Computational CyberSecurity in Compromised Environments

Special Cyber Operations

Research and Engineering

C3E Workshop Report, 2012

Contributors: Kevin O’Connell, Antonio Sanfilippo, Tamas Budavari,
Daniel G. Wolf and the C3E Team

Table of Contents

Executive Summary

The Special Cyber Operations Research and Engineering (SCORE) Interagency Working Group sponsored the 2012 Computational Cybersecurity in Compromised Environments (C3E) workshop at West Point, NY in September. The research workshop brought together a diverse group of top academic, commercial and government experts to examine new ways of approaching the cybersecurity challenges facing our Nation.

This was the fourth in a series of annual research workshops related to C3E, drawing upon the work of C3E efforts in 2009, 2010, and 2011 on adversarial behavior, models, data, predictive analytics, practical solutions and the need for understanding how best to employ human- and machine-based decisions in the face of emerging cyber threats. C3E holds as a central purpose the creation of an enduring community of interest who can continue to innovate on the analytic and operational challenges we face in light of these threats.

This year, we emphasized risk management and decision making in cyberspace by continuing to explore novel and predictive analytic approaches and visualization techniques. The ability to estimate the occurrence of future events using expertise, observation and intuition is critical to the human decision-making process. From a biophysical perspective, there is strong evidence that the neocortex provides a basic framework for memory and prediction in which human intelligence emerges as a process of pattern storage, recognition and projection rooted in our experience of the world and driven by perception and creativity. Human decision making can therefore be seen as a situation-action matching process which is context-bound and driven by experiential knowledge and intuition. However, despite the natural disposition of humans towards prediction, our ability to analyze, to forecast and respond to plausible futures remains one of the greatest intelligence challenges because of limitations on human reasoning due to cognitive and cultural biases. The objective of track discussions centered on how to assist analysts and policymakers in providing better cybersecurity analysis and response through the enablement of a human-based approach to decision-making that is unhindered by cognitive and cultural biases.

In support of augmenting human decision making, participants explored emerging “best practices” in visualization that effectively aid cyber practitioners in addressing real world problems. The massive volumes of data being collected in the physical and cyber worlds are dwarfing our abilities to assess, identify, characterize, and prioritize items, objects, and issues of interest. This is the antecedent of any action designed to anticipate or respond. New visualization methods, as well as those that help humans respond more effectively, are required, especially to help analysts orient and assess large areas of data. As we discussed at C3E 2010, effective cybersecurity will have to take advantage of intelligent use of both humans and machines for the things that they each do best.

C3E remains focused on cutting-edge analysis and analytics and understanding systems, networks, and how people interact with them. In addition, while C3E is often oriented around research, we have begun to incorporate practical examples of how different government, scientific and industry organizations are actually using advanced analysis and analytics in their daily business, and creating a path to applications for the practitioner. This is important to providing real solutions to address cyber problems, rather than remaining at the theoretical level. These ideas are summarized in this report, including detailed appendices. As such, they are ideas that the C3E workshop participants thought to be worthy of additional U.S. government, academic, and private sector attention.

Introduction

The Special Cyber Operations Research and Engineering (SCORE) Interagency Working Group sponsored the 2012 Computational Cybersecurity in Compromised Environments (C3E) workshop at West Point, NY in September 2012[1]. Once again, this workshop brought together a diverse group of top academic, commercial and government experts to examine new ways of approaching the cybersecurity challenges facing our Nation. Having the workshop at the US

[1] Prior C3E events were sponsored or co-sponsored by the Science and Technology Lead for Cyber at the Office of the Director of National Intelligence and the Chief of Trusted Systems Research at NSA. These organizations continue to be energetically involved in C3E.

Military Academy (West Point) also offered the opportunity to faculty and some of the cadets engaged in cybersecurity education to attend the various sessions.

This year, the extraordinary potential of the C3E community of interest (COI) was demonstrated, not only in continuing discussions about the deep analytic challenges of cyberspace, but by drawing upon a hypothetical time-sensitive operational challenge (known from here as “the identity challenge problem”) designed to focus attention on analytic methods and visualization techniques. The C3E COI met initially in April to review the identity challenge problem, and to identify initial analytic pathways and visualization approaches to attack it: these catalyzed our initial discussions at C3E. While the identity challenge problem was not a cybersecurity problem per se (it was based on an epidemiological challenge) it did highlight innovative approaches to finding tiny anomalies in big data that characterize many classes of problems, including those we confront in cyberspace.

As always, though the problems in cyber security are many and varied, and the types of expertise required to address them diverse, the group was concerned with a very specific part of the problem: how to enable smart, real-time decision making in cyberspace through both “normal” complexity and persistent adversarial behavior? Related to this was the challenge of presenting the necessary information to the decision maker in an effective visual manner, including the link between visualization and perception. The workshop was designed to draw upon earlier C3E efforts and advance new thinking about how to anticipate future challenges in cyberspace.

As in previous years, the workshop centered on topics that were selected based on their relevance to previous C3E themes as well as their potential importance in the context of addressing cyber challenges. The 2012 track session topics involved: (1) assessing decision making and risk management issues in cybersecurity; and (2) assessing the role of visualization in cybersecurity and its link to human perception.

The Identity Challenge Problem

C3E 2012 marked the use of an additional approach to understanding and developing novel approaches to cybersecurity: intense focus on a specific but hypothetical, time-urgent operational challenge to emphasize analytic methods and visualization techniques. This was a

C3E innovation in terms of strengthening the C3E COI over time as well as allowing intense focus – both in the period between a mid-year (April 25, 2012) C3E session and C3E, as well as during the C3E workshop itself – to focus on these approaches and their broader purposes, including their application to cybersecurity. For example, determining analytic approaches to deal with the enormous influx of data being delivered to cyber analysts has become an essential aspect of cybersecurity, and is the subject of intense U.S. federal government focus, including SCORE and OSTP's Big Data Research and Development Initiative. Consistent with this mandate, C3E has become involved in facilitating leading edge research in developing/recognizing novel approaches to cybersecurity challenges.

At our April event, the C3E COI initiated multiple research paths in response to a problem developed by C3E team in order to catalyze discussions around analytic methods and visualization techniques to respond to this challenge. The problem was scripted, the results and goals actually known. Participants were asked to determine the analytic methods and techniques that would aid in locating a fictional "individual with a deadly virus" who was moving around, with potentially catastrophic epidemiological consequences. During the one day event, participants began to brainstorm possible methods and techniques for locating the unidentified patient within a synthetic dataset of 345,987 nodes. However, in order to keep focus on generalizable concepts, the actual dataset was not disseminated until the end of the day. C3E has successfully maintained itself as an innovative and diverse community of interest not by determining the precise isolated solutions to each and every problem, but rather by proposing multiple generalizable solutions that could apply to a wide array of similar problems. While the answer may be interesting, it is more far more desirable, for C3E purposes, to identify the innovative analytic methods that foster the resolution of the initial challenge and the inevitable challenges of the future.

Introducing the identity challenge problem enabled C3E to finally shape state of the art research efforts being conducted by participants. As an extension of the April mid-year event, participants were asked to submit proposals concerning potential research projects addressing the identity challenge problem. Among those submitted, eight were selected. While each project proposed a unique methodology for the identity challenge problem, all were concerned with achieving the same final result; the identity and the location of the infected patient. This offered a measurable success metric that enabled each research project to be evaluated based on how their approach compared to other approaches in yielding new information and providing a generalized solution.. Based on final submissions, several research endeavors were selected to present their findings amongst peers at the C3E annual gathering.

By incorporating the element of competition, C3E demonstrated its ability to encourage research in developing unique approaches to cybersecurity challenges.. The value of competition is continually demonstrated by academic communities as an effective driver for inspiring innovation and excellence in students and faculty. Similar to C3E's endeavor, many cybersecurity gatherings around the world encourage participants by offering them the opportunity to prove the value of their research relative to other endeavors. Competition continues to reappear within so many cybersecurity endeavors because of its effect in in inciting high quality insight. Fortunately, the effects were once again proven by the rigor of the results concerning C3E's identity challenge problem. Although only four projects were selected to present at the September workshop, the selection proved difficult due to the exceptional research produced by each candidate. While each project did not necessarily yield the correct answer, each of the candidates successfully demonstrated a unique way of applying a generalized analytic methodology to detect a piece of information that would have gone undetected had it not been for their research.

The identity challenge endeavor was also unique in its effort to effectively bridge the gap between the classified and unclassified communities. The sensitive nature of cyber related endeavors has forced a wedge between those capable of developing solutions and those who are cleared to implement them in the real world. As a consequence of the lack of collaboration between these two communities, researchers remain ignorant of the specific issues that need their attention because of the restrictions faced by the classified community in divulging the specific challenges. In addition to weakening progress on both ends it has also restricted opportunities for individuals and organizations to build upon one another's findings. To help bridge this gap, C3E sought the opportunity to provide the research community with a rare glimpse into the types of challenges faced by the classified community every day and, similarly, to provide the classified community with unhindered observation of the innovative approaches circulating throughout industry and academia. Essentially, the effort proved the value of facilitating the bi-directional channel of insight that has long been desired by both communities.

As a community dedicated to inspiring actionable solutions, C3E efforts must also be evaluated for their overall functional benefit to cybersecurity. While the challenge problem itself is not directly related to cybersecurity, it provides a robust framework for locating small anomalies in a massive dataset using limited and incomplete data. The challenge of locating a single patient among hundreds of thousands of people is analogous to challenges faced cybersecurity practitioners responsible for locating a single anomaly amongst a large datasets. The rigor of the analogy was questioned during C3E discussions as the point was raised that the dataset involved was not representative of the volume of data handled by true cybersecurity practitioners. However, the intention of this approach was not discover the exact mechanism to be implemented, but rather to develop a set of approaches that, with further tailoring and exploration, have the potential to address our real world needs.

The inclusion of the Identity Challenge problem also acted as a catalyst for the two central topics of the September 2012 annual event: Decision Making and Risk Management; and Visualization.

Cybersecurity: The Practitioner's View

While theoretical concepts remain at the core to a better understanding of cyberspace, C3E continued its pursuit in 2012 of the many issues confronting cyber practitioners (e.g., information technology administrators, operations center managers, intelligence and law enforcement analysts) every single day. This year's discussion strongly emphasized the need to ensure that the analytic methods and technologies derived from C3E workshops are in direct response to practical limitations and needs. C3E has always kept practitioners in mind, but chose to bear down on them in a greater way this year and be more explicit about practical issues. Many of the researchers were surprised at the quantity and diversity of the cybersecurity challenges that a practitioner encounters in a given day. This exposure to actual real-world problems added a new dimension and positive impact on all the discussions at this year's C3E workshop.

Beyond the practical implications of the Identity Challenge Problem, three keynote presentations at C3E were dedicated to emphasizing the specific challenges of the cyber practitioner. The first presentation was MIT's Cyber Red/Blue simulation that took participants through a set of complex tradeoffs associated with multiple attacks in a time-sensitive environment for a corporate manager. The second talk, which was given by Carnegie Mellon, offered participants a better understanding of how to achieve resilience in command and control of operational networks. One final presentation provided detailed statistics and practical use cases on the daily activities and realities of cyberspace at Los Alamos National Laboratories (LANL). This talk offered insights on the practical realities of defending a critical enterprise to the largely theoretical research that is underway in the C3E community. Cybersecurity is a second-by-second challenge, with multiple tradeoffs and uncertainties. The LANL presenter emphasized that the goal of any cyber-program should always be to support the mission. If the cyber program is at odds with the mission, than the cyber program is simply not worth its while.

Achieving mission centric solutions require adherence to four imperative goals: minimizing the likelihood of an event, minimizing the mission impact, maximizing the speed to resolution, and learning from the event. By incorporating these goals into our processes, researcher and developers will improve the effectiveness of their innovations to address real world challenges. The concepts and theories derived from C3E workshops must have the practical capacity to someday translate into solutions that enable practitioners to achieve the four goals identified by the LANL practitioner.

With the experience of the Identity Challenge problem and the context of the practitioner in mind, we now focus this report on the work of the different tracks at the C3E workshop.

Decision Making and Risk Management

Because of the high speed at which change occurs in the cyber world, the rate at which protective action for cybersecurity needs to be taken is higher than in many other operational domains and it tends to have more immediate consequences. Decision making and risk management are therefore critical components of any cybersecurity approach. The C3E track discussion focused on challenges of decision making in cybersecurity, methods that address these challenges, and the evaluation and user acceptance of the technologies emerging from these methods.

Estimating the occurrence of future events using expertise, observation, and intuition is critical to the human decision-making process. Human decision making can be seen as a situation-action matching process which is context-bound and driven by experiential knowledge and intuition. Despite the natural disposition of humans towards prediction, our ability to analyze, forecast, and respond to plausible futures – whether in cyberspace or elsewhere -- remains one of the greatest intelligence challenges because of limitations on human reasoning due to cognitive and cultural biases. The objective of the C3E track discussions centered on how to assist analysts and policymakers in providing better cybersecurity analysis and response through the enablement of a human-based approach to decision-making unhindered by cognitive and cultural biases. In response to these human decision making limitations, there is growing interest in using visual analytics to help mitigate the effects of human biases by expanding sensory awareness and promoting human-to-human and human-computer collaboration. The massive volumes of data being collected in the physical and cyber worlds are dwarfing our abilities to assess, identify, characterize, and prioritize items, objects, and issues of interest. This is the antecedent of any action designed to anticipate or respond. New visualization methods, as well as those that help humans respond more effectively, are required, especially to help analysts orient and assess large areas of data. As we had discussed at C3E 2010, effective cybersecurity will have to use the best of the capabilities from humans and machines.

Our innate human ability to recognize patterns allows us to intelligently navigate through sizable datasets and form predictions based on various levels of analysis. While the human’s exploratory and pattern recognition capabilities may be unrivaled, our ability to apply analysis towards split second decisions remains lacking. This year’s discussion was focused on the perspective of cyber practitioners; those who are confronted with the need for quick, timely decisions every day but lack the methods and technologies to aptly respond. Eye opening statistics regarding the frequency of attacks revealed the dire need to enhance understanding of decision making and risk management processes. Where do the deficiencies lie? And what are the methods and technologies that sufficiently address the gaps?

Throughout discussions, participants reiterated the effect that inefficient decision making processes have on the ability of cyber practitioners to aptly respond to incidents. As the rapidity of attacks increase, there is a need to streamline the decision making process to quickly address current threats and possibly even to predict or project new ones. While a swift machine automated response may be enough to address some incidents, there is no substitute for the observation, analysis, and intuition offered by actual humans. (Remember that we are operating at the level of cognition here and nowhere near issues like authority or policy.)

An optimal solution might involve an automated system to handle well-defined (or previously characterized) attacks coupled with a human in the loop for the new and unusual incidents. The fact that we will rarely ever have the luxury of waiting for complete, uncorrupted date requires a flexible approach where we learn to make the best decisions possible with the data we have. As humans, we are further disadvantaged by our innate weaknesses that restrict our ability to analyze data objectively. As we continue to allow our own cognitive and cultural biases to distort reality, we effectively restrict the innate human qualities that positively serve our efforts.

Our biases also limit intelligent decision making because of the effects they have on situational awareness. How we perceive an image is strongly based on our own personal experiences. Cognitive biases shape how humans approach and interpret data and prevent us from dealing with unfamiliar circumstances. This raises the question: Given our unique individual experiences, how do we recognize something we have never seen before? Many make the argument for pushing advanced pattern detection technologies capable of massive parallel computing. A significant factor to consider is: As humans we are considered the epitome of pattern recognition. Can a machine ever truly replicate this? However, adopting technology as good at pattern recognition as humans is ideal because of the opportunity to avoid bias. While completely eliminating human bias is impossible we may be able to leverage the objectivity of technology to inform us what we are not considering. However, there remains a need for techniques to enhance person- machine collaboration to bridge across human and machine intelligence. C3E 2012 emphasized the need to move away from the conventional approach of interacting with machines and towards an open feedback loop where information flows both ways. Theoretically, everything the human does should drive automation and everything the automation does should drive human action. Instead of a machine coming back with a single response to an isolated question, machines should analyze questions to deduce other related questions/information that could guide the human decision making process.

Technology can provide solutions to decision making challenges in cybersecurity, but ultimately acceptance by the practitioner determines the effective impact of the technological solutions provided. Human-computer interaction methods that leverage human intuition and creativity through ambience intelligence and highly interactive interfaces can greatly advance user acceptance by promoting human-machine cooperation, as does the availability of transparent ways of evaluating the decision making workflow.

Visualization

As a new approach, C3E integrated the use of visualization techniques into the workshop dialogue by exploring the potential impact of these techniques on improving human decision making in cybersecurity. Visualization has arisen as a potentially game changing advancement in addressing cybersecurity challenges. The importance of visualization in the C3E community stems from the potential contribution that effective visualizations can have on improving human decision making. Effective visualizations present enormous sets of raw data in a way that easily lends itself to being understood by the human mind. Sophisticated visualization systems can act as a reminder of the information we chose to ignore, the places we forgot to look, and the questions we didn’t ask. As with other fields trying to capitalize on the advantages of visualization, the field of cyber security has the opportunity to become profoundly strengthened by a symbiotic relationship between robust visualization tools and trained analysts.

The ongoing development of visualization tools and techniques has led to revolutionary advances with how humans interact with and communicate data. They help our minds see patterns and identify outliers in a way that is simply impossible to achieve with simple text. Attributes such as data manipulation, model scalability, and details on demand enable visualizations to capitalize on our strengths and mitigate the effects of our weaknesses. They empower humans to expand our innate pattern recognition skills to their full potential by offering automations that are capable of augmenting, and not interfering, with the human visual and reasoning processes. Advanced visualization systems enable the user to correlate information from various sources and databases to explore “what-if” hypotheses. With their effective functionality and intuitive controls, visualizations have the potential to provide a dynamically representative environment in which an analyst can actively explore, communicate, and anticipate.

In preparation for this year’s annual workshop, C3E organizers conducted a review of scholarly and professional articles on the latest methods and techniques concerning information visualization. The articles revealed the immense level of interest that is currently directed at developing effective visualization techniques enabling cybersecurity. For example, efforts aim to aid analysts throughout the entire investigative cycle by first providing users with a high level view that enables situational awareness of the cyber environment at large. Users can then explore areas of concerns by drilling down to specific areas and tweaking features such as filtering and modeling until the clearest possible representation of the area is achieved. Ideally, exploratory functionality of these visualizations should not stop there. Tools should enable hypothetical exploration of data by allowing users to use their own reasoning and creativity to explore “what if” scenarios that could potentially lead to valuable insights.

In recognizing all that visualizations have to offer, the question then stands as to how much detail is actually possible? What are the current, most effective techniques, and how do we expand upon them? We have hierarchical node-link diagrams that provide us with tree-like images that we can traverse up and down; Arc diagrams that use pattern recognition algorithms to connect data; and Matrix views that can show us representations of clustering. But how do we optimize the techniques we have to empower the human’s exploratory qualities? How do we improve interaction, scalability, and even the use of color so that it enables efficient human cognitive processing? For example, it is widely accepted that providing multiple graphs and representations of the same data empowers users to exercise their pattern recognition skills. Correlations identified by users in one model are immediately transferred to other models of the same data. This not only allows users to see correlations from different perspectives but also inspires interactive data analysis where the visualization generates an interactive environment that allows both human and machine to build from one another’s findings. We must remember that the objective here is not to create visualizations that explore data autonomously, but rather to create visualizations that facilitate a mixed-initiative response of both machine and human.

At the foundation of improving visualization is first achieving an understanding of the human cognitive processes that are relied upon when interacting with a visual representation of data. This involves reaching across multiple social and scientific disciplines to achieve a better understanding of how concepts are stored and processed by the human brain. From here we can then approach challenges such as: How to improve models of human-machine cognitive processing? And what could this mean towards creating a predictive human cognition model? The key to enhancing the interaction between humans and machines is to first understand how they function individually, and then determine techniques that capitalize on the individual qualities of each. For example, our perception provides us with an initial impression of what is going on in a given environment. At this very important stage, how do we incorporate what we know about how humans perceive into how the machine chooses to display the information? To achieve this, we must rely on multiple disciplinary efforts to reveal the relationship between visualization and perception in both the human and machine learning worlds. This is essential to improving the usability of our visualizations for analysts. It is often overlooked that the minds who design/develop the visualizations are not the same minds who use them. So how do we successfully capture and present things in a way that are intuitive to the analyst? Bridging this gap requires visualizations that can tailor what they display to a specific audience. Whether it be a programmer, network analyst, or senior decision maker, the visualization should recognize that each of these groups is interested in something different.

SRI International’s “From Perception to Interest: A bRIGHT Approach to Interacting with Big Data" was a C3E keynote demonstration of a visualization technology that leverages cognitive psychology to enhance human-machine interaction. The tool, referred to as bRIGHT, increases efficiency and effectiveness for users that are cognitively loaded and heavily tasked. By recording what the user does and sees the tool stores meaningful application events at high abstraction levels and then uses the information to populate contextual models. The knowledge in a user’s contextual model is then used to favorably adapt his/her user experience: if a user is exhibiting a behavior that is similar to one observed earlier, the machine can retrieve likely next steps for this user, fetch necessary information needed in these steps, and present the user with a menu-like list of next steps that have been instantiated with information to the greatest extent possible. bRIGHT represents a tight feedback loop between the user and the cyber device.

Through a better understanding of human cognitive processes, we begin to fully appreciate the astounding potential that visualizations have in facilitating human reasoning and decision making. However, in addition to expanding upon the visualization techniques themselves, we also require improvements in the analysts who are trained to use them. Our persistent research in developing new exploratory visualization methods and technologies assumes that fact that we already have the experts to use them. It is easy enough to train an analyst to use a visualization tool, but how do we train them to be creative? We can give them the resources, but how do we know they will fully leverage the tool's exploratory capabilities? The only way to better ensure this is to transform the profession of cyber analyst so that we eliminate the misconception that being a cyber-analyst simply means relying on the machine to do most of the work. How do we revolutionize training so that it inspires analysts to creatively explore those "What if" scenarios that reveal undiscovered insights?

The September 2012 C3E gathering addressed several of the visualization topics described above. Participants were particularly interested in the role of visualization in constructing the narrative of a story. Ideally, visualizations will reveal through various displays for the user: the who, what, when, where, and why of any given cyber threat or cyber attack. However, in addition to being the most important aspect, the "why," is also the most difficult for visualization tools to currently convey. Visualizations will allow the analyst to navigate through the masses of data to discover the details of the incident. Combined with story generation capabilities, how do we create visualization techniques that augment humans in other ways such as by using automatic hypothesis generators, highlighting differences of opinion, and providing visual representation of uncertainty? Another factor that was strongly emphasized was the need to formally distinguish between visualization tools that enable communication and those that enable exploration. While strong communication visualizations aid in providing a basis for collaboration they are not the same as the exploratory visualizations that are used to reveal spontaneous insights. Simply relaying information does not capitalize on the creativity and intuition of the human mind. As humans try to understand data, their thought process is similar to a tree where one may traverse up and down the tree exploring options. However, such a progressive exploration first requires a visualization that captures and presents things in a way that reflects human intuition.

Hopes of capitalizing on the efficiency offered by machines are also dependent upon enhancing the level of confidence that humans have in machines. How do we evaluate the level of confidence that can be assigned to a visualization technique? And how do we improve this confidence? We must expand our efforts to explore the machine's ability to progress from visualization to perception. How a machine visualizes an event and then perceives that visualization will allow a greater understanding of the machine's decision making process; which will inevitably improve the confidence and relationship between human and machine.

Conclusions and Next Steps:

C3E 2012 provided new insights into analytic methods and visualization techniques of potential use in understanding, mitigating and even possibly eliminating cybersecurity threats. We have tried to summarize here some of the main themes and approaches taken from an extraordinary gathering of talent: an outline of additional ideas generated at the workshop is appended to the back of this report, organized by track.

Lessons Learned from C3E

In addition to the substantive contributions made within C3E workshops and the C3E COI, we offer a number of process developments that have made C3E an increasingly successful venture since inception in 2009.

First and foremost, the high-levels of diversity and quality of the participants proved essential to maintaining a productive, active interaction throughout the workshop. While initial conversations across a wide range of academic and professional disciplines can be chaotic and seem at cross-purpose, the encouragement of the group toward common taxonomy and language – typically ones centered on analysis, followed by more explicit dialogue on cybersecurity – and the good nature of participants typically catalyzes to a higher level discussion. Organizers have to strike a fine balance between direction-less dialogue and over-scripting.

Track selection and organization is also essential to achieving tangible results. Among the organizational techniques that proved beneficial included the selection of C3E veterans as track leaders and the division of large track groups into smaller sub-groups in order to maximize participation.   For C3E 2012, the spontaneous adoption of a “speed dating” process within the visualization track allowed participants to engage in brief one-on-one conversations regarding their specific research ideas and interests.   This process served to enhance familiarity among C3E participants and promote interactions that help members identify overlapping research interests.

The Identity Challenge Problem served as an excellent catalyst for substantive discussions around a very practical and operationally-oriented analytic problem. It served as an excellent basis for stimulating the continuing efforts of the C3E COI. In addition to serving as a real world example, the Challenge Problem also added a new dimension of competition to C3E by offering participants the opportunity to present and discuss techniques and solutions.

The inclusion of presentations by actual practitioners proved invaluable in stimulating the research community to understand and consider the daily challenges encountered in real world operations. These talks added a different perspective to the sometimes purely academic view of the daily cybersecurity challenges.

Next Steps

C3E continues to highlight current and emerging aspects of understanding developments in cyberspace as an essential element of cybersecurity.   From four annual C3E workshops, and, increasingly, the continuous efforts of the C3E COI, new theoretical, conceptual and practical approaches to cybersecurity are forthcoming, especially in the areas of analytic methods, analytic tools and visualization techniques.   Especially in the area of visualization, it is recognized increasingly as a place to do work, vice a mechanism to show the results of work.

Of course, need for effective cybersecurity increases daily, and therefore the activities and processes of the C3E community are continually evolving. Our great success with the use of the C3E COI to attack 2012 Identity Challenge Problem – admittedly, a fictitious, scripted problem of not quite “big data” size – has emboldened us to find other kinds of Challenge Problems, including current and operational ones heretofore unsolved.

The results of the C3E community are being recognized at many levels. We hope to publish a paper in 2013 (tentatively known as "What Have we Learned From C3E?") that speaks to the value provided by maintaining this community and offers a comprehensive view of the lessons learned incurred since its inception.

Appendix A: C3E 2012 Input Papers

C3E 2012 Content Development Participants	
Title	Exploring the Applicability of Formal Methods to Identity-Discovery
Participant	Matt Sottile
Organization	Galois
Summary	Researchers propose to explore whether formal methods approaches are adaptable to the C3E Identity Discovery Challenge. Successfully adapting formal methods tools would be significantly beneficial; as such tools are very efficient at handling and searching very large and intricately connected data sets. Their work will focus on developing flexible ways to model relationships between data in such a way that the efficient search-based Satisfiably Modulo Theories (SMT) tools can rapidly trim out complexities in the data set and prove or disprove identity hypotheses.
Title	SCALABLE METHODS FOR C3E CHALLENGE USING DOT PRODUCT REPRESENTATIONS
Participant	Sang Peter Chin
Organization	John Hopkins University, Applied Physics Lab and ECE Dept.
Summary	Researcher wishes to obtain a low dimensional representation of the C3E data by using Dot Product Representations*. Given the graph containing all the records in the challenge, Researchers will attempt to assign a vector to each vertex of the graph in such a way that the set of vectors capture the most important and essential structure of the graph. By applying dot product representation to the C3E graph, some notable structures are expected to emerge. Researchers will then investigate further the set of vertices in this group to look for the mysterious virus-carrying person. _*Random Dot Product Representation:_ A vector is assigned to each vertex (point). The graph shows that the probability of an edge uv between any vertices (points)u and v is some function of the dot product **u • v of their respective vectors.**
Title	C3E Identity Challenge Problem
Participant	Tamás Budavári
Organization	The John Hopkins University
Summary	Researcher proposed an in depth exploration of a variety of related activities

	that could potentially lead to a meaningful joint analysis. These activities would include cleaning up raw data using automated Data Scrubbing techniques, examining the Connectivity of Nodes, applying advanced statistical concepts such as Spectral Methods and Clustering, and using the analogies such as resistor networks.
Title	An Analytic Method for Solving the C3E Identity Challenge Problem
Participant	Sven Deitrich
Organization	Stevens Institute of Technology
Summary	Researcher proposes to develop and apply a variation of a botnet detection algorithm called the Dye-Pumping Algorithm (DPA) to analyze the data. They believe that the C3E ID Challenge data can be viewed as a complex network that has similar characteristics to modern botnets and advanced persistent threats. Their proposed approach includes using the edge and node tables to develop similarity metrics in order to build a connection or contact graph for the DPA and adapt the DPA algorithm according to the new attributes. For a simple visualization concept, the dye pumping algorithm shows where dye accumulates; thereby effectively highlighting potential targets. **Dye- Pumping Algorithm:* Once a graph is structured and generated, the DPA iteratively pumps dye from the seed node and distributes it to other nodes in graph. Then, the algorithm picks the node which accumulates more dye than a threshold. The distribution depends on a certain heuristic which estimates the likelihood of a specific node **A** being a bot given that node **B** is a P2P bot. For each iteration, each node in the graph will update its dye accumulation level.
Title	Finding the Needle in the Haystack: A Predictive Graph Analytic Approach
Participant	Antonio Sanfliippo
Organization	Pacific Northwest National Laboratory
Summary	Researchers propose a methodology that is based off the idea of graph simulations. Researchers propose to utilize graph modeling techniques to incrementally reduce the data and then use graph matching techniques to compare newly generated networks to the original network. As data is incrementally reduced, researchers will look for any mismatches between the networks that are intrinsically related to area of the disease outbreak
Title	Self-Organizing Information Matching for Identity Discovery
Participant	Van Parunak
Organization	Jacobs Technology, Inc
Summary	The goal of the proposed study is to demonstrate the applicability of self-organizing information matching to the C3E Identity Discovery challenge. The proposal concentrates on a solution that has scalability, robustness, speed,

	interactivity, and generality. Jacobs has developed and demonstrated a *scalable* approach to the information matching problem through the use of self-organizing agent methods that is *robust* to incomplete, inconsistent, and noisy data, *fast* in its retrieval due to its any-time characteristic that is capable of handling dynamically changing information from user live data feeds, and *general* in its assumptions about record attributes or relationships.
Title	Identity Challenge Processing
Participant	Warren Hunt Ivan Sutherland
Organization	ForestHunt, Inc.
Summary	This proposal describes a small, exploratory effort, to determine whether their tools for symbolic analysis and visualization can help with nuanced identity discovery. Researchers wish to develop and extend the mechanisms in their (ACL2-based) tool suite in order to capture databases as association lists that are linked together as graphs. Researchers wish to utilize existing tools for rapid query development and data output winnowing and presentation. Research will involve brief collaboration with an experienced analyst who would aid in supplying questions to be later converted into appropriate queries.

Appendix B: The Risk Management and Decision Making Track

Challenges

Situational unawareness caused by Irrationality

- Our experiences shape how we approach and interpret data
- These experiences create biases that narrow our focus so that we are unable to deal with unfamiliar circumstances
    - We are limited by our experience because they prevent us from stopping events that we have never seen before
    - Cybersecurity Unawareness causes
        - ***Blissful ignorance***
        - ***Blind spot bias***

  - Sometime I get information and since it feeds into my blind spot, I am more quick to accept it as true
- ***Misdirection***
  - Getting misleading information from the attackers
- ***Attack or problem is ignored until a solution can be found***
  - “Don’t bring me a problem without a solution”
- ***Hindsight Bias:*** a set of tools that worked before may not necessarily be effective today
- ***Framing Bias:*** using too narrow of an approach
- ***Deformation Professionelle***: looking at things according to the conventions of one’s own profession, forgetting any broader viewpoint
  - Computer science vs. a mission focus
- ***Normalcy Bias***: refusal to plan for or react to a disaster that has never happened before
- ***Herding Bias:***
  - Too much focus on specific cyber threat types because of widely reported cyber breaches events adherence to a specific approach to cybersecurity
  - People now find it easier to be biased because social media allows them to group themselves with those who also share the same views
  - Herd Instinct: adopting the majority view in order to feel safer and avoid conflict
  - In-group bias: give preferential treatment to those perceived to be part of one’s own group

- Difference between cybersecurity and other fields
  - An experienced pilot will take shortcuts in protocol and they will be right because they have experience. However, in cybersecurity, taking a shortcut may be wrong because of the constant change. This is an example of a static system vs. a dynamic system
- High volumes of data make it difficult to establish proper signal to noise ratio

    - Includes the:
        - Notational bias
    - Is the dashboard up to date?
        - Are we taking the correct measurements

Fear of Consequences

    - A heightened awareness of the escalation of false positive counts can hinder action
    - Should I shut this node down or not because of the mission
    - The operator is aware of the false positives that could occur which deters him/her from committing to an actual decision
        - For example: the challenger explosion
    - If you get too many false positives than the system losses its credibility

Too Much Transparency

- The transparency of cybersecurity is too strong (it is not tangible). It's so transparent that we can't even see it
- Until it hits massively, it will not be a cultural norm to worry about
- The transparency makes it difficult to know when exactly you've been hit. Whereas other infrastructures are clear, such as train tracks, cyber infrastructure is invisible.
- So the question is: how can we increase the tangibility of cybersecurity?
- Ex: Sometimes the info can still be in your repository but you don't even know if it's been shifted. Whereas if you move a table from a room, you would know it.

Bounded Rationality

- The physical Limitations of information processing
    - Scale: impact can be much wider than in other forms of attack
    - Velocity: Things happen so fast that we can't get inside adversary's OODA loop
    - Volume of Data: Needle in the Haystack and it is hard to get a workable signal to noise ratio.

Social Issues

- Limits of decision making both within an organization and between multiple organizations
    - Velocity: Limits of organizational decision making
    - Organizational conflict

Methods and Technologies

Expand Sensory Awareness

a. Technology
    i. We need massive parallel computing to enable advanced pattern detection capabilities that will provide better sensory awareness. Pattern recognition is currently hampered by computational technology (i.e. sequential operations, economic limitations). Humans are considered the epitome of pattern recognition. If we could make a machine as good at pattern recognition as we are then we would be golden because machines have no biases. We will never totally get rid of our biases but what we need is a way to fill in our blind spots (i.e. a method for expanding our awareness). In the future, we will need a machine to tell us what we are not considering.
    ii. Need advanced machine learning techniques to detect cyber behavior that enables the generation of alerts, warnings, and leading indicators
    iii. Need to leverage ongoing analytic advances in other sciences
        1. Biology Network modeling
    iv. Need techniques to describe and detect deception

Facilitate Person-person Collaboration –

b. To complement each other’s skills and mitigate each other’s bias
c. Methods:
    i. Integrate different types of experiences.
    ii. Overcome cultural differences within the organization by creating a managing role focused on this task

1. Need to reduce the struggle between those more concerned with policy and those more concerned with technology

iii. Need to Rotate the Team composition

d. Technology:

i. Gaming and Postgame analytics – experimental methods

ii. Crowdsourcing, wisdom of crowds: aggregation of social intelligence

iii. A game is a great technology to use because it allows people to learn from one another.

1. Stimulates Questions such as:

a. How do the cyber adversaries know they are under surveillance?

b. Person used example of when criminal let them know how they gave themselves away. This is very rare since normally you will never get feedback on how you gave yourself away. Having a real simulation allows you to get feedback from the attacker

Facilitate Person- Machine Collaboration –

e. Method:

i. To bridge across human intelligence (focus, insight, intuition, analytical thinking, deeper semantic analysis) and machine intelligence (intensive data processing, enhance data through annotation) – bootstrap

f. Technology:

i. Bring together game theory and other decision theoretic approaches together to enable cyber simulations

1. There is currently a heightened ability to simulate scenarios but we are simply not using them enough.

ii. Automating response and keeping the human in the loop

1. The approach should always be to let humans continually interact with the machine which will guide the human along the way to come to a decision.

iii. Need Interactive visualization

iv. ACH approaches to intelligence and decision making
v. Transparency in analysis and reasoning

Context Awareness - Purely technical and operational foci can be reconciled by creating context awareness

- Differing views are sometimes the result of the context being unclear. Context awareness is difficult because the context boundary is often flexible. Context awareness can be achieved by identifying decision making tasks and the timeframe that is needed per task.
    - Example: everyday has a different set of priorities when looking at war because the context changes hourly
- Technology:
    - Data Dictionary/Ontology for Cybersecurity
    - Network analysis and simulation
    - Modeling human intent
        - Looking at future models and Assessing alternative outcomes so that decision makers can have some insight into possible future events
    - Predictive/forecasting analytics with human interaction to steer analysis
        - Must be careful not to introduce biases (use a data driven approach)
        - Must be able to predict unseen contexts
        - Ontology-driven to achieve generality

Communication across different parts of the organization

- Streamline knowledge management process
    - Need to know who has access to which information in order to be most effective in addressing cybersecurity.
    - Hierarchical process
        - Information = bottom up
        - Decisions = top down
    - Technology
        - Implement knowledge management processes that guides strategic vs. tactical decision making

Coordination across Cybersecurity across organizations

- Create a community in which access to information about ongoing attacks can be accessed to obtain advice (i.e. a Network of Networks)
- Provide Awareness about friendly networks
- Problem for Companies:
    - Reluctance to share information for competitive reasons
- Technology
    - Data Anonymization
    - Networks of Networks:
        - A way to share network views and context
        - Provides ways of describing and representing a network
        - How comprehensive network knowledge has to enable information sharing.
            - Access the rate of change of the internet
            - Promote Self organization

Predictive technologies

- Need to move away from the conventional approach. Information must flow both ways. Everything I do, as a human, should drive the automation and everything the automation does should drive my actions.
- You need to know exactly what the machine is giving you
    - Computer typically understands what you "said" but not necessarily what you meant
- It is not yet an open loop (feedback) process
- Need a way to capture both the user mental model and the machine's automation context
    - Could show this is WHY you made those decisions
    - Can certainly determine how well you executed on your own processes
    - May be able to go back and change architecture
    - What went wrong? Was I missing a data source?
    - What are the gaps I need to fill
    - What we need is, essentially, an audit trail
    - Essentially creating a post-reality analytics

Appendix C: The Visualization Track

Main points

Create a Storyline

- The role of visualization is to create a storyline
    - We need different visualizations for communication and exploration
        - What is communication vs. exploration? This needs to be defined.
    - Visualization for communication of existing insight.
        - Does the visualization provide a basis for collaboration?
        - E.g., Edward Tufte's visualizations are attractive because they are carefully crafted, but that is time-consuming. Current 3D visualization of networks is not useful for insight. Most just show data as an "eye chart" without enabling any insights to what is really happening.
        - Visualization for analyst support. We are currently terrible at providing information for analysts to use to move from visualization to perception/understanding.
    - Exploratory visualization
        - Pictures help us think; allows a user to substitute perception for cognition. Why do we spend a lot of time looking at something for understanding? Why do we scribble and draw diagrams? Humans use it to free up "working memory." Study EEG to measure cognitive load
        - When users try to understand data, their thought process is like a tree, where one may traverse up and down the tree exploring options.
    - Need alternative storylines
        - To show when storylines start competing with one another
    - Is it possible to provide one or more storylines through visualization?
    - How do visualizations best support the analyst?
        - Sense making or situational awareness
    - Who, What, When, Where, Why?
        - The "Why" is the most important
        - But "Why" is the most difficult to support with tools

        - Example: “Why is the actor doing what they are doing?”
- Lots of choices to make about themes and insights. The early stages of getting ready to decide incite the following questions:
    - How do you know whether your information is accurate?
    - How can you be certain you have enough to pull the trigger and go ahead?
    - Challenge: how to avoid cognitive dissonance? How can you understand in an unbiased way?
        - Once you accept information, you become the owner of that information and you begin to reject information that is different.
        - **Efferent** readiness – Expectations can affect what you see and hear/perceive. Therefore your expectations will match what you believe.
        - Humans naturally do this, so how do we get analysts to overcome their biases?

Use context to ensure actions are realistic

We need to make better use of context; it plays a key role in sharing info between analysts; we need cognitive task models. Context of the attack is very important versus context of the user to see what’s normal.

Achieve Situational Awareness

SA provides a fundamental underlying understanding to help analysts comprehend risk. Situational awareness does not happen on the screen – your view of situational awareness maps into your own perception. Visualization helps support SA, which remains the fundamental focus, not just the visualization by itself.

Tailor the visualization to different roles

- The visualization changes based on the role of the observer

Visualization should support a crowd sourcing function

- Enable a stream of hypotheses to come in from end users
- Enable quick feedback

    - Example: The “like” icon on Facebook

Leverage end users as part of an incident response team

- In complex systems such as health systems and nuclear power plants, it is difficult for users to react effectively – e.g., in the Three Mile Island incident, users were overwhelmed with some 3,000 alarm warnings. This is not unlike how humans interact with cyber threats.
- The most important analyst in any country is the president or king/monarch. The second most important group is the national security council/president’s cabinet. How does the president get briefed on the tsunami of data affecting his decisions and what can we learn from that?

Align Visualization Timeline with Human Timeline

- Prevent a Visualization Timeline that is faster than what human beings are capable of perceiving
    - Timelines should be presented at many different levels of granularity
- The full continuum of enhanced human faculties, and their time scales, include from:
    - Visualization to perception – done in milliseconds
    - Perception to recognition – takes from milliseconds to minutes
    - Recognition to understating – takes minutes to hours (only occasionally seconds)
    - Understanding to insight – from hours to days
    - Decision and action – both may take longer depending on the situation
- There are three situations for visualization: 1) faster than human response 2) “real-time” so a user can respond 3) long term. For circumstances like dog fights in the air, visualization could be in the form of dials and gauges, which would look far different from visualization for phishing attacks, which have a longer-term form. There needs to be an appropriate response for each attack. Would users need different tools or could they utilize the same tool to have a short-term or long-term view? Some attacks may happen too quickly for humans to react, so the system must be designed to auto-respond. A real-time interaction would be like a Bloomberg terminal.

Leverage the younger generation

  - Changing the ways the younger generation expects to use computers and visualizations (i.e. touch screen)
  - Shape the younger generation so that interaction with these visualization come naturally. In many ways today's technology and culture is already doing this.
  - Visualization and perception represent the early stages of capabilities we want to enhance in humans – e.g., very important in intelligence work.

Need a way to capture and present things in a way that reflects intuition

- Visualization should include not just the perceptions of the eyes but other sensory inputs.
- How can we visualize the data so the analyst can make the right decision? There is not a "game over" situation; both the attacker and the security analyst never cease the confrontation.

Task Oriented Visualization

- Has been successful in numerous other fields
  - i.e. geospatial, business intelligence, avionics, financial decision making
- Visualization should be driven by:
  - What decisions face the analyst?
  - What actions the analyst may take?
  - What is the impact of a decision?
- How do we visualize the logical process rather than visualizing the process taken by the analyst?

Augmentation

- Augmentation of humans with hypothesis generators
  - Combined with story generation
- Highlighting difference of opinion
- Highlighting the difference that a small set of data may make
- Need a visual representation of uncertainty

    - How much uncertainty is there when we connect the fuzzy dots

Additional Points

- With a targeted attack, it cannot be predicted ahead of time. Attacks are broken apart into stages, which might be not detected individually. Warnings happen in the middle of an ongoing attack. The question is whether to be able to predict attacks, externally or by other methods. There are too many dimensions to consider.
- How do we define success? - To have reasonable answers for most of the problems we face and insights that help us move forward.
- High dimensional visualizations needs projection tools
- Streaming for embedding is not always available
- Automated hypothesis generators
    - Possibly replace humans
- Perhaps brain devices for interaction?
- Vendors provide feeds that need to be federated
- Taxonomy of threats/events for merging and filtering.

Appendix D: 2012 C3E Workshop Agenda

Sunday evening	
1900 – 2030	No Host Reception Brief Welcome Remarks by Brad Martin, Dan Wolf, and Kevin O'Connell
Monday	
0700	Continental Breakfast in the Grant Ballroom
0750	Transportation from The Thayer Hotel to West Point Campus - First Bus
0815	Last Bus to Thayer Hall

0830	Introductions West Point Welcome - BG Trainor - Dean of the Academic Board
0900	Challenge Problem Responses – Panel/Presentations The Identity Challenge Problem - Jason Bellone Panel Presentations *Panelists:* Sven Brueckner, Peter Chin, Antonio Sanfilippo, Ivan Sutherland
1045	Morning Break
1100	bRIGHT Visualization - Patrick Lincoln/Grit Denker - SRI International "From Perception to Interest: A bRIGHT Approach to Interacting with Big Data"
1200	Transportation back to The Thayer Hotel - First Bus
1215	Last Bus back to The Thayer Hotel
1215	Working Lunch
1245	Lunch Presentation – Dr. Joseph Halpern, Cornell University “Beyond Nash Equilibrium: Solution Concepts for the 21st Century"
1330	Plenary Introductions of C3E Track Themes Antonio Sanfilippo – Decision Making / Risk Management Patrick Lincoln - Visualization to Perception
1400	Introducing the Role of the Practitioner – Kevin O’Connell Nancy Crabtree, MIT LL – “Cyber Red/Blue Playable Situation”
1445	Introduction of Track Leads & Initial Track Session w/Break
1700	Challenge Problem Response – Poster Session
1830	Adjourn

Tuesday	
0700	Continental Breakfast in the Grant Ballroom
0830	Keynote Presentation - Michael Kyle, LANL “Cybersecurity: A Practitioners View”
0930	Track Work Continues /w Morning Break Decision Making / Risk Management Visualization to Perception
1200	Working Lunch
1220	Lunch Presentation - Ivan Sutherland, Portland State University “Computers Yet to Come”
1300	Track Work Continues w/ Afternoon Break Decision Making / Risk Management Visualization to Perception
1500	Keynote Presentation – LTC Michael Lanham, Carnegie Mellon University “Science of Security: A Complex Socio-Technical System Perspective” - Command, Control and Resilience in Networked Systems
1600	Plenary "Snapshot" Review of Track Work and Discussion
1645	Adjourn
17150	Bus departs for No Host Culinary Institute of America, Hyde Park, NY
1830	Dinner at CIA (Optional – No Host Event)
Wednesday	
0700	Continental Breakfast in the Grant Ballroom

0830	Looking Forward, Keynote Presentation Adam Greenfield, Urbanscale “On Public Objects: Connected Things and Civic Responsibilities in the Networked City”
0930	Track Work including Morning Break
1100	Plenary Review of Track Efforts Antonio Sanfilippo – Decision Making / Risk Management Patrick Lincoln and Tamas Budavari - Visualization to Perception
1145	Summary and Closing Remarks Brad Martin, Dan Wolf, Kevin O'Connell
1200	Workshop Adjourns
1200	Box Lunch