peering point with other major R&E networks. Experimental data repositories for these types of experiments often are denoted by their location, commonly referred to as a tier. Data stored at the experimental site itself is denoted a Tier-0 location, data moved to an offsite location is a Tier 1 location, and so forth. Serving as the "Tier-0" site for RHIC, and the Tier-1 site for USATLAS, a separate ultra-high-speed network enclave supports the explosive bandwidth requirements of the RHIC/ATLAS Computing Facility.

4.2 Computational Science Center (CSC)

As part of the of the BNL Environmental, Biological and Computation Sciences Directorate, the CSC brings together researchers in biology, chemistry, physics, and medicine with applied mathematicians and computer scientists to take advantage of the new opportunities for science made possible by modern computers. The remainder of this section briefly explains a subset of CSC programs that could almost immediately and directly support the IC's approach to analyzing cyber threats.

4.2.1 Graph Analysis Using Super Computers

Dealing with disparate high-dimensional unstructured data sets, researchers using Super Computing Hardware (i.e., IBM's BlueGene L, P, or Q) have the ability immediately to analyze large amounts of data without moving it between multiple nodes. This facility efficiently resolves these data into structures, clustered around a user-defined center point, thus making them ready for analysis.

4.2.2 Network Allocation

Controlling data-transfer through perimeter boundaries is a desirable ability. Researchers identify a means to dynamically reallocate network bandwidth. This reallocation can be based on any of the data's dimension, so allowing for setting priorities or for directional allocation.

4.2.3 Super-Computer Usage

IBM's super computers can monitor and record various states of the machine during its usage. Profiling researchers usage allows administrators to establish a fingerprint unique to the type of job, so helping to identify any misuse.

4.2.4 Social Networking Analysis

With their rising popularity, the field of social networks has become very prominent in research. The CSC' researchers focus their studies on smaller, niche networks. The fact that the network attracts clients affiliated with a particular niche implies that pre-filtering already has taken place. The researches goal is to build metrics establishing hierarchical relationships between data points already know to be associated with each other. This, in turn, can develop a predictive pattern and profile that could be used to anticipate future action. Using uniquely developed algorithms and BNL super computers, sorting through continuous, seemingly erratic data, becomes systematic, efficient, and meaningful in establishing hierarchies, patterns, and predictive predicates.

4.2.5 High-throughput Visualization

Visualization techniques written for the General Processing Unit (GPU) allow the researches to filter through tens of petabytes of data (10^{16}) and reduce noise, so to allow discernable patterns to emerge from a continuous data-flow.

4.3 Probabilistic Risk Assessments

Probabilistic techniques for safety studies of nuclear-power systems use models aimed to predict the likelihood of an event occurring that could trigger a cascade of events that will result in a low-probability outcome, such as an accident. This type of modeling is invaluable in identifying key areas in a complex system that can be watched or improved to prevent the occurrence of an adverse event or a cascading effect.

4.4 BNL/ITD Cyber Security Group

The practical needs of the BNL cyber security group have inspired some innovative approaches for protecting the infrastructure of the unclassified network. Although these are not necessarily advanced scientific techniques, they have advanced conventional technology to the new states needed to keep up with the growing demands of network defense. These advances provided security professionals with ways of rapidly accessing stored network data, employing various criteria as search terms. These terms produce either discrete results, or serve as a pre-filtering mechanism for more detailed analysis. Coupling this "hands-on" the network technology with more abstract advanced-technology modules may create the beginnings of a complex analysis system.

Specifically the BNL cyber security group has two ongoing projects that aid in yielding direct insights into computer usage as well as a methodology to rapidly extract networked information from current collected unstructured data. Both are described below.

4.4.1 Log-Aggregation Reporting Tool

One project resulted in a data aggregation tool that queries data in the log aggregator SPLUNK, i.e., is a software product widely used to make the logs of disparate systems readily available to network professionals. Although default SPLUNK interface is robust, the ITD tool takes BNL-

relevant input, and generates a coherent report describing a networked user's total activity on the network. The scope covers search terms input into search engines, web sites visited and the associated times, frequencies, and classifications (e.g., news, social, entertainment).

4.4.2 HADOOP

BNL currently averages about 2.6 TB/day of formatted, compressed records of all network-flow activity and retains about a year's worth of records. To keep up with requests to search for bad sources, and for forensics analysis of known break-ins, the cyber-security group established a HADOOP cluster. It uses 60 spare nodes that were scheduled for excess, and, at a cost of less than $5000, can search through all of the stored data for a known indicator in less than 11 minutes.

5.0 POISED FOR SOLUTIONS

There are three distinct reasons why BNL is ready to offer prompt solutions to the field of cyber-threat analysis:

1. Programs already engaged in world-class research encompass advanced cyber techniques,
2. An expansive, scalable infrastructure already is in place, and,
3. Close collaborations with other academic- and industrial-institutes offer expanded research opportunities, and innovative insights.

The previous section (Section 4) described programs at BNL that a have nexus to the field of networked computer analysis. The following subsections map these programs to the analysis of networked computers. The reader will be able to identify a possible path to pursue in terms of relevant projects, or even a series of projects all presenting a high probably of success. The following section (6) will suggest explicitly some of the more obvious programs that can be retrofitted easily from existing BNL programs, so casting more fidelity to specific outcomes. Again, other subtle applications may be inferred, leading to more precise outcomes.

As implied, and explicitly mentioned throughout this paper, employing a modular approach to complex problem-solving is paramount when architecting a successful solution. This lies at the foundation of many problem-solving disciplines; in computer science, "divide and conquer" algorithms work for large, complex data sets. When programing, long procedural routines are best broken into functions, or abstracted for reusable classes. In complex experiments, having discrete, reusable steps building towards a larger solution is the hallmark of efficiency and ultimately of success. Reasonably, the same would hold true for understanding nefarious behavior on networked computer. Thus, to "find a spy" or discover a pattern or a rare event, the rigor needed to create the steps in this process either will have to be invented, or borrowed and retrofitted from disciplines that already employ these methods. As any programmer understands, it is preferable to use good code that has been debugged and proven rather than starting from scratch.

Each of the following subsections can be series of projects when debating possible applications to networked computers. Much has been written about their relationship to the fields of study

discussed herein. Therefore, the descriptions given are not intended to illuminate the details of the techniques, but rather to show that BNL has significant world-class expertise in those fields and that can significantly impact our understanding the security of America's networks.

5.1 High- throughput Computing

Sophisticated analysis of big data-sets requires moving, storing, and processing large data sets. The physical act of doing so not trivial; it requires having infrastructure, expertise, and ultimately experience. The BNL computer infrastructure handles tera-bytes of data from various experiments, and serves as a Tier 1 data-repository for the ATLAS experiment. Bandwidth, reliability, redundancy and security all are considered when moving this amount of data either in bulk or during processing. Participation by BNL in its development and, as users, such solutions as the Open Science Grid, as well as 'home- grown' solutions for local throughput has given the Laboratory the ability to conduct research in areas once constrained by the amount of data. Prima-fascie, the amount of networking might seem trivial compared to that of the data needed to conduct physics experiment, but in terms of producing a system with the goal of effectively analyzing these data, both throughput and infrastructure quickly become constraining factors.

5.1.1 Triggering Systems

Filtering data based on some criteria is most effective and efficient the closer it is done to the source of the data. The efficiency of using a configurable front-end near to the genesis of the data makes the packaging of it very effective. Networks are flooded with data that could effectively be grouped into various subcategories. A hardware solution would separate the data without impacting the performance of the enterprise.

5.1.2 Science Demilitarized Zones (DMZs)

Networked Demilitarized Zone (DMZs), viz., a small neutral firewall for protecting the infrastructure of a company's computers, are effective ways to sequester incoming data to protect an infrastructure. This paradigm becomes challenging when there is a need for interactions and data flow between the DMZ and the insolated network, especially when the flow of information is large. In 2012, the USATLAS Tier-1 sent and received on average of 60 and 52 TB of data per day, with traffic peaking in both directions at a daily average of ~2 GBps. To support this growth, the size of the BNL WAN connection has grown over a thousand fold, with a greater acceleration of growth expected. In anticipating the near- to mid-term bandwidth requirements of other major collaborations (e.g., the NSLS-II, CFN) have deployed a 100-GBPS Science DMZ to provide an environment capable of supporting the expected unprecedented demands for bandwidth. On the BNL campus, all scientific buildings except the Medical building, are connected to a high availability, dually redundant, core network at a minimum of 1 Gbps (the NSLS-II and ISB are at 10 Gbps), with each an individual switch port on campus configurable up to 1 Gbps.

5.1.3 File Creation and Data Encapsulation

The physics experiments at BNL effectively employed methods of capturing, encapsulating, storing, moving, reconstructing, and analyzing large data streams. These techniques can be used

on the network's data streams in preparing segments and various groups for sophisticated analysis. Achieving analysis-ready data, of any type, requires establishing a multi-step process. Raw-data files of the complete data flow (or a triggered subset) are created over a fixed period of time. Calibration data, which moves a much slower rate than data in general, is continually recorded and stored with a comparable time-stamp. These files are stored in mass storage array, and eventually are put through a code sequence that merges with the calibration data, and constructs data structures that are accessible through a common framework containing built-in methods serving the needs of the analysis, and has an understanding of the data structures constructed by the reconstruction code.

5.2 Analysis - Patterns, System states, and Rare Events

The goal of the analysis is to understand the phenomenology of the debris resulting from time of collision. Thus, there is a process of working backwards in time from the recorded tracks of data.

Latent variables versus observable variables differ, but the underlying goal of using a linear-quadratic estimation is to establish a system state with a system rejecting noisy data is essential. The ultimate goal of data filtering, reconstructing with conditions, and with calibrations data is to effectively use advanced techniques found in software frameworks to identify find some significance in the data stream.

5.3 Predicative Probabilistic Modeling

The use of Probabilistic Risk Assessments, as designed for the nuclear power industry, can be adopted to look at the insider threat. Two approaches which are effective in mitigating insider threats, sociological and technical. A system can be analyzed with regard to its risk to a potential insider, and then precautionary maneuvers could be taken to remove those risks. In essence, the event being modeled is not an accidental human error or system failure, but intentional sabotage or infiltration. Hence, some input variables may be somewhat more complex to integrate into the model; however, the core of the model remains, that is, how will the system be affected by such an event. Accordingly, a quantified risk factor can be placed on each part of the system to determine whether an acceptable risk was achieved.

5.4 Cluster Finding

Communications within a culture can lead to information about the organization of that culture. The social networking work being done by the CSC coupled with the ability at BNL to create structure from unstructured disparate data fits into the social aspect of cultural cluster. Also, the work in physics that looks for clusters of particles can be applied to social strata to fine-tune clusters and identify vertices of an event.

5.5 Network Security

Subjecting network streams to various filtering and grouping techniques, based on any number of criteria, followed by graph analysis that can be performed on the BNL super computers, will generate a high-resolution picture and the organization of network traffic. Benchmarks can be

developed and anomalies are detected more easily in organized, segmented, and sorted pools of data.

5.6 Super Computing

Even with the sharp decline in price and the increase of power of consumer grade computers, super computers are a still a rare- and sought after-commodity. Furthermore, if this need somehow is connected to a nation's intelligence strategy, then stealing CPU cycles may likely become a part of their clandestine posture, as it likely will for non-nation states that have a need for supercomputing.

Rigorous procedures are in place for user verification and authentication to ensure that the users of these computers are restricted to those who can demonstrate a legitimate need for them, and are a part of a "known" organization. However, once an account is created, there is no simple way to know if the person granted the account is the person using it.

6.0 SAMPLE PROJECTS

This section contains some "back-of-the-envelope" ideas of what projects BNL could begin almost immediately. Each would require minimal additional resources, and each promises a high probability of success. This section does not offer actual proposals; however, the ideas expounded might be the seeds of a proposal or the genesis for ideas leading to one. Even though they are conceptual, some manifestation of completion for any of them would have tremendous positive impact on understanding the components of our nation's networks.

6.1 Data Encapsulation

In keeping with a modular approach to creating files of raw data that define some subset of network data, there are several problems. Capturing data from full packets often results in ones that are too large to keep for long times; further, encryption is a problem. However, a possible solution lies in having a set of distinct files that could be processed later, provided that their existence was meaningful, In contrast to capturing stored packets, headers or log files, of some aggregation of all these might be applied. Creative groupings, bound by a time stamp, would support coordination with other conditional events that may be included in the analysis. At present, network data is too unstructured to approach specific analysis in an efficient, meaningful way.

6.2 Framework

Although there are many network data-analysis tools, few, if any, give an analyst the ability to subject a data set to the advanced statistical and scientific methods needed, such as for example, to find a rare event based on tracking of events. Disparate- and voluminous- data resulting from a multitude of events must be sorted and organized. Advanced graphing techniques coupled with techniques for finding physics events could be grouped into an analysis framework. This framework could be used to establish a standard dataset which facilitates the ability to study and understand the intent of adversaries. The ultimate goal is to uncover attribution by discovering

identifiable patterns. Adding to such analysis, probabilistic methods and their outcome could be significant and far-reaching.

Several discrete steps are needed for this type of analysis, much of which has been written in a modular format so it can be reused. The idea of a software framework is that is takes blocks of code and brings them together for analysis. The differences in the type of event and type of needed action are only the boundary layers of a framework, the input and output. The internal workings are the statistical methods and scientific preparations used to ensure rigor in uncovering the hidden nature of the event. In short, much of the framework used for particle physics could be adapted and applied to analyzing network traffic.

6.3 Unstructured Data Analysis

As mentioned in Section 4.2.3, Super Computers could be used to build data-structures based on a user-defined dimension without moving the data. This allows efficient building of these structures, which, in turn, support iterative tearing down and rebuilding under different dimensions. Creating sorted groupings of networked data is an effective step toward discovering and understanding the nuances hidden in noisy network traffic.

6.4 Insider Threat – Complex System Analysis

Most analyses of insider threat focus on the social- and psychological-disposition of the insiders. Personnel records, behavior, and verbal indications all are valuable trip-wires in forecasting a potential threat. Unless sabotage is the goal of an insider, several more dimensions are needed for an insider to pose a threat. Placement in proximity to sensitive information on a system is one of them, and that can be assessed using system analysis in conjunction with probabilistic risk assessments.

6.5 Data Diversion

As represented by the BNL science DMZ, data can be directed using a fast real-time filtering technique. Other methods at BNL that can be looked into for directing network traffic, i.e., the RHIC triggering system and the CDC network diversion algorithm. It would be interesting to be able to redirect incoming attacks or probes without allowing them into our network and without letting them know they been diverted. Currently, such attacks or probes are blocked, but deterring them to a “safe area” would allow researchers and security professionals to study potentially malevolent traffic. Advanced techniques could allow this to happen without impact to Enterprise performance.

6.6 Cultural Clustering

Adding cluster-finding techniques to the techniques of filtering, grouping, and sorting described above would allow secondary grouping, or clustering of individuals. Graphing techniques could then be applied to the network data of that cluster to discover social strata unique to the cluster.

6.7 Super-Computing Usage Analysis

To prevent unauthorized use of our nation's super computers, benchmarking techniques could be developed along with an automated system alerting an administrator that there has been a deviation from the expected usage of the machine.

7.0 CONCLUSIONS

- National Laboratories have the ability and to assist in efforts to protect our nation's computer networks.
- Based on its unique research and experiments, BNL can add significant knowledge and address specific problems, issues, and questions that resonate in the world of national security.
- The relevant capabilities for BNL experiments are mature with respect to application to the IC. Retrofitting these techniques unique to BNL to fulfill national security needs can directed to small, modulized projects.
- These modular projects have a high probability of success and broad scope of usefulness across a wide range of national-security problems.
- An established infrastructure tailored to this type of work is an essential component for a project's success. This already is in place at BNL.
- Opportunities exist for collaboration with academic and industrial research partners.

BNL deals with huge volumes of data each day. Generating and harnessing these data, fuels the experiments that provide groundbreaking, world-class discoveries. Controlling and directing data that is generated elsewhere and sent though our networks may afford answers to many questions and mitigate the threats we find on our networks. Retrofitting and sharing techniques developed at Brookhaven add a vital piece in our nation's efforts to sustain the integrity of our intellectual-, social-, and economic- infrastructure.