Computational CyberSecurity in Compromised Environments

C3E Workshop Report, 2011

Contributors: Antonio Sanfilippo, Daniel G. Wolf, Kevin O'Connell,
Lonnie I. Carey, Jr., Thomas Longstaff, and the C3E Team

Executive Summary

The Science and Technology Lead for Cyber at the Office of the Director of National Intelligence (ODNI) and the Chief of Trusted Systems Research at the National Security Agency (NSA) co-hosted the 2011 Computational CyberSecurity in Compromised Environments (C3E) Workshop this past 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 third in a series of research workshops related to C3E, drawing upon the work of C3E efforts in 2009 and 2010 on adversarial behavior, models, data, the need for 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.

Predictive analytics was topic of the C3E 2011 workshop. The temporal and substantive dimensions associated with cyberspace are so challenging that warning, anticipation, and reaction are different than those from other threats, including aspects created by continuous, fast moving, or advanced persistent threats. Predictive analytics can be applied to understanding the presence of an adversary in a system or network, as well as the range of actions they could apply, and under what circumstances. Predictive analytics also allows the development of analytic models of normal, abnormal, and threat activities in order to anticipate, rather than react to developments.

This year, looking through the lenses of two potential foundations of predictive analytics – emergent behavior and the intersection of anomalous activities – C3E participants' highlighted observations and findings into a set of areas for further consideration, including:

- the importance of predictive analytics to the cyberspace security mission
- the demonstrated relevance of analytics to both theoretical understanding and practical application of cyberspace challenges
- the opportunities and challenges associated with "big data" including the relevance of perspective and analytic metaphor from other scientific disciplines
- data requirements and metrics for modeling cyber emergence
- issues related to model evaluation and interoperability
- conceptual approaches and real examples of the potential value of intersecting anomalies, and
- alternative perspectives on the attacker-defender calculus in cyberspace

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.

Table of Contents

Introduction

The Science and Technology Lead for Cyber at the Office of the Director of National Intelligence (ODNI) and the Chief of Trusted Systems Research at the National Security Agency (NSA) co-hosted the 2011 Computational CyberSecurity in Compromised Environments (C3E) Workshop this past 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 third in a series of research workshops related to C3E. It leveraged the efforts of the C3E 2009 workshop in Santa Fe, New Mexico on modeling adversary behavior as well as those of the C3E 2010 workshop in Santa Barbara, California on models, data, practitioner's needs 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.

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, such as we learned from the financial sector.

The workshop was an analytic workshop as much as it was about cyber security. Though the problems in cyber security are many and various, and the types of expertise required to address them are 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? The workshop was designed to draw upon earlier C3E efforts and advance new thinking about how to anticipate cyberspace developments.

Predictive analytics was the topic of the C3E 2011 workshop. The temporal and substantive dimensions associated with cyberspace are so severe that warning, anticipation, and reaction are different than those from other threats, including aspects created by continuous, fast moving, or advanced persistent threats. Predictive analytics can be applied to understanding the presence of an adversary in a system or network, as well as the range of actions they could apply, and under what circumstances. Predictive analytics also allows the development of analytic models of normal, abnormal, and threat activities in order to anticipate, rather than react to developments.

The C3E 2011 Workshop explored two key areas that potentially underpin the use of predictive analytics in cyberspace:

Emergent Behavior

The concept of "emergence", as used in modern science, usually refers to the complex behaviors that emerge from dynamic interactions between simple entities. Modeling emergent behavior in cyberspace can have a game-changing impact in the fight against cybercrime. The end of the 20th and beginning of the 21st Century have seen computers and networks permeate life and science, and embed us all into a new, strikingly complex system of computational networks. Such developments have broached groundbreaking advancements in the way we interact with infrastructure, institutions and other people, but have also created new vulnerabilities. Due to the ever growing complexity of computer networks, cyber threats quickly evolve so as to constantly present novel challenges. Simulating scenarios that embody the emergence of such challenges is the first step to developing a truly anticipatory approach to cyber defense. Several algorithms have been proposed to simulate emergence using as reference processes occurring in nature such as the flocking behavior of birds and the swarming behavior of insects. Some of these algorithms have been utilized to model cyber emergence with promising results. Among the questions explored at the C3E workshop were:

- Which are the best algorithms and natural-process metaphors to model cyber emergence?
- How do we harvest and calibrate evidence to inform models of cyber emergence?
- How do we test the validity models of cyber emergence?

The purpose of the Emergence track at the C3E Workshop was to achieve a better understanding of how to model emergent behavior in cyberspace to anticipate cyber threats.

Intersecting Anomalies

Within a compromised environment, anomalous behavior has been used to identify indicators of a wide variety of faults, both natural and malicious. Much of the research and development performed to address both the analysis of anomalies and the attribution of the behavior has focused on individual variations from specified or learned normal behavior along a single attribute or sensor (e.g., anomalous network traffic patterns). Human analysts often identify a series of intersecting anomalies, anomalies that may be related to the same behavior of interest, to gain a gestalt of the activity. These insights are particularly difficult to automatically identify and combine, but there may be techniques that could be applied if we consider the problem from a wide variety of potential sources intersecting about a common behavior. With our ability to handle big data, we have the opportunity to discover anomalies from a wider variety of sources.

Intersecting anomalies need not be strictly passive, but where there are hypothesized intersections of anomalies, we may generate specific stimulus events that help converge anomalies and more easily associate many anomalies to common adversary behavior. For example, cluster analytics may be appropriate to partitioning these intersecting anomalies and help to guide the appropriate generated events within the compromised system toward convergence. Some questions that we explored at the C3E workshop are:

- What are the optimal sets of anomaly sources and types that can be combined to identify an imbedded adversary?
- What algorithms are appropriate in discovering the relationships between different anomaly types?
- How does big data help to enrich our use of anomaly identification and analysis? Can we use scale to our benefit?
- Do intersecting anomalies help us to uniquely identify adversary behavior when most of what the adversary does is identical to normal behavior?
- How can we “game” the system to help deconflict/cluster anomalies that clarify the normal and anomalous system behavior?
- Can we combine multiple detected anomalies at scale and near-real time to address attacks in the real world? What are the barriers to successful application of such an approach?

Several assumptions influenced our approach to addressing this problem. Cyberspace is fast moving, so decisions need be instantaneous. Cyberspace is vast and complex, so cyber analytics must work with very large data sets. Cyberspace is constantly changing, and our understanding of that space must be continuously updated. Finally, there is the realistic assumption that gives our workshop its name:

cyberspace is often deeply compromised, so our analysis must be undertaken in full recognition of that adversarial dimension. These aspects formed the foundation of our work and our thinking.

Academics, government and industry officials spent two and one-half days focused on the implications of working in the compromised cyberspace environment, including the research and substantive developments that would be useful in enhancing our understanding of the operating environment and improving chances of successful and more secure computing.

During the workshop, experts first spent time in plenary session discussing the potential importance of predictive analytics and understanding how the threats have evolved in different ways, including the use of analytics to understand and sometimes mitigate or counter them. A panel on threat presented various perspectives on how the threat is changing. The group also discussed the challenges and opportunities associated with big data including the perspectives of different scientific disciplines and other areas. Experts from the fields of astronomy, behavioral sciences, biomedical informatics, computer sciences, financial fraud detection, and others brought perspective on the relationship of analytics and big data to cybersecurity, including ideas about new kinds of analytic tradecraft.

During the course of the two and one-half days, participants also met in two "track" sessions – one for emergent behavior and another for intersecting anomalies – to conduct tailored discussions and work through new concepts for evaluating cyberspace activities. Discussions were generally oriented around scientific and technology concepts and not around legal or policy issues.

The two and one half days also included keynote and topical speakers tailored to a particular aspect of the C3E workshop. Participants heard briefings and lectures about the role of predictive analytics in helping with the compromised cyberspace environment problem, the role of big data, government and industry perspectives on the threat and potential response, and advances in predictive analytics posited around the possibility that network penetration could be used to the advantage of the defender, thereby bringing new perspective to the attacker-defender calculus.

Participants highlighted observations and findings into a set of areas for further consideration, including:

- the importance of predictive analytics to the cyberspace security mission
- the demonstrated relevance of analytics to both theoretical understanding and practical application of cyberspace challenges
- the opportunities and challenges associated with big data including the relevance of perspective and analytic metaphor from other scientific disciplines
- data requirements and metrics for modeling cyber emergence
- issues related to model evaluation and interoperability
- conceptual approaches and real examples of the potential value of intersecting anomalies, and
- alternative perspectives on the attacker-defender calculus in cyberspace

These ideas are summarized and organized by track theme below. As such, they are ideas that the C3E community thought to be worthy of additional U.S. government, academic and/or private sector attention.

Cyberspace and the Role of Predictive Analytics

Predictive analytics is a broad term describing a series of statistical, data mining, and other analytic approaches used to identify trends, patterns and relationships among data that allow the user to develop analytic models and anticipate behavior. Within our construct, predictive analytics can be applied to activities undertaken by humans, machines, or combinations of both. One of the more common applications of predictive analytics of relevance to the C3E community is from the financial sector with the creation of credit scores, fraud detection, and their application to point-of-sale transactions.

Within the U.S. national and homeland security domains, cyberspace threats represent one of a new class of threats by which the ability to anticipate and warn is different from many other traditional threats. The time dimensions associated with cyberspace are so severe that warning, anticipation, and reaction are different than from other threats, including the temporal and substantive aspects created by continuous or advanced persistent threats. Predictive analytics can be applied to understanding the presence of an adversary in a system or network, as well as the range of actions they could apply, and under what circumstances. Predictive analytics also allows the development of analytic models of normal, abnormal, and threat activities in order to anticipate, rather than react to developments.

Predictive analytics can be applied across an entire network or system and include interactions with many different external actors and factors. Analytics can be applied across multiple time-frames (short-term to long-term), are efficient, and are capable of discovering novel patterns and relationships across data sets. Analytics can serve as a “cyber sixth sense” and play an important role in either automatic mitigation and response or triggering human assessment and response. The development of analytic models of our own and potential adversary behaviors, over time, allow for system-level understanding of what is going on as well anticipation, warning, response, and mitigation of anomalous behavior. Advanced models potentially allow us to discover what an adversary is doing – either from a persistent perch or in preparation for action – from the very first time of a specific action (e.g., reconnaissance, attack).

Far from the panacea that it is sometimes pitched as in the commercial world, predictive analytics have a number of aspects that must be carefully thought through to be effective:

- **Sources of error**: far from being a source of specific prediction in such an uncertain cyber environment, predictive analytics can be expected to provide 40-70% accuracy in assessing systems and networks. These errors can be narrowed down over time with appropriate revisit strategies, including incorporation of new data sources, including data that is deliberately provoked or queried from the network.

- **Temporality**: the selection of timeframe matters in predictive analytics, whether driving instantaneous decision and reaction or allowing the integration of data over longer timeframes. The selection of a specific timeframe will typically be oriented around the question than an analyst asks.

- **Granularity**: for many different cyber models, they exist at different levels of granularity, limiting the potential for model interoperability or worse, the creation of an “apples and oranges” effect (analytic failure). Further, cyber models will need to look at systems and networks at varying levels of granularity, with due consideration of the tradeoffs created between synoptic and precision views within a model. Integration over multiple data points can miss important error or detail if not organized carefully.

- **Context**: How should analysis be different on systems we control? How should they perform under attack, as opposed to good hygiene and diagnostics mode? How do we apply diagnostics in such a way that an adversary remains unaware of our analytic purpose?

Predictive analytics can be organized in such a way as to be hypothesis driven, where the hypothesis drives the strategy through which you look at and assess the data (not what one is looking for analytically or the best fit for the data at hand). Vantage points and focal distance need to be deliberate, and chosen in such a way as to provide coherence to analysis; multiple vantage points also provide the potential for more precise assessment. Emerging loci for predictive analytic assessment are in areas of system self-assessment (understanding "normal" over time) and swarm assessments of a given situation or system.

The Challenge and Opportunity of Big Data

One of the most important dimensions of the cyberspace challenge is the need to assess massive amounts of data, often times to identify and if possible characterize minute changes of anomalous behavior. This is made more complicated by the rapid evolution of the cyberspace environment and the lack of full understanding of normal, abnormal, or anomalous behavior. While this is an important challenge, it is by no means unique to the cyberspace issue alone: many of the other scientific and other disciplines invited to C3E -- astronomy, behavioral sciences, biomedical informatics, computer sciences, financial fraud detection, and others – point to the big data problem as holding as much promise as it does challenge for the development of models and predictive analytics. Within this section, we first identify some general ideas about the application of Predictive Analysis to Big Data, followed by experiences drawn from the other scientific and commercial areas. We conclude with ideas specific to the CyberSecurity Mission.

Big datasets refer to large, dynamic collections of information that become problematic for traditional tools and methodologies to manipulate. One might consider trillions of records, loosely structured, and often distributed as characteristic of Big Data. Often times, there are few known complex interrelationships known about and within the data, and connections are usually probabilistically inferred. Further, unlike traditional analysis – where one formulates a hypothesis and then tests it – predictive analytics allows, indeed, demands the development of many multiple hypotheses, some pointed to by structures within the data.

Predictive analytics has shifted during the past decade from a relatively simplistic approach using computationally intensive statistics to one involving data mining, knowledge discovery, and other techniques. Predictive analytics employs the development of a large number of models based on the structure of any particular data set, and then combines and deploys them into the data set. The models can then be modified based on operational results and then continuously repeated.

Today, gaps in computational capability exist in dealing with big data, especially in terms of capacity and speed. Current High Performance Computing Initiatives might not address all needs in this area. Related to performance is the need understand the power consumption aspects of predictive analytics as they are employed within systems. Large data centers, such as those employed by Google or Facebook, for example, require massive amounts of power for the kind of assessment that they can create in the commercial market.

Further, some decisions are tactical and require immediate calculation and response, such as those designed to do assessment, reaction, and response to an outside attack on a system. More strategic aspects of predictive analytics can be timed to coincide with slower cycle periods.

During the C3E Workshop, participants from other areas requiring predictive analytics in their business made observations about the emerging challenges associated with Big Data. Throughout the scientific disciplines, for example, data is growing exponentially and there is a convergence of physical and life sciences catalyzed through the use and understanding of Big Data. New approaches are needed that use both computing and statistics. Big Data sets result in a shift from hypothesis driven discoveries to data-driven ones. Scientists/analysts can now discover patterns or anomalies in the data and postulate a hypothesis to test the explanation. This is a shift in the analysis paradigm. Astronomy has always been data driven. In many cases astronomy data of interest is below the noise floor of the sensor. Real errors and covariances exist. New data intensive scalable computer architectures are needed for dealing with big Data. Storing and managing the large volumes of data is a major problem.

The astronomy community is conducting research in a number of relevant areas related to dealing with huge data sets and the ability to conduct analysis – including predictive analytics – within them, such as :

- Scalable architectures
    - Extreme database queries
    - Extensions of databases
- Data representation
- Characterization of phenomena (known and unknown)
- Discovery of phenomena
    - Rare outliers buried in noisy objects
- Comparing events and ensembles (stream processing)

Business also needs to deal with these phenomena in various ways, including development of practical applications that allow for instantaneous decision. Among them, the financial sector seems to be the most advanced in leveraging big data and predictive analytics for a wide range of activities from the new methods of fraud employed by criminals to allowing merchants to make instantaneous decision at point-of-sale about individual, and potentially fraudulent, transactions. Their approach includes Big Data sets to develop models and rules along with testing of hypotheses. These Big Data repositories also enable long term analysis. For real time processing, each user account is characterized be a small set of metadata parameters. This approach allows for scoring of each financial transaction in less than a second with the actual approve/reject decision made by the financial client. Another small dataset characterizes the current high priority risky devices for real time fraud monitoring. Cybersecurity researchers should examine these techniques as possible solutions for predictive analytics.

From a cybersecurity perspective, the data to be dealt with is messy, incomplete, corrupted, missing, and intentionally deceptive. Data from some domains changes so rapidly that static analysis is intractable. A recent cyber incident spanned a period of three days and involved hundreds of thousands of IP addresses. Flows of information estimated at 7 billion page views involving over 2500 servers exceeded 125Gbps of traffic.

The challenge for cyber is threefold: the internet is seemingly boundless and spans grow nearly instantaneously, individuals have the power of a nation state, and 80-90% of traffic flows over non-government networks. Given these facts, Workshop discussions focused on how to leverage Big Data as an asset to provide foresight into emerging threats. Yet important scientific, technical and intellectual developments must be navigated in order to employ predictive analytics effectively. Among the most important aspects discussed at the workshop was the need to create validated models, including an understanding of how those models interact and interoperate with others. These themes will be discussed below in the areas focused on emergent behaviour and intersecting anomalies.

The Emergent Behavior Track

The Emergent Behavior track addressed key issues:

- Data requirements for modeling cyber emergence
- Model interoperability and evaluation

Emergence models are important to predictive analytics for at least three reasons:

- Globalization and the increasing rate of knowledge sharing is escalating the asymmetric nature of threat vectors
- Analysts and policy makers increasingly ask for anticipatory techniques (what-if or possible futures to counter adversities and maximize opportunities) to aid decision making
- Emergence models help anticipate plausible future that have never happened previously (zero day attacks)

The Emergent Behavior track looked at various static and dynamic models for predictive analytics with an emphasis on emergence models since their structure changes over time. The advantage of this model is the future states of the system result from interactions from lower levels of the system. In particular, approaches of agent based models and multi-agent systems were presented. Brief discussions were presented on strong and weak emergence, with emphasis on the latter because of the notion of self-organization as a process of pattern formation rather than a strict conformity to direction from a higher level. Numerous examples of self-organization from biological swarming behavior were presented.

To model cyber emergence, the concept of digital ants was discussed. The analogy of ants searching for food (anomalous cyber behavior) and marking a trail (network topology) back to the food (location of the malicious activity). Ants searching food will follow these trails (source of the network malicious activity). As long as the food (bad behavior) exists the ants will follow the trail. This is a concrete example of how self-organizing processes can capture the emergence of cyber threats. Stigmergy (the indirect coordination between agents leading to modification of the agent's local environment) is the key to this self-organization.

From the discussion of biological stigmergy, three factors emerged for applying this model to the cyber domain. Each of these was discussed to understand the cyber environment context of the analogy. The three factors were:

- Topological structure of the environment
- Feature detectors leading to the discovery of network anomalies
- Hierarchical structure of agents

If the system can evolve dynamically using these factors, then it can stay ahead of the adversary. Using this model, emergent behavior can characterize both attackers and the defenders. Subject matter experts are key to developing the typology/taxonomy of cyber attacks. Given this approach, will humans trust this emergent cyber model?

With this as the basis, the track then proceeded to examine model interoperability, parameter calibration and evaluation.

Clearly, massive datasets facilitate the development and reliability of emergent models. Big datasets enable adjustments to the model based on actual operational experience. After the plenary discussions on big data, the track focused on a five level model of hierarchy offered by McCusker (Sonalysts) as a cyber taxonomy. This was used as a template for characterizing how agents at various levels of the hierarchy could communicate between levels to achieve self-organization for the cybersecurity domain.

As a result of all the discussions during the track session, the emergent behavior group had the following outcomes:

- Started a C3E emergent modeling community of interest
- Created a common understanding of base principles
- Established a working process for characterizing interaction across layers of cyberspace activity
- Defined base requirements and design for emergence modeling in the cybersecurity domain

The Intersecting Anomalies Track

The Intersecting Anomalies track used a three-phase working group approach to assess the following hypothesis:

By seeking the intersection of anomalies of multiple sensor kinds, we may be better able to identify situations of interest

The track was partitioned into three phases, covering anomaly identification, anomaly intersection, and the application of anomalies to big data. Ultimately, the teams developed six anomaly detection and prediction techniques and attempted to describe the following for given threat scenarios:

- For each scenario, consider how to create an analytic to combine the anomalies that would add clarity to the state and velocity of a threat
- Consider any predictive aspect of the anomaly intersection with regard to anticipating the threat's next action or the threat's action in response to a defensive action

Among the anomalies considered were ones at the system level to the user level, and involving malicious software, authentication failures, disguised/mis-represented transactions, beaconing and signaling as evidence of network surveillance, and others. Some anomaly intersections will indicate antecedents of other adversary actions. The combination of static and dynamic data sets were particularly interesting, as they were more expressive in the intersection of anomalies to provide greater context to detect the current and predict future state of a threat. This was similar to using offline data to enrich online anomaly detection. The adaption of anomalies over time seemed to indicate that self-learning data scoring systems were superior in detecting unusual or unexpected combinations of anomalies over the pre-disposition of expert knowledge.

As a result of their efforts, the intersecting anomaly track indicated that the original hypothesis was supported in several explicit ways, including that the hypothesis was supported in the identification of new situations of interest that might be identified by an intersection of features, none of which might be unusual or especially rare. Secondly we observed many types of anomaly intersections that apparently were better suited to identify situations of interest, most particularly the combination of anomalies derived from orthogonal datasets. Finally, we were able to identify many situations of interest that would be difficult to detect without the intersection of anomalies.

While each of the intersecting anomaly analytics described in this track were not meant to be directly implemented, there were a number of insights gained from the process of rapidly identifying these systems. The value of putting together anomalies that were not designed to work together initially was demonstrated through the identification of attack scenarios that would be much easier to identify and predict than without these anomalies. Future work could take some or all of these analytics and develop actual techniques that could prove effective at detecting a class of behaviors invisible to us today.

As a result of all the discussions during the track session, the Intersecting Anomalies group had the following outcomes:

- Started a C3E intersecting anomalies community of interest
- Created a common understanding of base principles
- Established six combinations of anomalous behaviour for further assessment from both a specific and conceptual basis
- Defined aspects or elements of a process for looking at intersecting anomalies as a basis for predictive analytics in cyberspace.

Conclusions and Next Steps

For the third consecutive year, under the sponsorship of the Office of the Director of National Intelligence and the National Security Agency, experts gathered to understand new ways to assess the human, data, and modeling aspects of a cyber environment that is compromised by persistent adversarial behavior. Predictive analytics, applied on a foundation of emergent behavior and intersecting anomalies, forms an interesting basis for continuing discussion and potential research attention in the pursuit of more secure computing in this complex environment.

While this environment will be very challenging, we believe it is a realistic portrayal of the future operating environment for academia, government and industry. Sophisticated techniques such as those considered here at C3E form the basis for both conceptual knowledge and practical applications, including how an improved description of what is normal and how to optimize the role of the analyst or operator relative to the machine, based on comparative competencies.

Massive, complex big data sources appear at first glance to be part of the challenge, but within this year’s continuing conversation about big data we found newer and more diverse approaches from other disciplines, as well as the leading edge of a scientific revolution – and its attendant analytic revolution(s) – that potentially transform the way we understand this and other complex problems confronting the Nation. And, as with the participants in the first two C3E workshops, we have found that modeling behavior of both human and machine activities forms the potential basis for predictive anticipation and warning of threatening developments in cyberspace. Both the emergent behavior looked at here and the potential for using multiple, intersecting anomalies as the basis for assessment will help improve those models and form the basis of response.

Calculating what to do and when to do it is, however, dependent on understanding much more than understanding the adversary's behaviour – it is highly contextual. As C3E continues to mature, we are increasingly thinking about issues that exist synergistically in the space between us and our adversaries, as opposed to thinking about them separately. We can and should use the power of computation, the big data and our understanding of networks to model a massive variety of phenomena in cyberspace, something we believe is essential to understanding the path to secure computing in the future.

For sure, the C3E efforts summarized here demonstrate that we are not without options for understanding the path to secure computing in the future. Our brief investigation provided the starting point for looking at how to use predictive analytics to understand the nature of the threat in cyberspace, in particular through the concepts of emergence and seeking a set of anomalies through which to conduct assessment. Predictive analytics will apply to an understanding of adversarial behaviour in cyberspace, whether normal, accidental, benign or otherwise, and the dynamic context that surrounds them. Given the time urgency of real-world developments in this area, we need to optimize the roles of humans and systems for what they do best, either by their very nature or by policy or system design.

As with many research workshops, the value comes not simply from the initial set of ideas put forward on a complex issue, but in the elaboration on that issue that comes from the continuing dialogue and critical assessment of ideas and approaches to mitigating or eliminating specific challenges. Going forward, our C3E website, http://www.c3e.info, will continue to provide the collaboration tools and ideas repository for contributors to share new and evolving approaches for advancing analytical cybersecurity and for maturing an enduring community.

We hope to gather again in 2012 to continue the conversation.

Appendix A: The Emergent Behavior Track Summary

Track Lead: Antonio Sanfilippo

The concept of "emergence", as used in modern science, usually refers to the complex behaviors that emerge from dynamic interactions between simple entities. Modeling emergent behavior in cyberspace can have a game-changing impact in the fight against cybercrime. The end of the 20th and beginning of the 21st Century have seen computers and networks permeate life and science, and embed us all into a new, strikingly complex system of computational networks. Such developments have broached groundbreaking advancements in the way we interact with infrastructure, institutions and other people, but have also created new vulnerabilities. Due to the ever growing complexity of computer networks, cyber threats quickly evolve so as to constantly present novel challenges. Simulating scenarios that embody the emergence of such challenges is the first step to developing a truly anticipatory approach to cyber defense.

Several algorithms have been proposed to simulate emergence using as reference processes occurring in nature such as the swarming behavior of insects, birds, fish and other animals. Some of these algorithms have been utilized to model cyber emergence with promising results. Further progress depends on how successfully we can address issues such as the following.

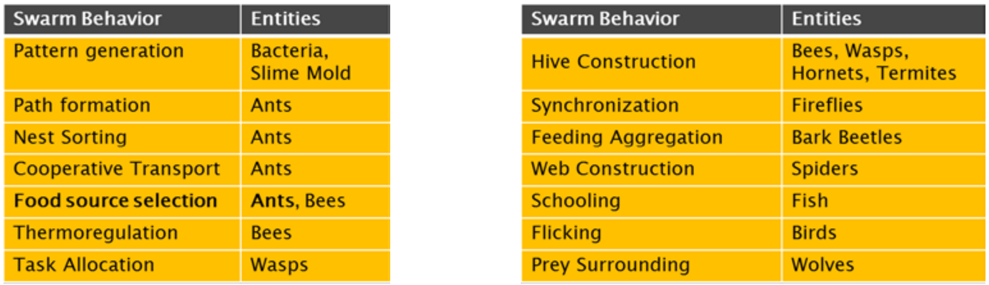
- Which are the best algorithms and natural-process metaphors to model cyber emergence?
- How do we harvest and calibrate relevant evidence from massive datasets to inform models of cyber emergence?
- How do we test the validity models of cyber emergence?
- How address interoperability questions at the levels of knowledge, algorithmic and operational integration?

Background

Globalization and the ever-growing rate of knowledge sharing keep escalating the asymmetric nature of threat vectors in the cyber world. Analysts and policymakers increasingly need predictive analytic techniques that help analyzing plausible future events to help counter adversities and maximize opportunities. Emergence models are particularly effective in predictive analytics as the help anticipate plausible futures that may have never occurred or been seen in the past.

Emergence models typically derive the future state of a system from multiple interactions among its lower level components. As typically invoked by multi-agent simulation and agent-based modeling algorithms developed in cognitive science and complex systems theory, these interactions are conceived as being fully self-organizational in nature so that the emergent property is ultimately reducible to its individual constituents (weak emergence). Alternative views have also been proposed, according to which the emergent property is irreducible (strong emergence) or only partly reducible (weakly strong emergence) to its individual constituents.

According to the notion of self-organization as understood in weak emergence, patterns at the global level of a system emerge solely from numerous interactions among the lower level components of the system that only use local information, without reference to the global pattern. There are numerous examples of self-organization in nature, as indicated in Table 1 with reference to the swarm behavior of animals.



Swarm Behavior	Entities
Pattern generation	Bacteria, Slime Mold
Path formation	Ants
Nest Sorting	Ants
Cooperative Transport	Ants
Food source selection	**Ants,** Bees
Thermoregulation	Bees
Task Allocation	Wasps

Swarm Behavior	Entities
Hive Construction	Bees, Wasps, Hornets, Termites
Synchronization	Fireflies
Feeding Aggregation	Bark Beetles
Web Construction	Spiders
Schooling	Fish
Flicking	Birds
Prey Surrounding	Wolves

Table 1 - Self-organization behaviors in nature.

Ant-based Model

Foraging patterns in ant colonies have been shown to provide a useful paradigm to model emergence in the cybersecurity domain [Haack et al. 2011]. Ants finding food mark the environment with pheromone trails during their return to the nest, and ants searching for food probabilistically follow these trails. Pheromone trails fade quickly overtime, unless they continue to lead to food. The ants find the shortest path to food, without communicating directly with one another and using signs left in the environment (e.g. in the form of pheromone trails) as guidance for action. For example, experimental evidence shows that when a colony of trail-laying ants is offered two equal food sources located at the end of two paths of different lengths, after sometime the vast majority of foragers converges on the shorter path [Beckers et al. 1993]. The insights of foraging patterns in ant colonies can be applied to model cyber-emergence using multi-agent simulations as follows.

- Food is equated to anomalous network activities
- The environment is formed by computer networks
- Digital ants look for anomalous network activities in the environment
- When anomalies are found, ants mark the network with digital pheromone trails
- Ants searching for network anomalies probabilistically follow these trails
- Trails fade quickly over time, unless they keep providing a source of anomaly.

Figure 1 provides a visual exemplification of this approach. The quadrants through which the “digital ants” move map into computer network components and the white arrows indicate the “digital trails” that the ants create in their pursuit of network anomalies. Trails (temporarily) lead to successful identification of anomalies, e.g. sudden spikes in network activities such as unusually high access requests to a database or service, which are represented by fire icons (a darker color indicates stronger anomaly.)

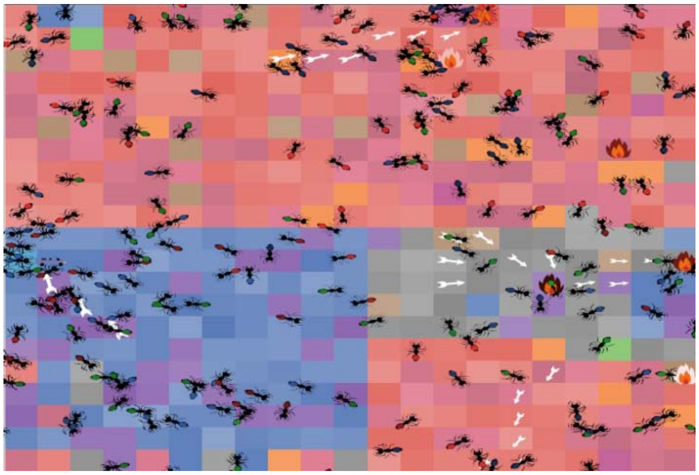
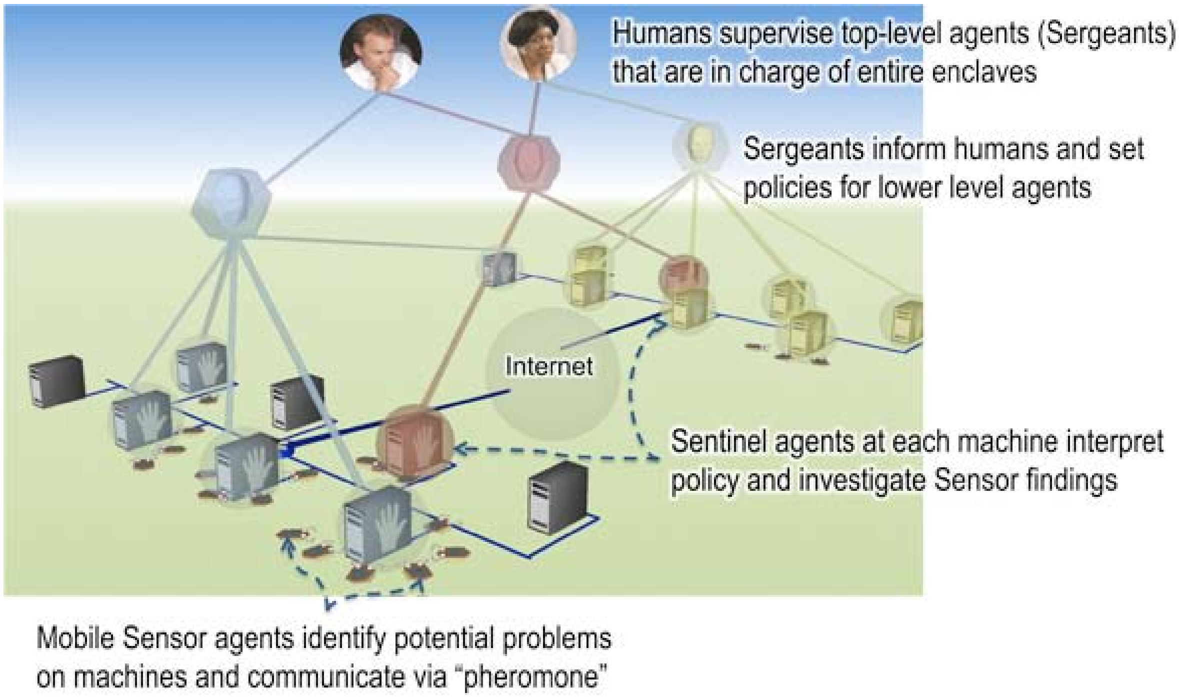


Figure 1: Multi-agent simulation that uses the foraging behavior in ant colonies as a paradigm to model cyber-emergence (adapted from Haack 2011).

The term *stigmergy* – derived from the Greek words στίγμα, stigma (“mark, sign”) and ἔ ργον, ergon (“work, action”) and originally introduced by Grassé (1959) – designates the concept that complex behavioral outcomes such as foraging patterns in ant colonies emerge from simple and indirect interactions among agents, mediated by the environment. An interesting elaboration of stigmergy is the possibility of organizing agents in a hierarchy. For example, in a multi-agent simulation such as the one described above (see Figure 1), we can define diverse classes of agent detectors with increasing levels of complexity and have them work cooperatively (e.g. complex detectors intervene when simple detectors spots a problem), as shown in Figure 2 where detectors are layered in the form of hierarchy of agents that detect anomalies at lowest layers and report possible attacks to humans at higher layer.



Humans supervise top-level agents (Sergeants) that are in charge of entire enclaves

Sergeants inform humans and set policies for lower level agents

Sentinel agents at each machine interpret policy and investigate Sensor findings

Mobile Sensor agents identify potential problems on machines and communicate via “pheromone”

Figure 2: Nested stigmergy in multi-agent Cybersecurity simulation, adapted from http://i4.pnnl.gov/news/digitalants.stm.

From the Ant-based Model to a General Model of Cyber Emergence

From the ant modeling sample discussed above, three factors appear to be central to modeling emergence in the cyber domain:

1. Topological structure of the environment (i.e. the network)
2. Feature detectors leading to the discovery of network anomalies
3. Hierarchical structures of agents.

Environment topology encompasses diverse types of interactions (e.g. network, social, behavioral) which lead to formation of parallel topologies. Parallel topologies enable concurrent modeling of the behavior of the attacker, the defender and other stakeholders, e.g. network agents that are hostile to the attacker or defender. Topologies are generated via interactions among lower level system components leading to self-organization. One important advantage of a self-organization approach to the generation environment topologies regards computational scalability and robustness [Angius et al., 2006]. Due to the logical simplicity of independent interaction as a mode of agency, stigmergic systems scale well to large numbers of agents and are robust against the loss of agents [Parunak, 2006]. This notion of environment topology opens a host of possibilities in how to define topologies where the attacker is constrained, e.g. creating controlled vulnerable topologies where we can observe the behavior of attacker.

One essential element of the stigmergic approach to modeling cyber emergence is that different "feature detectors" (e.g. detecting agents such as the digital ant in the ant model discussed above) must share information via the environment. Therefore, it is necessary to assess the robustness and veracity of detecting agents with reference to real world occurrences, e.g. to establish whether and to which extent detector agents can modify themselves to the point where they are ineffective or even subversive. One way to address this need would be to create a multitude of random detectors, and let them compete for survival in terms of their utility, i.e. using genetic or artificial immune system algorithms. Another important function to be determined for feature detectors is the rate of information transmission in the network environment.

Trusting Models of Cyber Emergence

An important question to address is how quickly humans can come to trust emergent cyber models based on algorithms that are inspired by biological processes such as the swarming behavior of social insects. While people are of course much more intelligent than social insects and do communicate, self-organizational processes that emerge from stigmergic activities have also been shown to be valid in characterizing human behavior. For example, Susi and Ziemke (2001), make a very convincing argument that stigmergy is germane to theories of social cognition such as activity theory, situated action, and distributed cognition. Akin to the notion of stigmergy, these theories focus on agent-environment interactions in providing a natural explanation for the "coordination paradox" that characterizes human activities that do not involve direct interaction and yet display cooperative behavior. Tummolini and Castefranchi (2007) provide a definition of stigmergy as the process of indirect communication of behavioral messages with implicit signals that can be generalized across any family of agents, including humans. According to this definition, the artifacts that agents leave in their trails and that trigger indirect communication are behavioral messages that inform about the presence of an agent (e.g. oneself), the opportunity for action, the result of an activity, and the agent's intention, ability, accomplishments and goals.

This novel human-to-human interpretation of stigmergy provides a game-changing paradigm for studying human self-organization. Specific examples discussed in the literature include processes such as traffic flow and trail formation, auction market systems, on-line auctions, elections, distributed document editing, status board, viral marketing, and intelligent highway systems [Parunak, 2006], as well as the development of open access information products (e.g. Wikipedia, WWW, recommender systems such as those in eBay and Amazon, Google page rank, open source software) [Parunak, 2006; Elliot, 2006; Heylighen, 2007; Robles et al., 2005]. Stigmergy has also recently been applied to security problems such as insurgency [Lugosky & Dove, 2011] and cyber security [Haack et al., 2011].

Interoperability and Evaluation

The pursuit of modeling cyber emergence through a stigmergic approach requires that issues of interoperability and evaluation be properly addressed to ensure operational adequacy and user acceptance.

Interoperability needs to be considered with reference to knowledge coalescence, algorithmic integration, and operational integration. Knowledge coalescence is concerned with the combination of heterogeneous elements of knowledge relative to cyber emergence to promote model fidelity and reliability. Algorithmic integration addresses the question of how to link across diverse modeling approaches (e.g. Bayesian nets, system dynamics, agent-based) in order to use the right tool for the right job and leverage legacy models of cyber emergence. For example, classification models of cyber anomaly (e.g. Bayesian nets) inferred from historical or simulation data can be used to calibrate agent properties (e.g. anomaly detection characteristics and thresholds) in agent-based models of cyber threat emergence. Operational integration focuses on user modeling and interfaces to enable analysts and policymakers to use models of cyber emergence interactively to test hypotheses and interventions through creative reasoning and competitive/collaborative work.

The evaluation of emergent cyber models requires answers to the following questions

- How do predictive cyber models compare with real life?
- How do we tease out emergent properties from predictive cyber models?
- What emergent properties generated by the model relate to cyber security? How do these properties relate to confidentiality, integrity and availability?

In term of prerequisites for model implementation, it is important to understand how to

- define requirements for emergence models
- generate concrete model implementations using hardware description languages (e.g. VHDL)
- move from hardware model descriptions to actual running systems.

For model evaluation, we need to establish

- how the concrete model satisfy the requirements
- how to perform V&V to map across concrete model and the actual running system
- how effectively does the behavior of system mimic real life.

Finally, can we predict how systems are designed, where they are headed, and their cyber security implications, and how do we decide which emergent properties predicted by the model are important for cyber security?

Big Data

As for any modeling endeavor, the use of massive datasets greatly facilitates and enhances the development and reliability of emergent models. Specifically to the nested stigmergy approach to emergence modeling we discussed above with reference to Figure 2, data are needed on observed and simulated self-organization patterns, and readjustment directives provided through feedback from users. For example, we may extract anomaly thresholds from network data to inform our low-level agents on what sort of potentially problematic events to flag, and then interactively drive the agents to focus on specific anomaly types to narrow down the search space, as indicated in Figure 3. Thereafter, from simulation to simulation, the same process can be repeated, using the simulated data as calibration source, as exemplified in Figure 4.

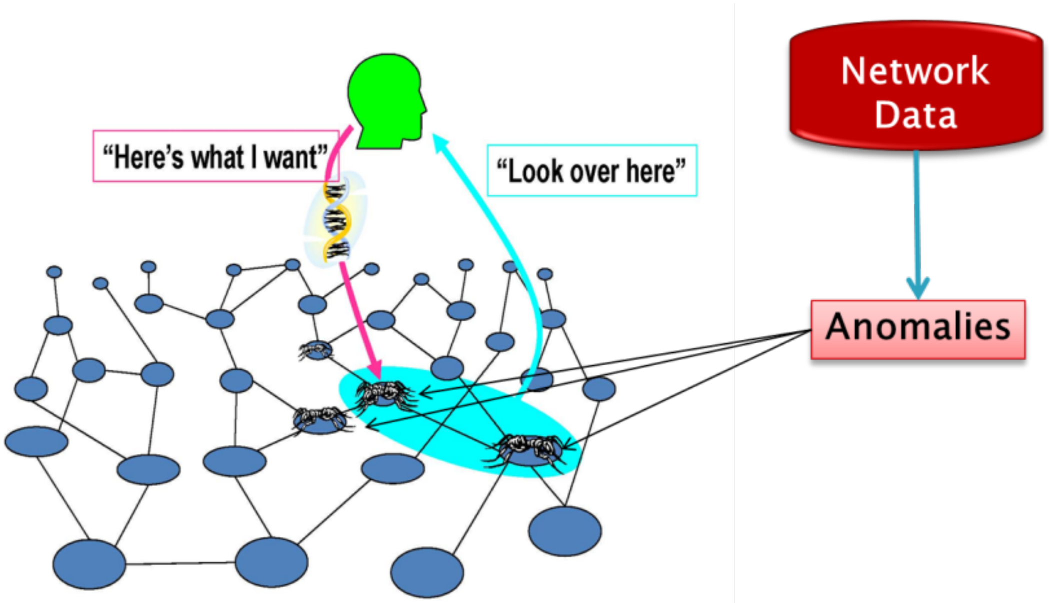


Figure 3: Use anomaly threshold values and dependencies learned from network data to direct agents' behavior in multi-agent simulations.

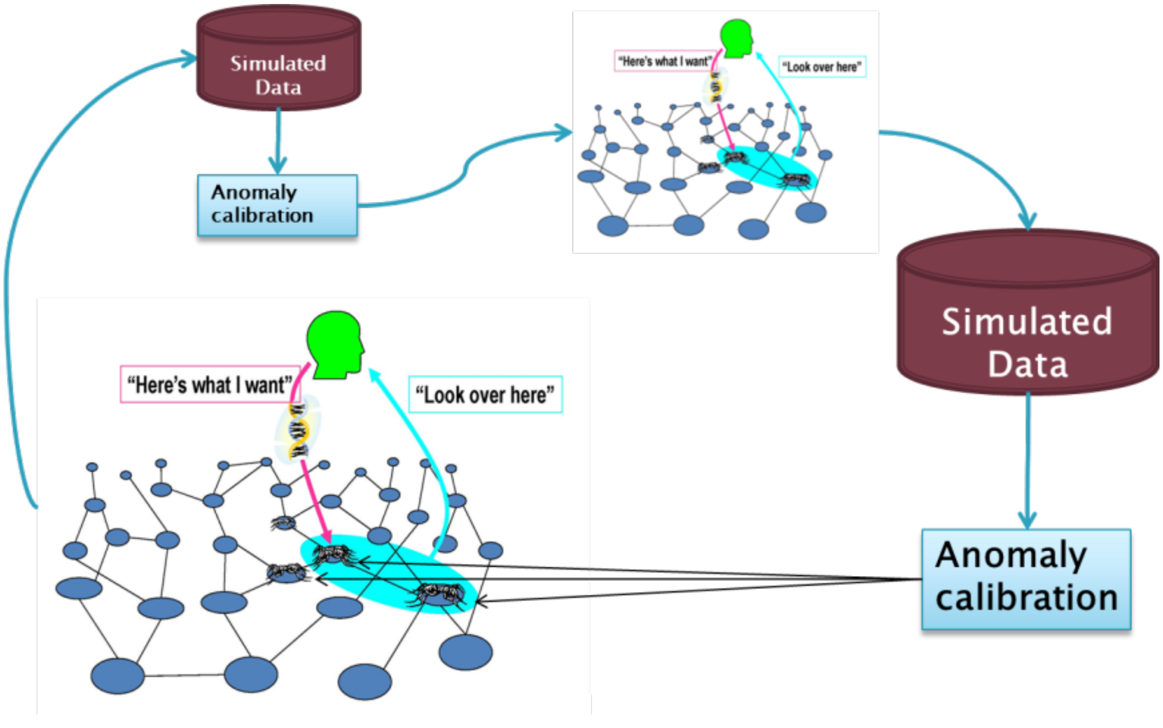


Figure 4: Use anomaly threshold values and dependencies learned from simulated data to direct agents' behavior in multi-agent simulations.

Putting It All Together

The nested stigmergy view of the approach to cyber emergence modeling that transpires from our discussion can be described as a multi-agent simulation environment where

- agents at the lower level can be instructed by the higher level agents or human users on how to directed their activities in their space
- the result at each simulation round is to remap the data into a new, higher-level space.

As exemplified in Figure 5, the ensuing framework is characterized by

- successive levels of refinement
- upward and downward information flow
- processes (emergent and otherwise) that transform one level into the next.

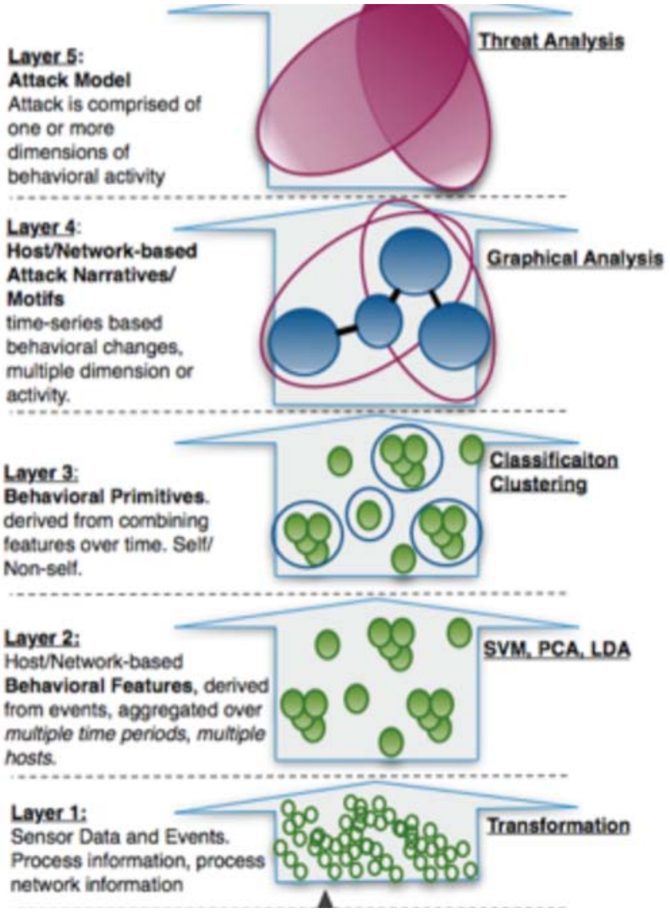
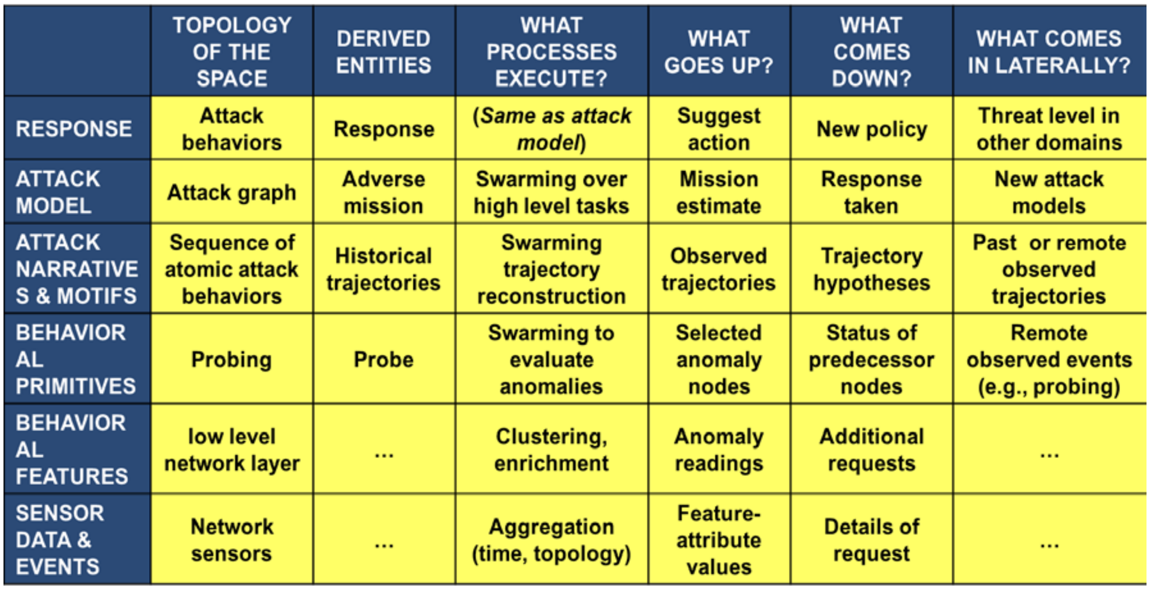

Layer 5:
Attack Model
Attack is comprised of one or more dimensions of behavioral activity
Threat Analysis
Layer 4:
Host/Network-based Attack Narratives/ Motifs
time-series based behavioral changes, multiple dimension or activity.
Graphical Analysis
Layer 3:
Behavioral Primitives.
derived from combining features over time. Self/ Non-self.
Classificaiton
Clustering
Layer 2:
Host/Network-based
Behavioral Features, derived from events, aggregated over multiple time periods, multiple hosts.
SVM, PCA, LDA
Layer 1:
Sensor Data and Events.
Process information, process network information
Transformation

Figure 5: Multilayer hierarchy in nested stigmergy.

Figure 6 provides a specification of the parameters and sample values for a modeling framework of this type. The bottom three topological sample values (probing, low level network layer, network sensors) and the bottom two processes (clustering enrichment, aggregation of time and topology) provide natural links to the activities of the C3E team on intersecting anomalies.



	TOPOLOGY OF THE SPACE	DERIVED ENTITIES	WHAT PROCESSES EXECUTE?	WHAT GOES UP?	WHAT COMES DOWN?	WHAT COMES IN LATERALLY?
RESPONSE	Attack behaviors	Response	(*Same as attack model*)	Suggest action	New policy	Threat level in other domains
ATTACK MODEL	Attack graph	Adverse mission	Swarming over high level tasks	Mission estimate	Response taken	New attack models
ATTACK NARRATIVE S & MOTIFS	Sequence of atomic attack behaviors	Historical trajectories	Swarming trajectory reconstruction	Observed trajectories	Trajectory hypotheses	Past or remote observed trajectories
BEHAVIOR AL PRIMITIVES	Probing	Probe	Swarming to evaluate anomalies	Selected anomaly nodes	Status of predecessor nodes	Remote observed events (e.g., probing)
BEHAVIOR AL FEATURES	low level network layer	…	Clustering, enrichment	Anomaly readings	Additional requests	…
SENSOR DATA & EVENTS	Network sensors	…	Aggregation (time, topology)	Feature-attribute values	Details of request	…

Figure 6: Sample parameters for a multilayer nested stigmergy approach to cyber emergence modeling.

References

Angius, Gianmarco, Cristian Manca, Danilo Pani, and Luigi Raffo (2006) Cooperative VLSI Tiled Architectures: Stigmergy in a Swarm Coprocessor. M. Dorigo et al. (Eds.): ANTS 2006, LNCS 4150, pp. 396–403, 2006. Springer-Verlag, Berlin, Heidelberg.

Beckers, R., Deneubourg, J. L. & Goss, S. (1993) Modulation of trail laying in the ant Lasius niger (Hymenoptera, Formicidae) and its role in the collective selection of a food source. J. Insect Behav. 6, 751^759.

Elliott, M. (2006) Stigmergic Collaboration: The Evolution of Group Work, M/C Journal, 9(2). Available at http://journal.media-culture.org.au/0605/03-elliott.php.

Grassé, P.P. (1959) La Reconstruction du Nid et les Coordinations Inter-individuelles chez Bellicosoitermes Natalensis et Cubitermes. La Théorie de la Stigmergie: Essai d'Interprétation du Comportement des Termites Constructeurs. Insectes Sociaux, 6, 41-81.

Haack JN, GA Fink, WM Maiden, AD McKinnon, SJ Templeton, and EW Fulp. 2011. "Ant-Based Cyber Security." In 8th International Conference on Information Technology : New Generations (ITNG 2011), April 11-13, 2011, Las Vegas, Nevada, ed. S Latifi, et al, pp. 918-926. IEEE Computer Society, Los Alamitos, CA. doi:10.1109/ITNG.2011.159.

Heylighen, Francis (2007) Why is Open Access Development so Successful? Stigmergic organization and the economics of information. In B. Lutterbeck, M. Bärwolff & R. A. Gehring (eds.), Open Source Jahrbuch 2007, Lehmanns Media.

Lugosky, Jena , Rick Dove (2011) Identifying Agile Security Patterns in Adversarial Stigmergic Systems. Insight 14 (2), International Council on System Engineering, July 2011.

McCusker, Owen (2010) Malicious Insider (MI) Behavioral Sensor Grid: A Pervasive Behavioral-Based Insider Threat Detection Capability Supporting Privacy Enhancement, BAA Proposal for DARPA CINDER Program (# DARPA-BAA-10-84), pp 2-3.

Parunak, H. Van Dyke (2006) A Survey of Environments and Mechanisms for Human-Human Stigmergy. In D. Weyns, H. Van Dyke Parunak, and F. Michel (Eds.): E4MAS 2005, LNAI 3830, pp. 163 – 186, 2006, Springer-Verlag, Berlin, Heidelberg.

Robles, G., Merelo, J.J. & Gonzalez-Barahona, J.M. (2005) Self-Organized Development in Libre Software: a Model based on the Stigmergy Concept, Proc. 6th International Workshop on Software Process Simulation and Modeling.

Susi, T. & Ziemke, T. (2001). Social Cognition, Artefacts, and Stigmergy. Cognitive Systems Research, 2(4), 273-290.

Tummolini, L., Castelfranchi, C. (2007). Trace Signals: The Meanings of Stigmergy. Lecture Notes in Artificial Intelligence, 4389, 141-156.

Appendix B: The Intersecting Anomalies Track Summary

Track Lead: Tom Longstaff

Within a compromised environment, anomalous behavior has been used to identify indicators of a wide variety of faults, both natural and malicious. Much of the R&D performed to address both the analysis of anomalies and the attribution of the behavior has focused on individual variations from specified or learned normal behavior along a single attribute or sensor (e.g., anomalous network traffic patterns). Human analysts often identify a series of intersecting anomalies, anomalies that may be related to the same behavior of interest, to gain a gestalt of the activity. These insights are particularly difficult to automatically identify and combine, but there may be techniques that could be applied if we consider the problem from a wide variety of potential sources intersecting about a common behavior. With our ability to handle "big data," we have the opportunity to discover anomalies from a wider variety of sources.

Intersecting anomalies need not be strictly passive, but where there are hypothesized intersections of anomalies, we may generate specific stimulus events that help converge anomalies and more easily associate many anomalies to common adversary behavior. For example, cluster analytics may be appropriate to partitioning these intersecting anomalies and help to guide the appropriate generated events within the compromised system toward convergence. Some questions related to this track are:

- What are the optimal sets of anomaly sources and types that can be combined to identify an imbedded adversary?
- What algorithms are appropriate in discovering the relationships between different anomaly types?
- How does big data help to enrich our use of anomaly identification and analysis? Can we use scale to our benefit?
- Do intersecting anomalies help us to uniquely identify adversary behavior when most of what the adversary does is identical to normal behavior?
- How can we "game" the system to help deconflict/cluster anomalies that clarify the normal and anomalous system behavior?
- Can we combine multiple detected anomalies at scale and near-real time to address attacks in the real world? What are the barriers to successful application of such an approach?

What we set out to do

The underlying hypothesis for the Intersecting Anomalies track is:

> *By seeking the intersection of anomalies of multiple sensor kinds, we may be better able to identify situations of interest*

To explore this hypothesis, the track was partitioned into three phases. Roughly, the three phases covered anomaly identification, anomaly intersection, and application of anomalies to big data.

In phase I, the track was partitioned into thirteen pairs of participants. Each pair was to describe two anomalies as completely as possible to include a description of the anomaly, the dataset that contained this anomaly, the interpretation of the anomaly (i.e., what does the anomaly imply about the state of the system), the threat vector that would likely cause such an anomaly, and the methods/analytics that would be used to detect the anomaly. The pair had 45 minutes to describe two anomalies as fully as possible given their experiences, and select one of the two to represent the work of the pair. The pair was to choose the anomaly that was most completely described and was most likely to combine with other indicators to predict the behavior of a threat.

In phase II, the track reformed into six groups of 3 or 4 participants each. These groups were selected by group leaders by selecting 4 of the anomalies from Phase I, such that all of the anomalies were selected by at least two groups, and no pair of individuals from phase I were together in the same phase II group. Each phase II group was required to use all of their selected anomalies to create a combined anomaly detection and prediction technique that incorporated each of them. The teams named each of their techniques and attempted to describe each of the following:

- The threat scenarios that may be indicated by the intersection of the anomalies
- For each scenario, consider how to create an analytic to combine the anomalies that would add clarity to the state and velocity of a threat
- Consider any predictive aspect of the anomaly intersection with regard to anticipating the threat's next action or the threat's action in response to a defensive action

The six teams were:

- PECAN
- Unmasking Coordination
- TERP
- Code Blame
- TVPA
- Endpoint Event Korrelation (EEK!)

In phase III, the group leaders from phase II led a discussion the included a description of their intersection anomaly analytic with a particular emphasis on the incorporation of "big data." The questions surrounding big data were:

- Can the detection be improved by big data?
- Does the detection scale to big data?

The first question attempted to evaluate each of the intersecting anomaly analytics fidelity. Does the system become more accurate and valuable with larger scales of data, or does the method create too many false indications with larger data sets? In all cases, the answer depended on implementation details of the individual implementations of the constituent anomaly systems more than the intersection of the anomalies. If the anomalies were more accurate with big data, the intersection was more valuable as well. In the case where some of the anomalies were sensitive to big data and others were not, the intersecting of the anomalies tended to improve the performance of the individual anomaly systems.

The second question addressed the computational complexity of the intersecting anomalies with regard to the size of the dataset. This particular question was broken down into two difficulties presented by scale to big data – the movement of big data through the communications infrastructure and the processing of the big data to discover and intersect the anomalies once there discovered. Each of the scenarios had some sensitivities to each of these difficulties, but could not be detailed in the time allotted.

Each of the intersecting anomaly analytics were then evaluated based on three criteria: Was the intersection effective, was it feasible, and is the intersection scalable. This was an attempt to summarize the questions arising from big data and from the description of each of the intersecting anomaly analytics. The track as a whole voted on each of the criteria after each of the group leaders presented their intersecting anomaly analytic. In addition, the track raised other issues that would need to be addressed in the implementation of this particular intersecting anomaly analytic. The results of this voting and brainstorming are described in Appendix D.

Observations from the Intersecting Anomalies Track

There were a number of observations and insights that were gained as part of this track at C3E. In particular, the creation of anomalies and intersections were very wide ranging, yet in each case this apparently random collection of anomalies could be put together into a plausible narrative. In each case, there were attack scenarios that appeared to take advantage of the combination of these collections of anomalies, even though they were not chosen to specifically detect these threats. This implies that taking advantage of a wide variety of anomaly detection systems in combination may hold significant promise in the identification of advanced threat that might evade any single anomaly detection system. This apparently supports the original hypothesis from the track.

In the identification of anomalies in phase I, anomalies were defined very broadly as simply some indication deemed interesting in the context of some multi-dimensional space. As they were simply features, they need not be "anomalous" in the semantic sense of unusual, but simply be features of interest. This helped to create a wide range of anomaly types that were not restricted to events of low occurrence. This observation is interesting in the context of most anomaly detection systems research, which mostly concentrates on unusual events rather than "features of interest." Focusing on features of interest, even if they are not unusual, may provide significant information to a detection system, as long as a mechanism such as the intersection of other features will extract actionable results.

The intersection of anomalies did help to prioritize among the individual anomaly detection systems. One observation was that by using multiple anomaly detection systems, we should be able to increase our ability to recognize false positives. In this case, the intersecting anomalies serve as context in the evaluation of individual anomalies to provide a better view of which anomalies are of interest. The subsets of intersecting anomalies may indicate classes of threats better than individual classification systems. This could also increase the detection rate of threats that can be observed in more than one dataset.

Some anomaly intersections will indicate antecedents of other adversary actions. This is the predictive capability that was hoped for in the formation of this track. Where the prediction is based on models, and intersection of anomalies can populate the models, there can be additional focus on downstream collection of specific indicators that would validate our predictions.

Another specific observation was that the combination of static and dynamic data sets were particularly interesting. These combinations tended to be more expressive in the intersection of anomalies to provide greater context to detect the current and predict future state of a threat. This was similar to using offline data to enrich online anomaly detection.

The adaption of anomalies over time seemed to indicate that self-learning data scoring systems were superior in detecting unusual or unexpected combinations of anomalies over the pre-disposition of expert knowledge. This observation held promise in the detection of not only zero day attacks against a single vulnerability (the goal of individual anomaly detection systems), but perhaps the detection of zero day attack classes. Different attack scenarios never before observed, predicted, or modeled.

Another observation was that the most value was gained in the intersection of anomalies arising from orthogonal data sets. These are sets with low statistical correlation and probably represent the best way to observe threat activity from multiple perspectives. The conjunction of anomalies from these orthogonal datasets tended to indicate signal (true detected anomalies) rather than noise.

Finally, there was an observation that even with intersecting anomalies, the analytic for intersecting and detecting these anomalies was still distinct from the response in our scenario. Each anomaly alone is not sufficient to trigger response, but the correlations of different anomalies yield high confidence that the intersection does represent threat activity. This provides an additional opportunity to be predictive, if you can stimulate the adversary to produce activity that would lead to additional intersecting anomalies. As always, there is a tradeoff here between understanding the attack behavior and stopping the attack proactively to reduce the severity of potential damage.

Conclusion

In conclusion, the intersecting anomaly track indicated that our original hypothesis, "by seeking the intersection of anomalies of multiple sensor kinds, we may be better able to identify situations of interest," was supported in several explicit ways. First, we expanded the notion of anomaly to simply be a feature of interest, the intersection of which was able to identify situations of interest. This immediately indicated that our hypothesis was supported in the identification of new situations of interest that might be identified by an intersection of features, none of which might be unusual or especially rare. Secondly we observed many types of anomaly intersections that apparently were better suited to identify situations of interest, most particularly the combination of anomalies derived from orthogonal datasets. Finally, we were able to identify many situations of interest that would be difficult to detect without the intersection of anomalies.

While each of the intersecting anomaly analytics described in this track were not meant to be directly implemented, there were a number of insights gained from the process of rapidly identifying these systems. The value of putting together anomalies that were not designed to work together initially was demonstrated through the identification of attack scenarios that would be much easier to identify and predict than without these anomalies. Future work could take some or all of these analytics and develop actual techniques that could prove effective at detecting a class of behaviors invisible to us today.

Appendix C: C3E Workshop Agenda

C3E C3E Computational Cybersecurity in Compromised Environments

Office of the Director of National Intelligence
National Security Agency
ODNI/NSA

Sunday, September 25

Afternoon – evening Participants arrive

5:30–7:00pm	Reception
6:30pm	Brief remarks by Brad Martin, Kevin O'Connell and Dan Wolf

Monday, September 26

7:30-8:30	Continental Breakfast	
8:30-9:00	Introduction and Welcome	Dan Wolf, Kevin O'Connell
	Brief review of C3E 2009 and 2010	
	Opening Remarks and Workshop Objectives	Brad Martin
	Detailed Review of Agenda	Kevin O'Connell
9:00-10:00	Cyberspace and the Role of Predictive Analytics	Patricia A. Muoio
10:00-10:30	Morning Break	
10:30-12:00	Setting the Scene: Threats and Analytics	Mike Fisk Scott Zoldi Owen McCusker
12:00-1:30	Lunch with Speaker	Dawn Meyerriecks
1:30-2:00	Track Introductions	
	What is Emergence?	Antonio Sanfilippo
	What are Intersecting Anomalies?	Tom Longstaff
2:00-5:30	First Working Track Sessions (includes Afternoon Break) Format – Separate Track Panels & Discussion Sessions Output –Discussion summaries, Reporter Notes	
5:30	Workshop adjourns for the day	

Tuesday, September 27

7:30-8:30	Continental Breakfast	
8:30-9:00	Summaries of Monday's Track Sessions	Antonio Sanfilippo, Tom Longstaff
9:00-10:00	Big Data	Alex Szalay, Bob Grossman
10:00-10:30	Data Intensive Computing for Intelligent Applications	Chris Oehmen
10:30–11:00	Morning Break	
11:00–12:00	Plenary Panel on Big Data Format – Group Review and Discussion Output – Group and Reporter Notes	Mike Bender, Alex Szalay, Bob Grossman
12:00-1:30	Lunch with Speaker	Gary M. Jackson
1:30-2:00	Introduction of Big Data in Emergent Behavior and Intersecting Anomalies	Antonio Sanfilippo, Tom Longstaff
2:00-4:30	Second Working Track Sessions – Consequences of Big Data (includes Afternoon Break) Format –Separate Track Panels & Discussion Sessions Output – Discussion Summaries, Reporter notes	
4:30-5:00	Points of Convergence and Divergence	Dan Wolf, Kevin O'Connell
7:00-9:30	Dinner with Keynote Speaker	Pat Lincoln

C3E C3E Computational Cybersecurity in Compromised Environments

Office of the Director of National Intelligence
National Security Agency
ODNI/NSA

Wednesday, 28 September

7:30-8:30	Continental Breakfast	
8:30-10:00	Summary Session	Dan Wolf, Kevin O'Connell
	Emergent Behavior	Antonio Sanfilippo
	Intersecting Anomalies	John Launchbury
10:00-10:30	Morning Break	
10:30-11:30	Advances in Cyberspace Predictive Analytics	DuskoPavlovic
11:30-12:00	Wrapping C3E from Inception	Brad Martin
	Where Do We Go from Here?	
12:00	Workshop Adjourns	