SANDIA REPORT
SAND 2009-0805
Unlimited Release
Printed February 2009

Mathematical Challenges in Cybersecurity

Daniel M. Dunlavy, Bruce Hendrickson, and Tamara G. Kolda

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94AL85000.

Approved for public release; further dissemination unlimited.

Sandia National Laboratories

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

NOTICE: This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government, nor any agency thereof, nor any of their employees, nor any of their contractors, subcontractors, or their employees, make any warranty, express or implied, or assume any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represent that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government, any agency thereof, or any of their contractors or subcontractors. The views and opinions expressed herein do not necessarily state or reflect those of the United States Government, any agency thereof, or any of their contractors.

Printed in the United States of America. This report has been reproduced directly from the best available copy.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone:	(865) 576-8401
Facsimile:	(865) 576-5728
E-Mail:	reports@adonis.osti.gov
Online ordering:	http://www.osti.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd.
Springfield, VA 22161

Telephone:	(800) 553-6847
Facsimile:	(703) 605-6900
E-Mail:	orders@ntis.fedworld.gov
Online order:	http://www.ntis.gov/help/ordermethods.asp?loc=7-4-0#online

SAND2009-0805
Unlimited Release
Printed February 2009

Mathematical Challenges in Cybersecurity

Daniel M. Dunlavy and Bruce Hendrickson
Sandia National Laboratories
P.O. Box 5800, Mail Stop 1318
Albuquerque, NM 87185

Tamara G. Kolda
Sandia National Laboratories
P. O. Box 969, Mail Stop 9159
Livermore, CA 94550

Abstract

This white paper is a response to a recent report on cybersecurity submitted to the U.S. Department of Energy (Catlett, 2008). We discuss what we see as some of the major mathematical challenges in cybersecurity. The document is not intended to be comprehensive, but rather the articulation of a few key themes. We have organized our thoughts into three challenge areas: modeling large-scale networks, threat discovery, and network dynamics.

1. INTRODUCTION

Several recent reports have highlighted the vulnerability of our nation's computer and communication infrastructure (PITAC, 2005; Goodman and Lin, 2007). Hardly a week passes without some new attack or compromise being reported in the national news. There is an urgent need to improve software quality and security policies and practices, and to remain ahead of the rapidly increasing capabilities and sophistication of cyber criminals and adversaries. A significantly more secure cyber world requires transformative changes in how we make and use computer systems.

But transformation requires a degree of understanding about the cyber environment that is currently lacking. As in many other fields, we believe that mathematics must play a key role in improving our understanding and helping to design alternatives. The complex behavior of networks emerges from the interplay of comparatively simple components, which is the hallmark of a complex system. Advanced mathematical models are essential to understanding and taming this complexity. Mathematical ideas must also play a key role in sifting vast data streams to identify emerging threats.

While other agencies are addressing immediate threats, we propose that DOE's research take the long view of tackling fundamental research problems in the cybersecurity domain, as suggested by the PITAC report (2005):

> We urgently need to expand our focus on short-term patching to also include longer-term development of new methods for designing and engineering secure systems. Addressing cyber security for the longer term requires a vigorous ongoing program of fundamental research to explore the science and develop the technologies necessary to design security into computing and networking systems and software from the ground up. Fundamental research is characterized by its potential for broad, rather than specific, application and includes farsighted, high-payoff research that provides the basis for technological progress.

In this white paper, we discuss what we see as some of the major mathematical challenges in cybersecurity. The document is not intended to be comprehensive, but rather the articulation of a few key themes. We have organized our thoughts into three challenge areas: modeling large-scale networks, threat discovery, and network dynamics.

2. MODELING LARGE-SCALE NETWORKS

One of the major challenges of the Internet or any complex network is the requirement for explanatory and predictive models. Such models enable us to:

- Simulate realistic networks at different scales for testing algorithms for network defense,
- Detect anomalies that don't conform to the model and thus potentially represent an intrusion or other network problem,
- Compress or extract portions of the network graph to make it amenable to advanced analysis that cannot be performed at larger scales, and
- Sample real-world networks in statistically meaningful ways (the standard statistical assumption of independent and identically-distributed random variables generally does not hold on graphs).

For the Internet, there are numerous graphs of interest, including the underlying physical network, the network of hyperlinks between web sites, the (dynamic) network of actual communications, and even email communication networks through which viruses can spread. Some common measures of a graph that a model would seek to emulate are the following (Chakrabarti and Faloutsos, 2006):

- Distribution over the entire graph of node in-degrees and out-degrees, which typically conforms to a power law distribution or some variant. Other distributions that have been used include the distribution of the entries of the leading eigenvector (or singular vector) of a matrix representation of the graph.
- Graph diameter, i.e., the greatest distance between any two nodes in the graph. Other closely related measures are the average distance between node pairs and the *effective graph diameter*, which is the greatest distance that connects most of the graph (say, 90%).
- Community structure, i.e., how is the graph organized. A typical specific measure is the number of *triangles* in the graph, though this is difficult to compute. It is also possible to incorporate geographic information into the community structure.
- Evolution of any of the above measures over time. For example, some real-world studies have noted that the diameter shrinks as the graph evolves.

These are just a few examples of graph measures, and we expect that the Computer Science community will continue to develop more measures as needed for different applications. Many of these measures are difficult to compute for large-scale graphs and present algorithmic challenges in their own right. However, the larger question is how to create a model with a relatively small number of parameters that captures a given set measurements from a real-world graph. Thus far, most research has focused on models that capture *just one attribute* of the graph, such as degree distribution.

One of the most interesting graph generator methods developed so far is R-MAT (Chakrabarti et al., 2004) because it simultaneously considers multiple measures. Unfortunately, the methods that have been developed thus far for fitting the R-MAT model to real data are numerically inadequate (based on unpublished studies at Sandia and elsewhere) and the model still has several shortcoming (in-degrees and out-degrees of nodes are highly correlated, which is not always the case in the real-world). Another promising model is based on hierarchical structure (Clauset et al., 2008) and can be used to detect anomalies; however, the model is not scalable to large graphs and detecting anomalies is numerically infeasible for all but very small problems.

Given an assumed underlying model of a network, we need methods for comparing two networks and measuring their similarity. Ideally, two networks generated via the same parameters would be detectable as such. In this way, we can compare real-world networks to determine if they are similar or not. This has application in detection of malicious sub-networks on the Internet.

Another issue with network models is that they are usually iterative, and so generating a large-scale network from a given model can be expensive. One class of approaches starts with a very small graph (2-4 nodes) and then grows, perhaps doubling the number of nodes at each iteration. Other approaches start with a set of unconnected vertices, and add edges one at a time. Since graph construction can be part of the process of fitting a model, efficient methods are needed for accelerating the process.

We have thus far assumed that our networks are entirely homogeneous, i.e., all nodes and all links represent simple objects. However, more realistic models would further include attributes on the nodes and edges. For example, it we are modeling the Internet, some nodes might represent routers while others represent computers. How does the modeling change when there are different types of nodes?

And we have only alluded to the problem of time-evolving networks. If we consider network traffic, for example, each edge could represent a single message passing between two computers. Thus, at any given instance, the graph would be different depending on which computers are communicating. Moreover, computers come on- and offline. What is an anomaly in such an evolving network? Can we describe “normal” behavior over time?

3. DISCOVERING CYBER THREATS

The discovery of threats in computer networks is a critical task in the area of cybersecurity. The sheer volume and heterogeneity of the data, the temporal nature of evolving threats, and the skew between normal and malicious behavior all pose major challenges in the analysis of computer network data. Solving the following problems is key to developing a long-term proactive approach to cybersecurity.

- *Malicious code detection* – Detection methods beyond simple signature detection are required for long-term proactive detection and analysis. Moreover, methods are needed that can identify mutations or variations of malicious code with high accuracy and low false positive rates.
- *Malicious behavior detection* – Methods are needed for aggregating information (locally and across networks) to detect complicated multi-stage attacks, analyzing systems for identification of potential vulnerabilities, and detecting rare events.
- *Malicious code attribution* – Methods are needed for determining the source of malicious code or behavior through analysis of network topology and/or traffic, but this is extremely difficult in the presence of IP spoofing, a large number of compromised machines, mutating malware, etc.

The most pressing mathematical challenges associated with the problems listed above center around the issues of adaptive data modeling (e.g., Stokes et al., 2008; Fries 2008; Ahmed et al. 2007; Stolfo et al. 2007), large-scale network data analysis (e.g., Kopek et al,. 2007; Vallentin et al,. 2007), and modeling temporally evolving network data for real-time analysis and prediction (e.g., Chan et al,. 2006; Seleznyov and Mazhelis, 2002). Commercial methods for detecting malicious code and/or behavior typically use rule-based or statistically modeled signatures for identifying threats. New methods and modeling techniques designed to facilitate real-time adaptation are needed to transition from a reactive to a preventative approach to threat detection. Such adaptation requires modeling of temporal data, modeling in the presence of missing data (missing either by error or by design in the case of changes in policies, standards, or protocols), and modeling of data that is not readily available on purely local systems (e.g., in the case of global attacks).

In the case of large-scale data (i.e., traffic data from large networks in real-time), analysis may only be tractable through the use of data sampling. This leads to uncertainties that must be propagated through the entire modeling and analysis pipeline to best inform security analysts and decision makers. Furthermore, such sampling methods would need to adapt to the changing characteristics of the network (including physical changes such as additions of nodes and software changes based on policy updates), malicious code, and malicious behavior.

In the case of temporally evolving data analysis, data structure or features may change over time, and models must be able to adapt to such changes. Useful or discriminating features in network data or malicious code may change over time, requiring dynamic modeling and thus dynamic techniques for determining how long data is useful for solving particular tasks. Research into the temporal analysis of discrete changes (e.g., network topologies, policies, protocols, malware) will also be required to facilitate such dynamic modeling.

Pertinent areas of mathematics include machine learning, mathematical optimization, statistics, probability, and linear/multilinear algebra. These areas, along with specific tasks identified with each area are listed below.

- *Machine Learning*
  - Online learning methods for dynamic modeling of network data and malware.
  - Modeling data with skewed class distributions to handle rare event detection.
  - Feature selection/extraction for data with evolving characteristics.
- *Optimization*
  - Large-scale optimization for graph searching/analysis and model parameter estimation (e.g., in learning models and descriptive analysis models).
  - Sub-optimal optimization in the presence of data sampling or other uncertainties.
  - Optimization under uncertainty applied to discrete models and/or data.
- *Statistics & Probability*
  - Data analysis in the presence of missing values.
  - Modeling uncertainty: reliability and risk versus noise and sampling error.
  - Sampling streaming or distributed data, and determination of how and to what extent such samples differ from real data.
- *Linear Algebra*
  - Updating matrix- or tensor-based models when data is added or changed.
  - Simultaneous (joint) modeling of heterogeneous data types and relationships.
  - Nonnegative models that may lead to better interpretation.
  - Tensor models for analyzing complicated, multi-way relationships in data.

4. NETWORK DYNAMICS AND CYBER ATTACKS

Many cyber attacks work by spreading malware to a large number of vulnerable machines. While the details may vary (e.g., whether a human needs to be tricked into making a mistake, or the propagation happens automatically), this style of attack expands along linkages in a social or technological network, infecting some fraction of nodes as it goes. For these kinds of broad-target attacks, rapid propagation is important since cyber defenders are likely to add protections once the malware is detected and characterized.

Improved mathematical models are needed to understand the spread of infections on networks. There is a whole domain of modeling questions associated with the evolution of cyber threats. For example, we need to determine what properties of the network and the malware determine the eventual limits of the infection. Likewise, patches may be propagating at the same time, so we need to determine how these competing processes interplay. Mathematically, this requires the evolution of coupled equations over time on a complex network structure, but the structure of the network is not known in detail and may be evolving itself. We need to model this lack of knowledge and understand its consequences for predictions. Ideas from game theory or dynamical systems may help defenders anticipate properties of future threats and consequently alter the interplay between attackers and defenders. Novel statistical methods are required to make reliable predictions. Researchers have developed powerful models for studying the spread of infectious diseases, and lessons from these models undoubtedly have relevance in the cyber domain.

This particular example reflects the need for improvements in our understanding of complex phenomena. Global properties emerge from the interactions of relatively simple networked components, and we need to understand the evolution of such properties. Insights will likely involve ideas from graph theory, dynamical systems, complexity theory, and the emerging field of complex networks. Models of varying scale and fidelity will be useful for different purposes, and multiscale methods will be necessary for efficiency. Within the reality of limited knowledge, we need to quantify and bound uncertainties in the outputs of the model. Uncertainty quantification has received considerable attention in recent years for physical and engineering simulations using partial differential equations, but it is not at all clear how to apply existing methods to the complex topology of networks, and to mathematical models that may not be founded on differential equations.

A related topic is the need to make decisions within a network, even though some nodes may be untrustworthy. For example, a cyber early warning system could involve detectors on dispersed machines that work collaboratively. If some of these machines become compromised, can the larger system continue to serve its function? The mathematics of Byzantine agreement and distributed systems has much to say about such problems, but these ideas need to be extended to situations of direct relevance to cyber security (Ryan, 2001).

A final challenge that permeates many aspects of cyber security is the enormous difficulty of writing error-free software. We believe that mathematics has an important role to play in reducing the incidence of software (and hardware) errors. Recent advances in automated model checking and the application of formal reasoning methods have had a major impact on the design of hardware (Clarke et al., 1999). New programming methodologies need to be developed that exploit similar ideas to improve software quality and robustness.

5. CONCLUSIONS

The challenges associated with cyber security are profound, and game-changing advances require research into new domains. The insight and rigor that mathematical ideas can provide are critical to progress on many of the hardest problems. In this paper we introduce mathematical areas that we believe to be of relevance. In summary, we have discussed the following areas.

- **Modeling large-scale networks**
  - Development of sophisticated mathematical network models that accurately emulate real-world networks, including extension to networks with attributes on the nodes and edges and to time-evolving networks.
  - Statistical techniques for comparing networks.
  - Methods for efficiently computing or estimating graph characteristics such as diameter.
  - Methods for compressing graphs or discovering interesting sub-networks.
  - Optimization and statistical methods for parameter fitting in network models.
  - Statistical numerical methods for likelihoods of an observation given a particular model.
  - Numerical acceleration methods for generating instances of a network from a model.
- **Discovering cyber threats**
  - Online learning methods for dynamic modeling, modeling data with skewed class distributions, and feature selection for data with evolving characteristics.
  - Large-scale optimization and sub-optimal optimization in the presence of data sampling or other uncertainties, and optimization under uncertainty applied to discrete models and/or data.
  - Data analysis in the presence of missing values and modeling uncertainty.
  - Sampling of streaming or distributed data.
  - Efficient and effective algebraic model creation and updating.

- **Network dynamics and cyber attacks**
    - Modeling the spread of infections and on complex networks.
    - Game theoretic or dynamical systems techniques for the evolution of cyber threats.
    - Mathematical models for the emergence of global behavior on networks.
    - Uncertainty quantification of network models.
    - Extending the mathematics of distributed systems to cyber security problems.
    - Mathematical techniques for validating software to reduce coding errors.

ACKNOWLEDGMENTS

We would like to thank Philip Kegelmeyer and Ali Pinar for helpful comments and suggestions.

6. REFERENCES

T. Ahmed, M. Coates and A. Lakhina, *Multivariate Online Anomaly Detection Using Kernel Recursive Least Squares*, Proc. IEEE INFOCOM 2007, pp.625-633, May 2007.

C. Catlett (ed.), *A Scientific Research and Development Approach to Cyber Security*, Report submitted to the U.S. Department of Energy, December 2008.

D. Chakrabarti and C. Faloutsos, *Graph Mining: Laws, Generators, and Algorithms*, ACM Computing Surveys, Volume 28, March 2006.

D. Chakrabarti, Y. Zhan, C. Faloutsos, R-*MAT: A Recursive Model for Graph Mining*, SIAM Data Mining, 2004.

J. Chan, J. Bailey, and C. Leckie, *Discovering and Summarising Regions of Correlated Spatio-Temporal Change in Evolving Graphs*, Proc. Sixth IEEE International Conference on Data Mining, pp.361-365, December 2006.

E. Clarke, Jr., O. Grumberg, and D. Peled, *Model Checking*, MIT Press, 1999.

A. Clauset, C. Moore and M. Newman. *Hierarchical structure and the prediction of missing links in network*s. Nature, Volume 453, May 2008.

T.P. Fries, *A Fuzzy-Genetic Approach to Network Intrusion Detection*. Proc. GECCO Conference Companion on Genetic and Evolutionary Computation, pp. 2141-2146, 2008.

S.E. Goodman and H.S. Lin, Editors, Computer Science and Telecommunications Board, *Toward A Safer and More Secure Cyberspace*, Washington, DC: National Academies Press, 2007.

C.V. Kopek, E.W. Fulp, and P.S. Wheeler, *Distributed Data Parallel Techniques for Content-Matching Intrusion Detection Systems*, Proc. Military Communications Conference, 2007.

PITAC, *Report to the President, Cyber Security: A Crisis of Prioritization*, President's Information Technology Advisory Committee February 2005.

P.Y.A. Ryan, *Mathematical Models of Computer Security*, In Foundations of Security Analysis and Design: Tutorial Lectures, R. Focardi and R. Gorrieri (eds.), pp. 1-62, Lecture Notes in Computer Science, 2171, Springer-Verlag, 2001.

A. Seleznyov and O. Mazhelis, *Learning Temporal Patterns for Anomaly Intrusion Detection*. Proc. ACM Symposium on Applied Computing, pp. 209-213, 2002.

J.W. Stokes, J.C. Platt, J. Kravis, and M. Shilman, *ALADIN: Active Learning of Anomalies to Detect Intrusions*, Microsoft Research Technical Report MSR-TR-2008-24, March 2008.

S. Stolfo, K. Wang, and W. Li, *Towards Stealthy Malware Detection*, in Malware Detection, eds. M. Christodorescu, S. Jha, D. Maughan, D. Song, and C. Wang, pp. 231-249, 2007.

M. Vallentin, R. Sommer, J. Lee, C. Leres, V. Paxson, and B. Tierney, The NIDS Cluster: Scalable, Stateful Network Intrusion Detection on Commodity Hardware, RAID 2007