Statement of Research and Teaching Interests

Erik M. Ferragut
http://erikferragut.me
ferragutem@ornl.gov

December 2013

Research Interests

My main research interest is in improving the security and dependability of network, computer, and cyber-physical systems through the development and application of data analytics and machine learning. With the growing dependence of critical infrastructure (e.g., financial systems, energy grid, and medical devices) on connected networks, the potential damage from network- and computer-based attacks continues to increase. At the same time, sophisticated, highly motivated criminal organizations and nation states present a growing threat. Attackers consistently outmaneuver prevention, detection, and remediation defenses.

In my role as lead for research-operations integration at the Oak Ridge National Laboratory, I have seen that the challenges faced by real-world network defense analysts are not addressed by conventional analytic methods. For example, analysts are often overwhelmed by the warnings and alerts emitted by their intrusion detection systems. In practice, these alerts are ignored until after an attack has been discovered by other means; only then do they use the alerts to support a forensic investigation. New, fundamentally different and more effective approaches are needed.

My approach to improving security is based on improving the intelligence of the defense systems by developing new data analytics and machine learning algorithms for integrating available data and making more effective decisions in the presence of uncertainty. My first step is always to understand the real-world problem and its constraints. Next, I take the time to carefully reformulate the problem and develop a plan for validating the results in order to help ensure that the algorithms and systems produced will be practical and relevant. A common concern is that the system will only have access to incomplete, noisy data, but will need to infer some unknown information (e.g., presence of an attack). Only at this point do I develop and apply data analytics, typically extending existing methods to address the specific challenges of the problem. Through experience, I have learned that this approach helps ensure operational relevance, while still providing significant opportunities for scientific development.

My main areas of recent research are *anomaly detection* for network intrusion defense and *resilience analysis* for cyber-physical defense. I also worked on new machine learning algorithms for *schema matching* of heterogeneous structured data, which, although not a security problem, is relevant to security problems where disparate structured data sets must be integrated.

Anomaly Detection

Intrusion detection systems typically search for particular patterns associated with known malicious software (malware) and behaviors. However, these signature-based methods can only identify attacks that have been previously observed and for which a signature has been created. Authors of

new malware, who have access to the malware detection software, can modify and test their code to ensure it will not trigger an alert. Furthermore, malware authors have tools to automatically re-implement their software using different machine instructions to avoid detection; some malware even re-implements itself. Anomaly detection has been proposed as an alternative to signature-based methods since it can potentially detect attacks that have not yet been observed and analyzed. Sponsors, such as the Department of Homeland Security (DHS), Department of Defense (DOD), and the Intelligence Advanced Research Projects Activity (IARPA), are very interested in effective intrusion detection methods that are not based on signatures.

In practice, however, anomaly detection systems are not widely used for two main reasons. First, they typically produce too many alerts to be useful. In an attempt to detect as many attacks as possible, they are often designed to be highly sensitive. As a result, they produce a large number of false alerts. In cases where the sensitivity can be changed, a series of trial and error adjustments is required to achieve an acceptable alert rate. Second, since alerts are often predominantly false positives, the time it takes an analyst to decide whether to discard an alert dominates analysts' time. The alerts often lack the context analysts need to process them quickly. As a result of too many false alerts with too little context, analysts have been known to disable or ignore alert systems.

In my research, I developed a new theory for defining and constructing anomaly detectors based on rigorous probability theory that ensures that the rarest events necessarily will be the most anomalous, an intuitive notion not always satisfied by heuristic approaches. I reformulated the problem in the context of probability theory, and I defined a new anomaly detection protocol, called *meta-rarity* [3]. I then proved that meta-rarity is regulatable in that it is possible to set the sensitivity in advance so that the false positive rate will remain below a pre-specified, operationally feasible level. Moreover, I proved that given a correct probability model, the false positive rate will be exact wherever possible, a strong result previously only available for very restrictive cases. In my theory, anomaly detectors can be constructed for *any* probability distribution (outside of some unusual mathematical counterexamples not encountered in practice). Because the anomaly scores are defined in terms of tail probabilities, the scores can be meaningfully compared across different detectors. So, by constructing a large collection of complementary anomaly detectors, it is possible to get coverage over a wide range of possible anomaly types. The diversity of detectors provides an immediate source of context: the "reason" for the anomaly is given by which detector flagged it. My research provides the first of its kind theory for general construction of anomaly detectors with built-in regulatability of false positives and with direct support for providing alert context. The anomaly detectors I have proposed generalize the Mahalanobis distance, a common approach to anomaly detection that allows for false positive regulation, but only applies to Gaussian distributions and does not enable comparison of distributions on different dimensions. In contrast, my theory allows comparison not just across dimensions, but between Gaussian and non-Gaussian distributions [5].

I have also developed two methods for automatically generating anomaly detectors. This is important because having a broad collection of detectors can significantly increase the system's ability to detect unexpected anomalies. One approach I devised takes a detailed probabilistic description (model) of the data and creates an anomaly detector for each of its various parts, including marginal and conditional distributions. This approach is especially well suited to handling large, complex, and structured data [4]. Another method of automatic detector generation takes into account the potential multiscale properties of the data. For example, a machine's IP address can be grouped with others that play the same role in the network (e.g., other DNS servers), that have the same prefix, or that are assigned to the same Autonomous System. Behavior that may appear anomalous in one context will not necessarily appear anomalous in the others. I have proposed a data exploration tool that would be generally applicable to structured data and would facilitate

analysis at multiple levels of aggregation [1].

Although my focus has so far been primarily on developing the theory, I have also helped integrate the approach into a working prototype (called Situ) at ORNL. We discovered all of the major events as well as some artifacts of the data synthesis in the Visual Analytics Science and Technology (VAST) 2012 intrusion detection data challenge [9]. I also created a plan for applying the anomaly detection theory to physical systems [7].

The next steps for this research are to (1) develop algorithms to discover anomalies at multiple scales, (2) apply the methods to more real networks and cyber-physical data, and (3) augment the existing methods with computational tools. I have been funded as PI for a competitively awarded Laboratory Director's Research and Development (LDRD) project at ORNL for October 2013 to September 2015 ($390K per year) to develop anomaly detection specifically for complex networks exhibiting behavior at multiple scales. One useful project, suitable for an undergraduate student, is to implement a tool for computing anomalies within multiple levels of data aggregation. This data exploration tool would enable quicker turn-around on analyses of structured data, and it would also support the simultaneous LDRD work toward incorporating multiscale behavior. Another useful project would perform a direct comparison of my theory (meta-rarity) with common non-probabilistic anomaly detection approaches such as Local Outlier Factor.

Cyber-Physical Security

A controller is a device that sends signals to actuators to keep a physical system's state close to a prescribed state. Control systems manage many common and critical systems, such as factories, nuclear power plants, energy distribution networks, communications networks, vehicles, and medical devices. Control systems' increased Internet connectivity enables improved management and reduced maintenance costs. However, the combination of network attacks and control theory on physical systems creates a new kind of security risk that is only beginning to be understood. In addition to network-based attacks, which is traditional cyber security, and physical perturbations, which is traditional control theory, there is a third category: network-borne physical effects on the control system. For example, an attacker might exploit network access to create a fake sensor signal leading to a physical meltdown. My interest is specifically in this third category, an area of research also of great concern to DHS and the Department of Energy (DOE). This research has been in collaboration with Seddik Djouadi (University of Tennessee) and Alex Melin and Jason Laska (ORNL).

First, I took the intuitive terms of state awareness and operational normalcy, which are the constituent parts of resilience, and provided a mathematical framework to characterize them [10]. The more rigorous approach allows for quantitative measures of whether resilience is being achieved, a new contribution to the field. This measurement motivates a systematic characterization of attacks that considers the effect of loss of confidentiality, integrity, and availability separately on each communication channel within a control system and the resulting impact on system resilience.

Second, I focused on the impact that denial of service attacks on control signals can have on resilience [8]. In a common set up, the controller has a number of channels with which it sends actuation signals to the system being controlled. If some of these channels are disabled (i.e., an availability attack) then the system may lose controllability, a technical concept describing whether the controller can eventually move the state to an arbitrary reference state. Controllability and its dual concept, observability, are foundational concepts in control theory.

By analyzing the algebraic properties of a linear control system, I was able to derive a simplified characterization of controllability, reproducing a known result, but possibly with a new proof. I then showed how to reduce the question of post-attack controllability to a combinatorial problem.

I showed that this problem is NP-complete by relating it to a set covering problem but that an approximate answer can be computed quickly. This research enables a new method for computing the resilience of a system's controllability and observabilty against particular availability attacks. By applying the results to sample design problems, I showed that my new resilience measurement can be used to design physical control systems that are more resilient to network and computer attacks.

Third, I analyzed various sensor perturbation attacks, considering both finite energy and bounded attacks. I showed that for finite energy attacks the optimal attack (in the technical sense of linear quadratic control theory) requires a "kick" to the system where the energy is concentrated in one or more short time intervals when the system is most vulnerable [2]. The insights resulting from this research impact security analysis of worst-case scenarios. They also show the significant advantage gained by an adversary who knows the system dynamics and the control algorithm. These results outline an approach to building high-impact attacks.

In future work, I plan to explore the central role played by the advantage provided by information in adversarial control theory. In particular, the defender's knowledge of the system dynamics may enable more accurate attack detection, whereas better knowledge for the attacker may allow for highly effective, but subtle, attacks. One project a student could undertake immediately would be to implement my attack and detection ideas on an existing ORNL control system test bed. This project would validate the theory in a real system, suggest new directions, and provide a useful demonstration.

Schema Matching

Technological advancements have made data collection and storage easier and more affordable. By making connections between data sets, researchers hope to reveal new insights: connecting a customer database to a demographic data set could provide ideas for finding new customers; connecting a pharmaceutical catalog to prescription histories could enable automatic checking for unwanted drug interactions; and connecting financial transactions with social network information could lead to detecting insider trading. However, identifying similar fields between databases remains a daunting task when the data sets are collected by different organizations at different times for different purposes. The names of the data fields and how the records are organized into tables (i.e., the database schemas) can vary considerably from one dataset to another. Existing methods for schema matching require domain experts to painstakingly compare data and field names between the data sets in search of field matches.

In collaboration with Jason Laska (ORNL), I proposed, developed, and validated an automated approach to schema matching. Rather than base the matching on field names, which are often misleading, I focused on the values within the database. These values, thought of as strings of characters, exhibit characteristic patterns depending on the field type. For example, names are generally comprised of letters and spaces, whereas addresses often begin with digits and tend to be longer. Previous attempts to use these differences involved manual construction of a number of features followed by classifier training. The features one might develop, such as the number of digits in the first five positions and whether the length is greater than ten, were created manually based on domain experience. The classifier required having matched schemas (i.e., ground truth) to train on. Requiring ground truth for training is burdensome and often infeasible.

In my solution, I created three classes of nonparametric Bayesian probabilistic models. A separate instance of each model class was trained on each field in the database. It was then possible to use simple rules of inference to determine whether two fields come from the same model or not [6]. My models provided consistently better matching capability in comparison to more traditional models,

including parametric models and models trained using maximum likelihood estimation.

The resulting solution has had significant impact. It has been implemented and used by ORNL to match federal medical data with state data in support of the Affordable Care Act. This work was sponsored by the Department of Health and Human Services (DHHS) Center for Medicare and Medicaid Services (CMS). The results were favorably received, and CMS has sponsored my proposed follow on work: a $1M task to apply the same probabilistic modeling approach to identifying and reporting data quality issues. In particular, I will be combining these probabilistic models with my anomaly detection theory to detect data anomalies under the hypothesis that data quality issues, such as a phone number in a name field, will be anomalous.

Currently funded work in this research area is to develop more advanced models that better identify particular data quality issues. I also plan to create new models for field dependencies and multi-record anomalies, which will enable the discovery of more complex types of data errors. Continued application to medical data will be used to validate the approach. In addition, I believe the ideas of probabilistic modeling for schema matching and data quality will have applications in security. One promising direction is in reconciling the malware records for different antivirus systems, which will contribute to better understanding the malware space. Another interesting application is to match schemas across network and host sensors to support automated integration of new network security data, which would enable a more extensible defense intelligence system.

Summary and Future Directions

I have made theoretical and practical contributions to a number of research areas. These research areas are tied together by the unifying goal of improving the security and dependability of network, computer, and cyber-physical systems. I foresee that data analytics research and problem-driven applications will continue to be mutually supporting threads for future work. One promising area of future research will be to use topological data analysis (TDA) to explore local-to-global structure relationships to address data analytics and control theory problems, which may support attack discovery on the power grid. I am beginning to develop and apply multiscale data analysis methods to dynamic graph data with the goal of analyzing ORNL's power distribution data. I have also begun collaborating with Fernando Schwartz (University of Tennessee) on developing a TDA-based classifier. Another direction is to apply multiscale analysis to learning control system dynamics from observations, which is the challenging task of system identification. Existing methods for system identification tend to produce highly complex models. I hypothesize that topological data analysis, because it infers global structure from data, will be a powerful tool for producing high fidelity but low complexity system approximations, provided that it can be applied to this new context. I am currently collaborating with Seddik Djouadi (University of Tennessee) to pursue this.

A second direction of future work is to explore the relationship between anomaly detection and one-class classification. In anomaly detection, events that are rare are flagged as anomalous. In one-class classification, a classifier is constructed given only examples of one class, and then new examples are classified as being from that class or not being from that class. By equating anomalies with the other classes, it is possible to view both methods as computing the same thing. However, the one-class classifier perspective provides a clear direction for validation while my anomaly detection theory provides significant theoretical advantages. Understanding the relationships between these two perspectives may allow for combining the benefits of each. Furthermore, it may provide a principled way to validate anomaly detection in cyber security, which is currently a controversial issue. In conclusion, I plan to pursue a number of opportunities for improving security through better data analytics and machine learning.

Teaching Interests

I have been fortunate to be able to teach at two universities and mentor at multiple laboratories. My first teaching experience was as a Ph.D. student at the University of Michigan, Ann Arbor where I was solely responsible for a section of a Pre-Calculus course. I later had the opportunity to teach cyber security in the graduate school of the University of Maryland University College, an online open university. I learned that on-line courses can make quick student-instructor communication difficult to achieve, but they allow each student far more time to thoughtfully participate in the course. I am interested in combining the benefits of on-line courses with the courses I will teach.

In addition to teaching, I have also had the pleasure of mentoring three post-doctorates, a post-masters, a post-baccalaureate, five undergraduates, and a high-school teacher. I enjoy working with mentees to plan their research. In the case of undergraduates, I found that a very clear purpose and direction is important. With post-doctorates, the challenge is to give them enough direction to make sure they can contribute quickly and meaningfully while still encouraging creativity. There is a special moment I treasure when the student becomes a collaborator, when we are both working at the whiteboard, trading a marker back and forth as we propose solutions and work out counterexamples. These exchanges can open the student's eyes to appreciate that there are questions that nobody knows the answers to and that their ideas are important. I am re-invigorated by these discussions, and the resulting insights have even led to published results.

My experience across four government and industrial computational science laboratories will be tremendously useful in guiding students' academic and career development. Students I have met are very curious about what it is like to apply their coursework in work environments. They wonder about big picture questions, such as which topics are the most important, as well as details, such as how many hours researchers work and whether people tend to work in groups. Not only can I answer students' questions, but I can also identify areas where modifying the curriculum could help prepare students with more in-demand skills. Employers I have worked with increasingly demand computer security and data analysis. With professional research spanning these areas, I can provide relevant scientific, engineering, and management experience to help shape the student experience.

I am most interested in teaching security and data analytics, broadly defined. This includes cyber security, data analysis, machine learning, cryptography, quantum computing, information theory, and coding theory. I would be especially interested to teach a course on existing solutions for cyber security problems focusing on data analytics, unsupervised learning, and anomaly detection. Such a course would clarify connections between cyber security, machine learning, and data analysis.

One of my teaching goals is to combine coursework with student competitions. High profile competitions exist in cyber security, programming, and data mining. However, competing in these competitions requires significant preparation for most students. By combining that preparation with a credited course, the university will be able to field a better team, and the students will gain more from the experience. Such an application-driven course would complement traditional topic-driven courses by pulling from, motivating, and inter-relating a variety of different topics. Competitions depend on teamwork, speed, and a breadth of skills. These are exactly the capabilities that employers demand. I believe that a good showing at one of these competitions can significantly raise a computer science department's profile, especially with employers. Student competitions will support the university's ambitions by strengthening student camaraderie, bringing attention to the department, and highlighting student talent.

To summarize, I offer a university a unique perspective with my professional experience in research laboratories. I have enjoyed teaching and mentoring in the past. I look forward to expanding my teaching experiences to new cyber security and data analytics courses, and eventually to include a competition preparation course.

References Cited

[1] Bogdan Denny Czejdo, Ferragut, Erik M, John R Goodall, and Jason Laska. Network intrusion detection and visualization using aggregations in a cyber security data warehouse. *Int'l J. of Communications, Network and System Sciences*, 5(29):593–602, 2012.

[2] Seddik M Djouadi, Alexander M Melin, Ferragut, Erik M, Jason A Laska, and Jin Dong. Finite energy and bounded attacks on control system sensor signals, 2013. Preprint.

[3] Ferragut, Erik M. Meta-Rarity: Operationally relevant anomaly detection, 2013. Preprint.

[4] Ferragut, Erik M, David M Darmon, Craig A Shue, and Stephen Kelley. Automatic construction of anomaly detectors from graphical models. In *IEEE Symposium on Computational Intelligence in Cyber Security (CICS), 2011*, pages 9–16. IEEE, April 2011.

[5] Ferragut, Erik M, Jason Laska, and Robert A Bridges. A new, principled approach to anomaly detection. In *11th International Conference on Machine Learning and Applications (ICMLA), 2012*, volume 2, pages 210–215. IEEE, 2012.

[6] Ferragut, Erik M and Jason A Laska. Nonparametric Bayesian modeling for automated database schema matching, 2013. Preprint.

[7] Ferragut, Erik M, Jason A Laska, Bogdan D Czejdo, and Alexander M Melin. Addressing the challenges of anomaly detection for cyber physical energy grid systems. In *Proceedings of the Eighth Annual Workshop on Cyber Security and Information Intelligence Research*, 2012.

[8] Ferragut, Erik M, Jason A Laska, Seddik M Djouadi, and Alexander M Melin. Quantitative cyber-physical resilience: Metric and algorithm, 2013. Preprint.

[9] Lane Harrison, Jason Laska, Riley Spahn, Mike Iannacone, Evan Downing, Ferragut, Erik M, and John R Goodall. situ: Situational understanding and discovery for cyber attacks. In *IEEE Conference on Visual Analytics Science and Technology (VAST), 2012*, pages 307–308. IEEE, 2012.

[10] Alexander M Melin, Ferragut, Erik M, Jason A Laska, David L Fugate, and Roger Kisner. A mathematical framework for the analysis of cyber-resilient control systems. In *6th International Symposium on Resilient Control Systems (ISRCS), 2013*, pages 13–18. IEEE, 2013.