DASI

ANNUAL REPORT
2015

DASI

Mission

The Distributed Analytics and Security Institute (DASI) is part of the High Performance Computing Collaboratory at Mississippi State University. DASI is dedicated to excellence in the areas of distributed computing, big data analytics, cyber security and critical infrastructure protection. The primary focus of the DASI is to coordinate, facilitate and expand research activities across academic and research units involved in the application of high performance distributed computing platforms to the areas of big data analytics, cyber security and critical infrastructure protection. DASI also provides science-based strategies aimed at increasing our ability to process large volumes of data in a highly secure way over a secure infrastructure.

Vision

Develop capabilities for Mississippi and the U.S. in high performance computing that include three problem areas of big data analytics, cyber security and critical infrastructure protection.
Research and assess the ability to do the above activities over a geographically distributed platform.

MISSISSIPPI STATE UNIVERSITY™
DISTRIBUTED ANALYTICS
AND SECURITY INSTITUTE

Contents

Letter From The Director

Dear Colleagues and Friends,

This annual report represents the accomplishments of the Distributed Analytics and Security Institute (DASI) at Mississippi State University during its first full year in existence. DASI was created in mid-2014 to take on a major research role in big data analytics and cyber security. This report highlights current projects, accomplishments, and a diverse research faculty and staff doing work to further efforts in cyber security and big-data analytics through the use of high performance computing.

One of the great technical challenges of today is the vast amount of data generated in our society. Analysis of this data is resource intensive and requires forward-looking science and high performance computing to accomplish. At DASI, we are analyzing many types of data in a variety of ways. Data types being analyzed include network flow, malware, multi-spectral, and social media data. Research methods include analysis of malware, visual analytics, autonomic computing, and network analysis.

Cyber security is a key component of national security. DASI is working to solve today's cyber security challenges as well as developing innovative solutions to prepare for tomorrow's cyber defense. As you will see in this report, we have projects related to control system security, network security, malware analysis, visualization of cyber data, and many others. Other efforts in the works relate to security of embedded systems, like those used in both surface and unmanned aerial vehicles.

There are many exciting things happening at DASI, and we are preparing for growth in the future.

All the Best,
Dave Dampier

Anomaly/Event Detection in High-Velocity Streaming Data

Researcher(s): Somya D.Mohanty and Mahalingam Ramkumar

CHALLENGE

With an ever-increasing necessity of smarter threat detection systems in cyber security, a data-driven approach towards pre-cognition of events/anomalies can provide time-critical reactivity to counter complex and real-time threats. In such high throughput systems, traditional batch processing approaches are insufficient to provide real-time insights into the data. Such high-velocity data streams require Big-Data based parallelization strategies for detecting trends. Architectures modeled after such an approach, can leverage the predictive capabilities to build smarter threat/intrusion detection systems with a proactive approach towards security of systems.

CURRENT PRACTICE

Traditional models of trend detection are applicable to a wide variety of security systems such as intrusion detection, security event management, and threat detection systems. However, most of them are based on parametric statistical models in which the single feature of bursty traffic is compared to a pre-determined threshold. Traditionally Poisson processes are used to establish a baseline Poisson rate for a sliding window, and any deviation from the baseline is categorized as an event. Another significant portion of approaches rely on signature-based event-detection, where a previously identified event pattern is matched with a new pattern to report occurrence of an event.

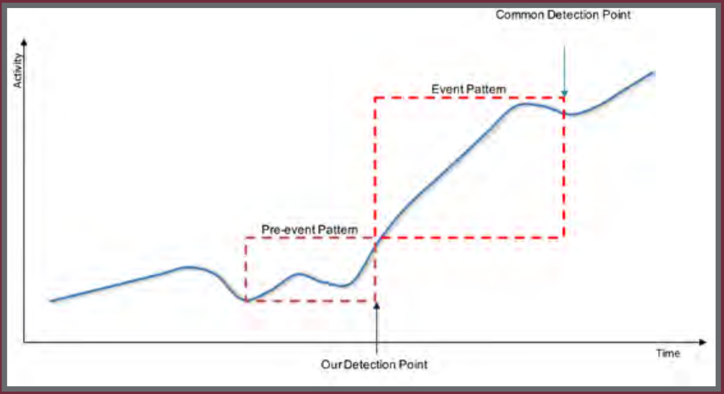
The weakness in both the methods is the inability to identify events before they become a trend. With the onset of high throughput systems generating large volumes of high velocity data, such traditional approaches cannot be used for reacting (enforcing necessary security policies or counter measures) to cyber attacks in real-time.

TECHNICAL APPROACH

The research aims at developing real-time multi-dimensional event/anomaly prediction systems based on pre-event data-trained machine learning models. More specifically, the proposed research plans to create a highly scalable supervised machine-learning models based on the training of pre-event patterns leading up to an event. The training set for the models will be extracted from the test datasets (described later) and stored with a binary classification of: 1) non-event patterns, and 2) event patterns. An occurrence of a new pattern from the streaming data can then be compared with the previously classified patterns, and distance vectors from each of the older patterns will be computed.

> Developing highly parallelizable architectures for ananomaly/event prediction in high-velocity streaming

Figure 1: Anomaly/Event detection in current state-of-art vs the proposed research approach

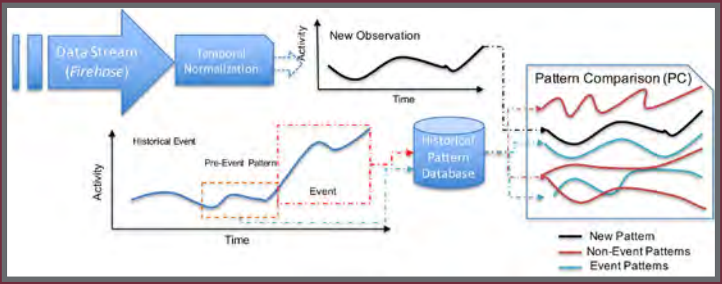


These distance vectors (Euclidean) will serve as the input to the machine-learning model for codification/classification of the class to which the observed pre-event pattern belongs to. Using the classification, the model will then be able to predict the upcoming trend (event/non-event, type/pattern of event) of the new observation.

Towards the overarching goal of the project, the research challenges that need to be addressed are:

- Identification of features/dimensions in traffic for extraction, normalization and prediction
- Identification of machine-learning models for high-dimensional data
- Parallelization and scaling strategies using High Performance computing at Mississippi State University (HPC²) test bed
- Evaluation of online learning schemes for dynamic models
- Describing a generic pipeline for universal applicability

Figure 2: Data-driven approach towards anomaly/event prediction



IMPACT

The research envisions the multi-dimensional non-parametric models can be applied to gain critical insights in the trends of other large-scale streaming data sources. For example, anomalies in sensor systems, web-servers, honeypots, packet routers, etc. can be tracked to determine increases in potential hacking activity. Online-learning models can be utilized to create models that can learn in real-time instead of relying on sets of pre-trained features. Applicability of the prediction model to develop smarter intrusion/threat detection systems can also be evaluated. Other application areas include implementation of game theory algorithms on top of the prediction models to develop self-adjusting intelligent decision engines (for example, traffic congestion, load-balancing severs, etc.). Using a scalable/distributed machine learning approach will also help researchers understand and establish paradigms for the applicability of disparate machine learning algorithms in large-scale streaming data.

Autonomic Security Management Framework for High Performance Computing Systems

Researcher(s): Sherif Abdelwahed and David Haglin

CHALLENGE

Securing data and applications in high-performance computing (HPC) systems is challenging, particularly due to the complex and large-scale nature of such systems, the open operational environment needed to support external access, and the variety of network protocols and network interfaces that characterize HPC infrastructure. All of these factors can introduce the potential for illicit cyber penetration. The issue at the heart of HPC security concerns, however, is the potential for malicious users in multi-tenancy environments and in the sharing of resource pools on the same physical platforms. Intruders can exploit massive resources in order to mount sophisticated and damaging attacks, gaining access to critical data and applications.

In order to address these concerns and to meet the challenge of ensuring continuously available and trustworthy HPC resources we are developing of a model-based autonomic security management framework for HPC systems that integrates system control, security analysis, and auto-response mechanisms into a model-based management framework to automatically identify and mitigate potential security intrusions and maintain a functional system.

CURRENT PRACTICE

Traditional detection techniques have addressed a portion of system attacks, but have not provided effective techniques to protect against application attacks. Current detection methods adopt predominantly reactive approaches, using signature-based and/or anomaly-based detection. They are typically designed in an ad hoc manner and only for specific system or operating conditions. These approaches are also labor-intensive and challenging to manage. In addition, with the exponential growth in the volume and sophistication of cyberattacks, it is no longer possible to track the signature of new intrusions. Thus a new innovative solution is required to solve this dilemma. In this project we are developing autonomic computing based approach to make HPC systems self-protected with minimum or no involvement from system engineers or administrators.

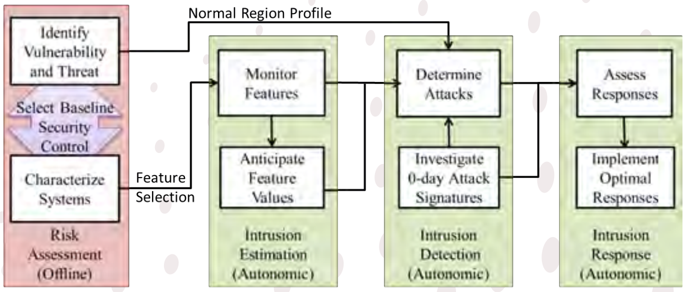
TECHNICAL APPROACH

This research develops an autonomic security management system for HPC systems by extending our current technologies including the predictive performance management system and our earlier work on security management that have been developed as part of our PNNL funded project to:

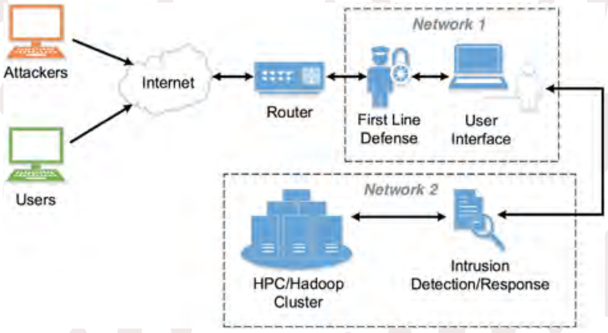
Utilizing model-based techniques and tools to develop effective and proactive security management structure for high-performance computing systems.

- Develop efficient algorithms and mechanisms for accurately identifying anomalous events triggered by attacks and malicious access.
- Develop stochastic models for system security levels based on monitored system/network performance parameters/events and provide the ability to accurately characterize current state and perform risk and impact analysis of potential attacks in real-time.
- Develop proactive mechanisms to detect and characterize cyber-attacks and deploy autonomic responses to disrupt and mitigate the impacts of such attacks on the system, and recover the system back to normal operation.
- Utilize the existing MSU and the Idaho Bailiff HPC testbeds to conduct the proposed research and demonstrate pre-deployment capabilities.

We are producing monitoring and behavior analysis tools to collect information about the current operational states of HPC systems and their applications, a risk assessment model to evaluate the overall vulnerability of the HPC system, attack prediction and early warning systems, and ultimately, autonomic control and security management structure for HPC systems. This security management structure provides an optimal plan of action that includes a sequence of control responses to mitigate and protect against cyber-attacks while maintaining the resilience of the underlying cloud infrastructure.


Identify Vulnerability and Threat
Select Baseline Security Control
Characterize Systems
Risk Assessment (Offline)
Normal Region Profile
Feature Selection
Monitor Features
Anticipate Feature Values
Intrusion Estimation (Autonomic)
Determine Attacks
Investigate 0-day Attack Signatures
Intrusion Detection (Autonomic)
Assess Responses
Implement Optimal Responses
Intrusion Response (Autonomic)

A model for self-protection system a coordinated processes of real-time monitoring, data processing and analysis, intrusion detection and classification, and automated system protection.



Enhanced security for an experimental HPC cluster testbed with two demilitarized zones (DMZs), Network 1 and Network 2. The first-line of defense (e.g. firewall) are installed in Network 1 to protect from known attacks. External requests pass the first-line of defense are sent to Network 2.

IMPACT

The set of technologies — theory, models, and algorithms — produced as part of this work will convert a significant number of intrusion detection and incident response tasks into systematic and semi-automated processes using concrete mathematical models and proven reasoning and optimization techniques. The project is anticipated to produce:

- Monitoring and behavior analysis tools to collect information about the current operational states of HPC systems and their applications,
- Risk assessment model to evaluate the overall vulnerability of the HPC system.
- Attack prediction and early warning system.
- Autonomic control and security management structure for HPC systems.

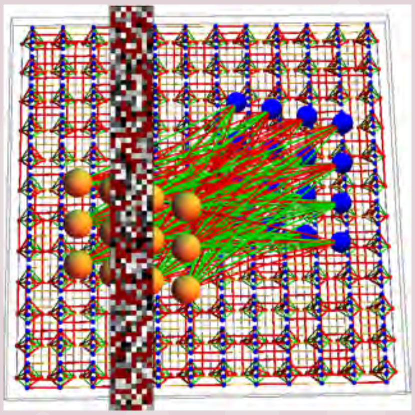
By proactively detecting attacks and responding to them at an early stage we will be able to significantly mitigate their impacts and provide effective mechanisms for system recovery. As such, the proposed research has the potential to make a major improvement in the effective protection of critical data and applications hosted in HPC systems.

Boltzmann Machines

Researcher(s): Mark A. Novotny and Yaroslav Koshka

CHALLENGE

In our connected world there are a number of bad actor who share the Internet with all users. Computers and other connected devices, including the ever-expanding Internet of Things, are wired through technology that make up the network backbone and network connections. All traffic passes along this network. It would be extremely advantageous to analyze the flow of network traffic, and even more advantageous to be able to classify the traffic over a particular network as normal traffic or as traffic from a bad actor. However, this type of network traffic classification is extremely difficult. One difficulty is network traffic at 100 Gbits/s or greater would need to be analyzed in real-time. Another difficulty is that as one detection and classification method becomes successful at removing traffic from bad actors the bad actors change their strategy.



A schematic of the theme of the development effort. Network traffic is the top picture representing the flow of bits in a network, here the colors are hexadecimal. A Boltzmann machine with 16 visible units (yellow) and 16 hidden units (blue) is the middle schematic, with colored connections being positive (green) or negative (red). The bottom schematic shows the Chimera lattice of a 1152-qubit D-Wave 2X. The D-Wave quantum computer plus high-performance classical computing provides the compute power to allow machine learning to be applied to cybersecurity.

CURRENT PRACTICE

The research community is converging on the paradigm that network traffic classification can only be successful by using machine-learning techniques. One of the main emerging machine-learning techniques is Boltzmann machines. A Boltzmann machine is a stochastic neural network wherein some units (visible units) are set to the bits of the data being sampled and some units (hidden units) are free to equilibrate to a stochastic set of configurations. A Boltzmann machine is a stochastic neural network wherein some units (visible units) are set to the bits of the data being sampled and some units (hidden units) are free to equilibrate to a stochastic set of configurations.

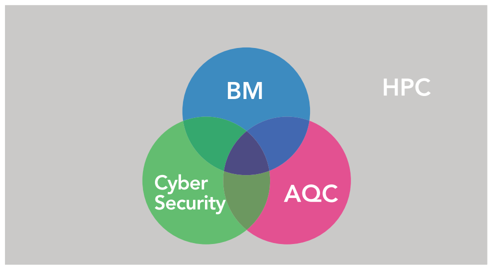
Use Boltzmann machines, as implemented on an anadiabatic quantum computer and on a massively parallel classical supercomputer, to analyze and classify network traffic in real time

The main difficulty is to determine the set of parameters for the Boltzmann machine that optimally gives the classification of the data. In other words, the way the machine learns needs to be efficient. One type of connection, called the Restricted Boltzmann machine (RBM) allows only connections of the visible-to-hidden type. A learning algorithm for the RBM can be written in a fashion such that an adequate way of approximating the required calculations for machine-learning of a RBM is possible. RBMs have been used in many types of environments that require machine-learning, including classification of network traffic.

We are at the dawn of a possible revolution in computers -- quantum computers may soon transition from laboratories to having a major effect on what calculations are feasible. Quantum computers rely on quantum principles, particularly quantum entanglement and quantum tunneling, to perform calculations that are impossible on classical computers. In technical jargon, classical computers can perform calculations in a class of problems labeled P, while quantum computers can perform calculations in a much larger class of problems labeled NP. Everyone assumes that P is contained in NP, but that NP is different from P. Quantum computers may change society as much in the current century as the wide availability of classical computers did in the last century. Currently one commercial quantum computer company in the world, D-Wave Systems. D-Wave makes a special-purpose type of analogue machine that is called an adiabatic quantum computer (AQC). An AQC is designed to solve one type of NP problem. This type of NP problem should have consequences for Boltzmann machines and hence for network traffic classification.

TECHNICAL APPROACH

The technical approach is to utilize Boltzmann machines, as implemented on a D-Wave AQC and on a massively parallel classical supercomputer, to analyze and classify network traffic in real-time. This involves efficient parallel programming of the classical supercomputer, devising and implementing novel algorithms made possible by availability of a D-Wave AQC, and applying the Boltzmann machine to network traffic classification and analysis. Mathematically our technical approach can be written as the intersection of four fields: Boltzmann Machines ∩ Cybersecurity ∩ High-Performance Parallel Computing ∩ Quantum Computing.



A Venn diagram of the task. The goal is to enhance cybersecurity (green). Machine learning methodologies are utilized, specifically Boltzmann Machines (BM) (blue). To enable solutions of hard problems, Computations are performed in a HPC (High Performance Computing) environment (gray). Calculations that are hard for HPC are enabled by AQC (Adiabatic Quantum Computers). The approach is at the intersection of these areas (purple).

IMPACT

Quantum computers hold the promise both of making all current computer and network security obsolete and of enabling heretofore impossible unbreakable security measures. Using an AQC to perform solutions of NP problems, and coupling these to machine learning may enhance cybersecurity related to network traffic. The impact would be in making our increasingly connected world more secure.

Cyber Phylogeny

Researcher(s): Wesley McGrew and Dae Glendowne

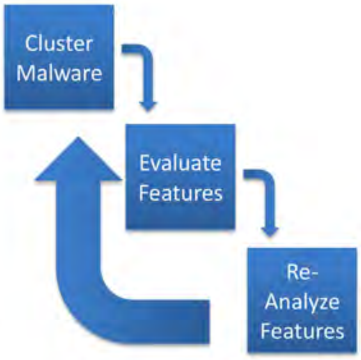
CHALLENGE

In the current environment of advanced nation-state and highly organized criminal actors creating malicious software, attribution is a problem that must be addressed in order to respond to these threats. We recognize the need to improve methods for identifying and attributing attackers' intent on disrupting network operations. Although some attackers use hands-on techniques, most successful attacks use at least some automation in the form of malware, to facilitate the invasions. We need to better understand this malware in order to best prevent future attacks.

CURRENT PRACTICE

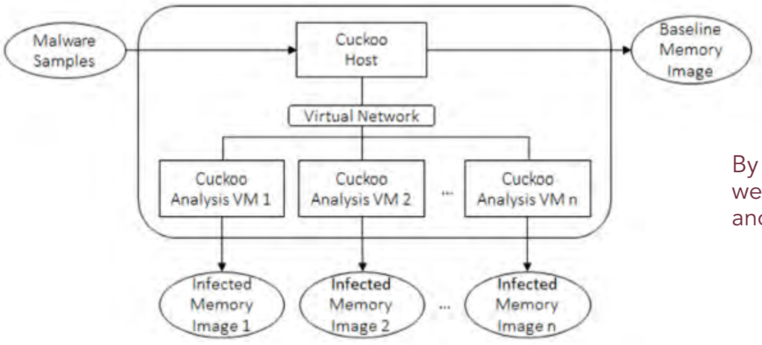
In Year 1, we have developed a procedure for gathering the dynamic data needed to perform machine learning on malware samples. In this procedure, memory images are created at runtime for large numbers of samples. These memory images are then processed to extract the needed data. In Year 1, this data included indicators of DLL injection, which was used with existing machine learning algorithms to classify malicious DLLs in memory images.

Throughout Year 1 we've continued to look into new bio-based technologies that can be applied to our malware characterization and/or detection. We've discovered several new methods that could be applicable to getting "better" results and have started to incorporate some of the "easier" ones such as using HMMR (Hidden Markov Models) for motif generation. We've also discovered some new mathematical and statistical methods that can be used to optimize our process in both speed and precision.



Iterative process for identifying features for machine-learning clustering

Developing a phylogeny of malware samples, characterizing their lineage, and informing attribution



By generating memory image data for feature extraction, we can identify relationships between malware samples and build a "family tree."

TECHNICAL APPROACH

We are identifying features of malicious software that can be used by machine learning algorithms to help identify the presence of malicious software in the memory images of sandbox environments, and classify that malware based on the feature set. With a collection of these features, relationships between malware samples can be identified and a "family tree" may be built. We are performing the collection of data, identification of features, and application of that information to new samples in a high-performance computing environment.

A Boltzmann machine is a stochastic neural network wherein some units (visible units) are set to the bits of the data being sampled and some units (hidden units) are free to equilibrate to a stochastic set of configurations.

Overall, our approach for Year 2 revolves around the cooperation between teams at MSU and PNNL and integrating the techniques developed at both while still moving forward. We intend to explore and implement options for utilizing the output of the PNNL techniques as features for classification in the MSU process, as well as use the MSU dynamic data capture as a potential way to bypass issues with packed and obfuscated code.

IMPACT

With the ability to leverage high-performance computing resources, analysts will be able to take current sources of potential new malware samples and quickly classify those samples by potential authorship or common source. New malware samples may be grouped with existing samples that have a high degree of similarity, giving analysts more information on specific threat actors. This benefits those who wish to attribute cyberattacks by giving them more data that can be examined for operational mistakes by the attackers. This raises the likelihood of finding the source of an attack or widespread malware infection.

Biosequence analysis will enhance the malware analysis process in operational environments by organizing the space of malware into a rigorous categorical structure. This has two immediate benefits. First, it will significantly reduce the computation required to recognize malware because the family tree allows us to represent many individuals using a single family recognizer. Second, because it is tolerant to variation, it introduces the possibility to catch previously unseen attacks simply by their similarity to known attacks.

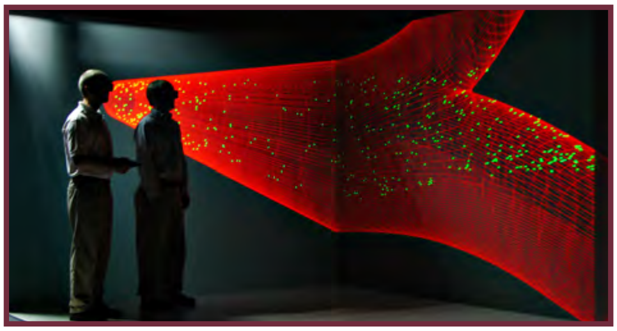
Exploring 3D Immersive Visualization for Malware Analysis

Researcher(s): Jean Mohammadi-Aragh and Derek Irby

CHALLENGE

In our highly digitized world, protecting data stability and security is crucial. Protection entails understanding and preventing potential malware threats that could compromise sensitive and confidential information.

The productivity of malware analysts is limited by screen space and lack of visual support to help identify and track patterns. Viewing malware on a two-dimensional monitor is much like looking at the sky through a set of binoculars—one can only see a small section at any given moment. However, bigger displays and more pixels will not solve all issues. To understand and reverse-engineer malware, it is important to analyze the code in a context-rich environment including the code itself, the code block structure, memory accesses, cross references, and more. This project addresses current malware analysis issues by investigating different ways of integrating contextual information into a single three-dimensional (3D) display for improved malware analysis.



This 3D visualization system interacts with a disassembler/code analyzer that can be displayed in multiple virtual realitysystems like the CAVE environment.

CURRENT PRACTICE

Analysts examine malware by converting binaries into human-readable assembly language and stepping through code using programs such as IDA Pro and Radare. While graphical interfaces do exist, limited screen space and visual tools place large cognitive demands on analysts. Current methods of analyzing and navigating disassembled code of complex cyber data can be overwhelming to an analyst who identify patterns, operations, and functions within the disassembled code while keeping track of patterns, operations, and functions that cannot be displayed simultaneously.

Developing 3D immersive and interactive visualization system for malware analysis

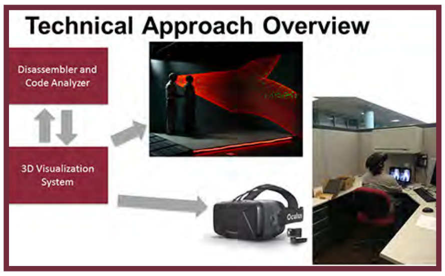
TECHNICAL APPROACH

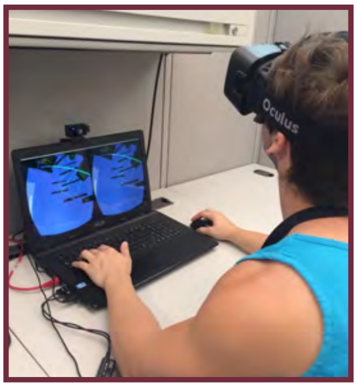
We leverage 3D immersive visualization to address challenges of analyzing and navigating disassembled code of complex cyber data. We hypothesize that utilizing 3D immersive displays will 1) diminish the screen real estate issue, 2) help users find information quickly and accurately within a context-rich environment, and 3) support analysts' efforts of grouping and classifying malware. Our approach focuses on six subtasks: data preparation, multi-level exploration, comparative visualization, interaction techniques, immersive visualization hardware, and evaluation.

Initial development and evaluation of the malware visual analysis system will occur in the Mississippi State High Performance Computing Collaboratory Virtual Environment for Real-Time Exploration (VERTEX), because of its support for collaborative-viewing of immersive visualizations. The VERTEX is a cave automatic virtual environment- (CAVE)-like display that tracks user movement and is controlled by a handheld device. While the collaborative nature of the VERTEX makes it suitable as a development and training platform, cost and size make it less than ideal for wide-scale deployment to individual malware analysts. For deployment of the immersive static malware analysis system, we are exploring the Oculus Rift, an affordable wearable display. Both displays have a large field of view that supports the incorporation of more accessible data.

In a large-scale system with multiple data sources, developing techniques to interact with and navigate through the data is necessary. This effort will use tracking devices for the development of 3D gestures that can be used in concert with analog and digital controls to explore malware data in an immersive 3D environment.

To gauge the performance of the 3D visualization, evaluations will be conducted with expert users. We aim to answer three questions in these evaluations: 1) Does 3D immersive display improve the human understanding of the malware data? 2) Do interactions including the tracking and the 3D gestures facilitate understanding and collaboration? 3) How effectively can the users gain insight from their malware data with the 3D immersive display? Alternative malware analysis methods will be compared to our 3D immersive visualization system.





Our system also can be used and displayed using the Oculus goggles on a Desktop platform.

IMPACT

Our proposed systems will address issues surrounding visualizing the large amount of code per malware sample, and ultimately, visualizing the larger phylogeny of multiple samples. The development of this system complements the work already underway in the Cyber Phylogeny MSU/PNNL project led by Dr. Wesley McGrew, and will provide part of the interface that analysts will use to work hands-on with the Cyber Phylogeny results and data. Additionally, similar to our success developing practical tools and training in other domains, we anticipate the immersive 3D malware analysis system will serve as a platform for training new analysts.

Geo-Inspired Parallel Simulation (GiPC)

Researcher(s): Michael S. Mazzola,
David Haglin, and Tomasz Haupt

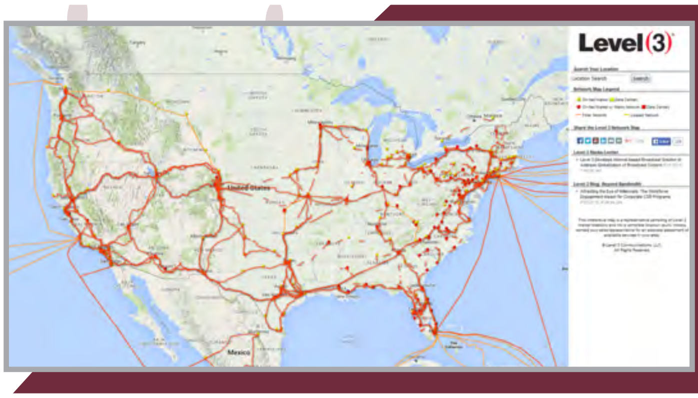
CHALLENGE

It is known in the field of power systems simulation that problems can be positioned (partitioned) in ways that take advantage of the geographical and physical properties of the system. However, partitioning for fast execution has many degrees of freedom in very large and complex system simulation.

CURRENT PRACTICE

There are many examples of partitioning large graph-based problems, and a base of proven and industry-accepted software tools for doing so. One such tool is METIS, and a variation known as hMETIS, which is a hypergraph partitioning tool popular in the integrated circuit design community. The output of hMETIS is determined by the weights placed on the unpartitioned graph. The specification of this weighting is ultimately tied to the speed of execution realized in the parallel-computed solution. In complex graphs, it is difficult to find partitions that optimally use the properties of the cluster computer. This issue is linked to the nature of the problem being solved. For example, improvement was shown in computational performance when a custom partitioning scheme influenced by the geographic properties of a power transmission system was used as compared to a hMETIS partition without consideration of these factors[1]. This demonstrates that performance improvement is available to partitions posed with the right link between cluster computer properties and the properties of the problem being solved.

The problem with existing methods is that there is no means for automatically finding partitions of networks that optimize the similarities of the physical problem with those of the cluster computer. For example, fast dynamics need shorter simulation time steps to converge, whereas slower dynamics can get by with longer time steps. A multi-objective optimization process is needed that can discover appropriate solutions for partitioning the problem in each case required.



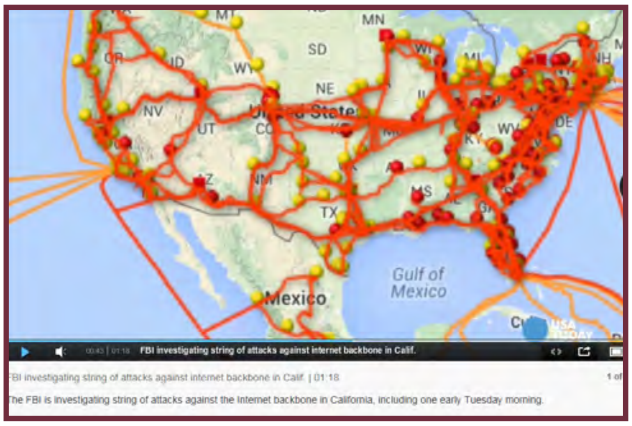
Geographic properties of networks serving human populations parallel the construction and efficient use of cluster computers.

Develop solutions for partitioning and solving graph-based physical simulation problems using geographic information about the physical network

TECHNICAL APPROACH

The Geo-inspired parallel computing (GiPC) software demonstrated in the second year of this task is designed to discover with a high degree of automatic operation good solutions to the partitioning problem using the similarities between the geographic and physical properties of a real system and the design and performance of a modern cluster computer. The case example is an auto-generated graph consisting of the edges and nodes of subsystems representing smaller physical networks connected together to form a final graph of millions, hundreds of millions, or potentially billions of nodes in size.

GiPC is a multi-objective optimization engine that uses hMETIS to create candidate partitions and then uses open-source software and a simulation manager framework developed by MSU students to test the partitions to compare performance against objectives. New partitions are created to improve observed performance. The framework is intended to use the Graph Engine for Multi-threaded Systems (GEMS) to search for nodes and edges with geo-tags that will produce candidate partitions with attractive properties, resulting in feasible convergence times for the optimization engine on problems differing widely in graph size. The analytical target cascading (ATC) method excels at cascading the optimization problem into sub-problems that run in parallel, thus giving the optimization process a chance to solve ever larger graph partitioning problems with larger allocations of cluster computer resources, keeping the execution time tolerable while scaling the size of the graph. An application of the software will be demonstrated with a use-case tied to the real-time simulation of cyber-attacks on the power grid.

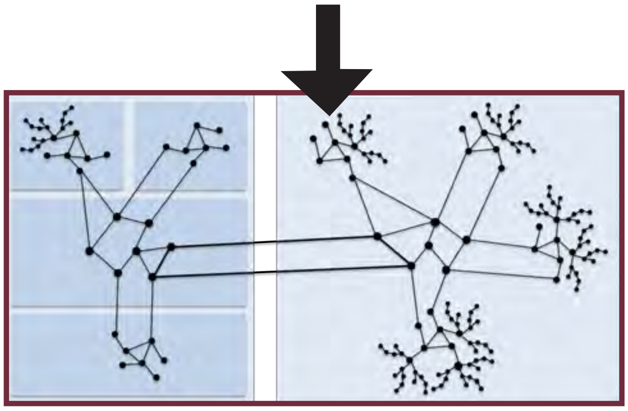


FBI investigating string of attacks against internet backbone in Calif.

FBI investigating string of attacks against internet backbone in Calif. | 01:18

1 of

The FBI is investigating string of attacks against the Internet backbone in California, including one early Tuesday morning.



The Internet backbone can be represented as a graph, where each relay can be seen as a node (vertex) and each branch line can be seen as a connection (edge).

IMPACT

Finding good solutions to very large and complex problems often appears impossible, and is impossible by manual means. Modern multi-objective optimization theory takes such impossible tasks and makes them tractable. GiPC relies on generalized techniques developed for parallel computing. The result is confidence that when the process is complete, a partition exhibiting good computational performance has been found that is better than the alternatives. The use of geo-inspired processes is a key to finding these alternatives with a reasonable expenditure of high-performance computing resources.

Graph Walk Optimization

Researcher(s): Ioana Banicescu and Nitin Sukhija

CHALLENGE

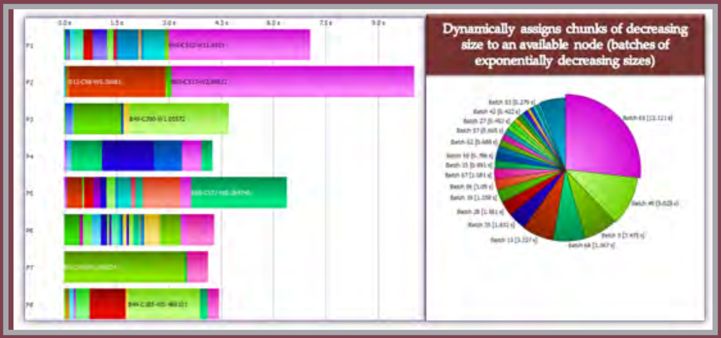
Data collection and analysis is rapidly changing the way scientific, national security, and business communities operate. Data analytics applications, specifically those involving graph analytics, have received increasing attention over the last several years. The performance of these applications is essential, even sometimes critical, to achieve the objectives proposed by the domain areas making use of them.

The complexity of working with applications involving big data requires computing in parallel and distributed environments. Scheduling big data computations on parallel non-dedicated heterogeneous systems, where the computing resources may differ in availability and reliability, is a challenging task requiring resilient scheduling methods for an efficient execution.

CURRENT PRACTICE

Many research efforts have been attempted to optimize performance. These optimizations include improving performance (per core), increasing scalability of their execution in parallel and distributed environments, and dealing with dynamically changing large data sets. Moreover, the sharp increase in number of computing units is likely to continue, which translates in expected growth in the failure rate and corresponding decrease in the mean time to interruption (MTTI) of the computing system. Therefore, frequently occurring resource failures will drastically affect the execution of big data applications running on high performance computing systems. In order to hide the occurrence of faults, or the sudden unavailability of resources, fault-tolerance mechanisms (e.g., replication or checkpointing-and restart) are usually employed. These optimizations are especially important with applications involving big data.

To employ scheduling algorithms along with fault tolerance mechanisms for achieving high performance on parallel graph walks, while addressing both data scaling and resilience during the analytic search queries over large graphs.

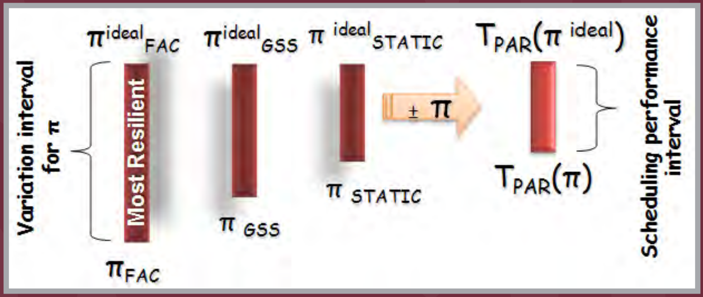


Dynamic scheduling enables load balancing of parallel tasks.

TECHNICAL APPROACH

In this work, we propose to enhance the flexibility and resilience of GEMS's current implementation for performing parallel graph walks (conceptualized as nested loops), which only support basic static heuristics to schedule an exponential number of parallel tasks (result of executing loop structures) in heterogeneous computing environments. First, this approach will be achieved by incorporating a node level scheduling functionality inside the GMT layer of GEMS library. This will assist with fine grain parallelism of computational tasks by dynamically scheduling of parallel tasks with diverse computational needs spawned by the graph walks across the cores (workers) inside the nodes, in order to avoid performance degradation due to load imbalance (where query data is highly skewed) at cluster level. Secondly, we plan to add provision of fault tolerance to the GMT layer so that the data pertaining to a failed node is not lost, and the parallel tasks scheduled on a failed node can be rescheduled, thus adding resilience to GEMS stack in the presence of resource failures.

The ongoing research work will address significant challenges faced by algorithms used to parallelize graph walks in order to achieve robust performance. The scheduling algorithms will take into account various properties of graph problems, as well as those of the system on which the query is being executed, such as: the unstructured and highly irregular nature of data, node failures, and failure detection and data recovery.



Resilient scheduling (that can handle the largest variations of π (perturbations)) results in having lowest impact on the performance of the parallel graph walk.

IMPACT

By analyzing and developing functionality for reliability of graph walk schedules in GEMS library, an increasing number and more complex big data analytic queries can be answered in the required set delivery times. Furthermore, the results of the analysis can be stored and learned, to enable dynamic selection of the most resilient scheduling algorithm from a portfolio of scheduling algorithms for a given execution scenario.

Large-Scale Graph Analytics and Risk Modeling for Detecting Malicious Cyber Nodes

Researcher(s): Dr. Hugh Medal,
Dr. Mohammad Marufuzzaman,
Dr. Song Zang, and Dr. Linkian Bian

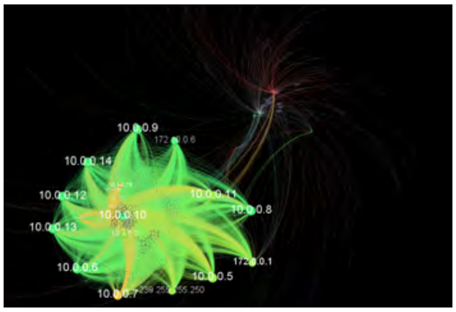
CHALLENGE

Although many approaches exist for detecting malicious nodes in a cyber network, most of these approaches assume a static attacker who does not adapt behavior according to the characteristics of the detection algorithm. However, research has shown that these methods are vulnerable to manipulation or evasion by an attacker. For example, a botmaster planning a DDoS attack may explicitly try to evade detection by modifying the behavior of the botnet so as to "poison" the data observed by an anomaly-based detection algorithm. Thus, there is a need to develop detection approaches that explicitly account for an evasive attacker. Some anti-evasive approaches have been developed, but none of these use graph analytics. While graph-analytic approaches show promise due to their ability to capture complex relationships between nodes, they can be computationally expensive.

TECHNICAL APPROACH

Our project goal is to quantify the effect of evasion on graph-based detection algorithms. In our experiment, we are generating a simulated NetFlow dataset with billion nodes and edges on shadow II with MPI programming. We will replicate the characteristics of dataset which is presently available like VAST dataset. We will then inject malicious activity into the dataset, we will begin with non-evasive and later extend with evasive malicious activity. We will implement method for monitoring graph characteristics to detect non-evasive anomalies. We will compare the detection performance of our technique e.g. success and false alarm rate for non-evasive and evasive malicious activity.

Developing large-scale, graph-based methods for detecting evasive cyber adversaries



Network for VAST dataset

CURRENT PRACTICE

Detection techniques are classified into two main categories: those based on setting up honey-nets and/or Intrusion Detection systems. The later one can further be subdivided into anomaly-based and signature-based detection systems. Anomaly-based detection systems can be either host based or network based. But regardless of the efforts by the researchers, bot detection remains a challenging task because bot developers continuously adopt advanced evasion techniques to make bots stealthier.
Very few works have been done on this sector.

EXPECTED ACCOMPLISHMENT

We are expected to develop anti-evasive monitoring and detection tool: source code and documentation of verification. We are also supposed to apply this tool to multiple type of graphs and check its generality. Our long term goal is to develop new anti-evasion dynamic graph-based anomaly detection algorithms and new anti-evasion dynamic graph-based malicious node detection algorithms.

IMPACT

If our project is successful, it will result in increased understanding of how evasion can degrade the effectiveness of malicious node detection tools. As a result, companies and institutions will be more aware of the effect of evasion, so that they can take steps to mitigate its effect.

Netmapper Project

Researcher(s): Bob Reese,
Philip Akers, and Patrick Pape

CHALLENGE

The Netmapper Project helps with the problem of network complexity and reproducibility. In this age of the Internet-of-Things, even a small business network can get complicated over a short period of time. Networks today are much larger than just computers connected together and can contain multiple routers, VPNs, firewalls, switches, printers/FAX/copy machines, wireless devices, NAS storage boxes and computers. In addition to all of the above devices, the network may contain devices with different operating systems such as Windows and Linux. Whether this network is administered in-house or is out-sourced, it can be very difficult to keep track of what is attached to it and what effect attaching something new to the network will have. Netmapper can scan these devices, determine connectivity and operating system type, and in some cases even log in to the device to determine what services and software are installed on the device.

CURRENT PRACTICE

Networks are generally built by hand with an initial planning stage to resolve issues like IP address ranges, subnets, and connection to the outside world. Testing usually involves testing on the actual network since network virtualization is still fairly new. Parts of a network may be virtualized and tested such as a server or small subnet and there may be a small physical testbed.

TECHNICAL APPROACH

The Netmapper program can map out a network to show what devices are attached to the network and produce a visual roadmap of the network for an administrator to examine. Netmapper will also allow the administrator to save the network scan output. A difference check of a saved output can be compared against a later scan to see if there are any changes to the network over that period of time.

Netmapper, a tool to aid in determining network topology, visualization, and virtualization.

IMPACT

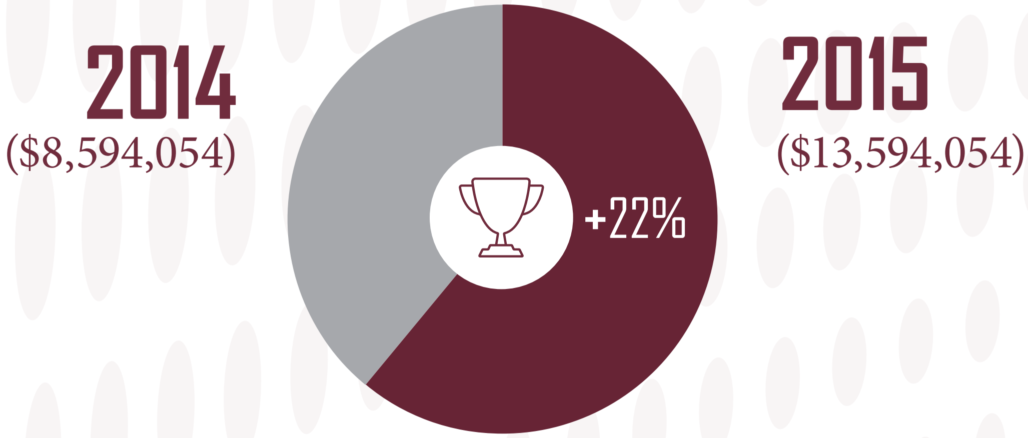
The Netmapper program allows the administrator to understand how the network is put together and see where problems might be occurring in the network due to what is connected at or near that point. It can show where potential bottlenecks or weak points are located in the network and allow the administrator to design contingency plans for the network.

Netmapper allows the administrator to see what devices might have been added to or removed from the network. It also shows what devices are accessible on the network. If a device that was previously accessible is still present but no longer accessible to Netmapper, it could be indicative of a problem. If there is a new device on the network that is unknown, it should be identified and understood.

Network stressing might involve overloading the network with traffic or doing intrusion testing to see how it responds to a new set of firewall rules. It might also involve adding a new Domain Controller or DNS server to the network to see how it will respond. A virtual network allows the administrator to do this kind of testing without interfering with the users on the live network.

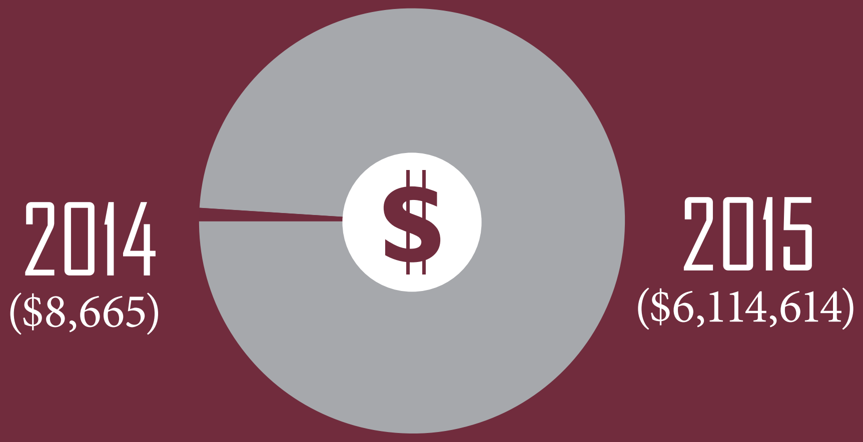
DASI by the Numbers

11 PROJECTS



AWARDS

EXPENDITURES



DASI Administration

Dave Dampier

Director of DASI and Professor of Computer Science and Engineering

James Fowler

DASI Associate Director for Analytics and Professor of Electrical and Computer Engineering

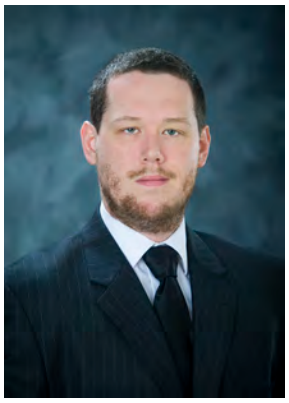
Sherif Abdelwahed

Interim Associate Director of DASI and Associate Professor of Electrical and Computer Engineering

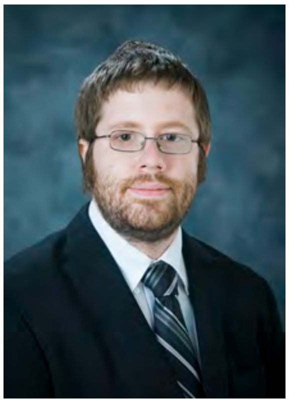
DASI Research Faculty

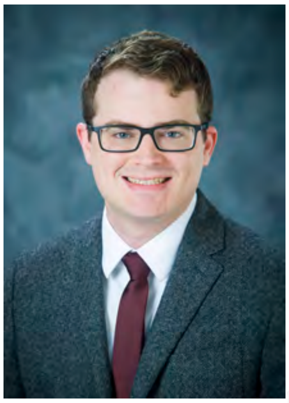
Uttam Adhikari
Associate Research Professor

Joe Crumpton
Associate Research Professor


Dae Glendowne
Associate Research Professor

Puntitra Sawadpong
Associate Research Professor


Wesley McGrew
Associate Research Professor


Patrick Pape
Associate Research Professor

Jian Shi
Associate Research Professor

Nitin Sukhija
Associate Research Professor

Research Associates

Phillip Akers

Mississippi State University
MSEE/MSCS, Hardware and
Software Security

Kendall Blaylock

Mississippi State University
MSIS, Digital Forensics

David Lee

Mississippi State University
MBA/MSIS: UM JD, Cyber
Security Policy

Cody Miller

Mississippi State University, Ph.D.
Candidate in CS, Virtualization,
Malware and Forensics

David Mudd

Mississippi State
University, MSECE, SCADA

DeMarcus Thomas

Mississippi State University,
MSCS, Ph.D. candidate
in CS, Machine Learning
and Security

DASI Affiliated Faculty

Jean Mohammadi-Aragh, Mississippi State University, ECE
Ioana Banicescu, Mississippi State University CSE
Linkan Bian, Mississippi State University, ISE
Rob Crossler, Mississippi State University, MIS
Yong Fu, Mississippi State University, ECE
Drew Hamilton, Mississippi State University, ORED/CBI
Thomas Haupt, Mississippi State University, CAVS
Kent Marett, Mississippi State University, MIS
Hugh Medal, Mississippi State University, ISE
Yaroslav Koshka, Mississippi State University, ECE
Mike Mazzola, Mississippi State University, ECE/CAVS
Mark Novotny, Mississippi State University, physics/CCS
Mahalingam Ramkumar, Mississippi State University, CSE
Bob Reese, Mississippi State University, ECE
Merrill Warkentin, Mississippi State University, MIS

Publications

Peer-Reviewed Journals

Fairley, J., Thompson, S.M., & Anderson, D. (2015). Time-Frequency Analysis of Flat-Plate Oscillating Heat Pipes. International Journal of Thermal Sciences. Elsevier. 91, 113-124.

Horstemeyer, M., Hughes, J. M., Sukhija, N., Lawrimore, W. B., Kim, S., Carino, R.L., & Baskes, M. I. (2015). Hierarchical Bridging Between Ab Initio and Atomistic Level Computations: Calibrating the Modified Embedded-Atom Method (MEAM) Potential (Part A). Journal of Materials. Springer US. 67(1), 143-147.

Hughes, J. M., Horstemeyer, M., Carino, R.L., Sukhija, N., Lawrimore, W. B., Kim, S., & Baskes, M. I. (2015). Hierarchical Bridging Between Ab Initio and Atomistic Level Computations: Sensitivity and Uncertainty Analysis for the Modified Embedded-Atom Method (MEAM) Potential (Part B). Journal of Material. Springer US. 67(1), 148-153.

Johnston, A. C., Warkentin, M., & Siponen, M. (2015). An Enhanced Fear Appeal Framework: Leveraging Threats to the Human Asset through Sanctioning Rhetoric. MIS Quarterly. 39(1), 113-134.

Miller, C., Glendowne, D., Dampier, D., & Blaylock, K. (2014). Forensicloud - An Architecture for Digital Forensic Analysis in the Cloud. Journal of Cyber Security and Mobility. 3(3), 231-262.

Ormond, D., & Warkentin, M. (2015). Is This a Joke? The Impact of Message Manipulations on Risk Perceptions. Journal of Computer Information Systems. 55(2), 307-327.

Pan, S., Morris, T., & Adhikari, U. (2014). Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems. IEEE Transactions on Smart Grid. 99, 1.

Pan, S., Morris, T., & Adhikari, U. (2014). Classification of Disturbances and Cyber-attacks in Power Systems Using Heterogeneous Time-synchronized Data. IEEE Transactions on Industrial Informatics. IEEE. 3, 662.

Shi, J., Amgai, R., & Abdelwahed, S. (2015). Modeling of Shipboard MVDC System for System Level Dynamic Analysis. Electrical Systems in Transportation, IET. IET. 4, 1-12.

Shropshire, J., Warkentin, M., & Sharma, S. (2015). Personality, Attitudes, and Intentions: Predicting Initial Adoption of Information Security Behavior. Computers and Security. 29, 177-191.

Peer-Reviewed Conference Abstracts

Warkentin, M., Vance, A., & Johnston, A. C. (2015). Introduction to the HICSS-48 Minitrack on Innovative Behavioral IS Security and Privacy Research. Proceedings of the 2015 Hawaii International Conference on Systems Science. Kauai, Hawaii.

Peer-Reviewed Conference Papers

Adhikari, U., Pan, S., & Morris, T. (2014). A Causal Event Graph for Cyber-Power System Events Using Synchrophasor. 2014 IEEE Power Energy Society General Meeting (PESGM). National Harbor, MD: IEEE.

Adhikari, U., Morris, T., & Pan, S. (2014). A Cyber-Physical Power System Test Detection Systems. 2014 IEEE Power Energy Society General Meeting (PESGM). National Harbor, MD: IEEE.

Babaei, M., Shi, J., Abdelwahed, S., & Zohrabi, N. (2015). Development of a Hybrid Model for Shipboard Power Systems. IEEE Electric Ship Technologies Symposium (ESTS). Old Town Alexandria, Virginia: IEEE.

Borges-Hink, R., Beaver, J., Buckner, M., Morris, T., Adhikari, U., & Pan, S. (2014). Machine Learning for Power System Disturbance and Cyber-attack Discrimination. 7th International Symposium on Resilient Control Systems. Denver, CO, USA: IEEE.

Crumpton, J., & Bethel, C. L. (2015). Validation of Vocal Prosody Modifications to Communicate Emotion in Robot Speech. 2015 International Conference on Collaboration Technologies and Systems. Atlanta, GA: IEEE.

Crumpton, J., & Bethel, C. L. (2014). Conveying Emotion in Robotic Speech: Lessons Learned. 23rd IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN 2014).
Edinburgh, UK: IEEE. 274-279.

Dampier, D. (2015). Building a Successful Cyber-security Program. Proceedings of the Southeast Cyber Security Summitt. Huntsville, AL.

Glendowne, D., Miller, C., McGrew, R., & Dampier, D. (2015). Towards a Feature Set for Classifying Malicious DLLs in Windows. 2015 IFIP Working Group 11.9 Conference on Digital Forensics. Orlando, FL: IFIP.

Mutchler, L., & Warkentin, M. (2015). How Direct and Vicarious Experience Promotes Security Hygiene. Proceedings of the 10th Annual Symposium on Information Assurance (ASIA). Albany, NY. 2-6.

Pape, P. R., & Hamilton, J. (2015). Generating Interest in Cybersecurity through High School Digital Forensics Education. 7th Annual Southeastern Cyber Security Summit. Huntsville, AL.

Price, S., Anderson, D., & Luke, R. H. (2014). An Improved Evolution-Constructed (IECO) Features Framework. Symposium Series on Computational Intelligence. Florida.

Shi, J., Abdelwahed, S., & Zhu, W. L. (2015). Development of a Control-based Performance Management system for Shipboard Power Systems. Electric Ship Technologies Symposium (ESTS). Old Town Alexandria, Virginia: IEEE.

Sharma, S., & Warkentin, M. (2015). Exploring the Role of the Temporary Workforce on Information Security Policy Compliance. Proceedings of the 9th Annual Symposium on Information Assurance. Albany, NY.

Thornton, Z., Mudd, D., Morris, T., & Hu, F. (2014). Virtual SCADA Systems for Cyber Security. 2015 IEEE International Symposium on Technologies for Homeland Security. Waltham, MA: IEEE.

Web Publications

Nelson, L., Godwin, D., & Abbott, C. F. (2014). Distributed Analytics and Security Institute Website - 2014.

The Distributed Analytics and Security Institute is a member center of the High Performance Computing Collaboratory at Mississippi State University.

The information contained in this publication is for the annual reporting period beginning July 1, 2014 through June 30, 2015.

Writer and Editor: Diane Godwin, MM, APR; HPC2
Designer: Robbyn Weaver, HPC^2
Photography: Mississippi State Office of Public Affairs

Discrimination based upon race, color, religion, sex, national origin, age, disability, or veteran's status is a violation of federal and state law and MSU policy and will not be tolerated. Discrimination based upon sexual orientation or group affiliation is a violation of MSU policy and will not be tolerated.

DASI

David Dampier
Professor of Computer and Science and Engineering
Director of Distributed Analytics and Security Institute
dampier@dasi.msstate.edu

MISSISSIPPI STATE UNIVERSITY™
DISTRIBUTED ANALYTICS
AND SECURITY INSTITUTE

DASI