SAND2013-2535P

Counter-Adversarial Data Analytics (CADA) Proposal No. 13-1376

Investment Area: Research Challenges

Project Intent: Discover

Duration: 2 years

Classified: N

Principal Investigator: W. Philip Kegelmeyer, 08900

Project/Portfolio Manager: Curtis Johnson, 05635

1 Overview/Abstract

1.1 Problem Statement

Sandia makes critical use of data analytics in defense of national security. Our adversaries therefore seek to sap, even suborn, those analytics. Through understanding our methods, they seek to produce data which is evolving, incomplete, deceptive, and otherwise custom-designed to defeat our analysis. Further, we cannot prevent them from doing so. We live in a changed world, in which we frequently *must* depend on data over which our adversaries have unprecedented influence.

In this LDRD, then, we will develop and assess novel data analysis methods to counter that adversarial influence. We will also generate implementations and at least one prototype deployment in support of a Mission Challenge.

We thus face paired problems. First, we must do data *science*, discovering generalizable and quantifiable counter-adversarial principles. Second, our national security mission requires methods that are *relevant*, applicable to analytics that matter, with realistic assumptions, useful uncertainty assessments, and practical implementations.

1.2 Creative and Innovative Nature of R&D

Being simultaneously scientifically rigorous and practically relevant seems challenge enough. Adding the fact that we are also trying to counter dedicated, agile, intelligent adversaries who will closely observe and learn from our behaviors makes this a high risk LDRD indeed.

Yet, if successful, the payoff will be immense. The ability to anticipate and defeat attacks on the data analytics at the heart of, for example, cyber security, stockpile assurance, counter proliferation, and biotechnology for threat detection is critical to a range of Sandia missions.

Further, though high risk, Sandia is positioned for success. We have staff expertise in machine learning, statistics, game theory, and network analysis. We have prior experience in attacking our own algorithms. We have direct access to pertinent national security data. And we have sufficient existing integration, knowledge and mutual respect among our research and application organizations to have already successfully transitioned ideas from research results to operational changes.

2 Proposed R&D

2.1 Technical Approach and Leading Edge Nature of Work

Basic Terms and Background One of the most powerful, subtle, and broadly applicable data analysis methods is machine learning. Indeed, machine learning's broad applicability is what recommends it for attention here, as important aspects of this LDRD's activities are developing *broadly* generalizable methods of dealing with adversarial attention, and *applying* them to at least one Mission Challenge.

For the sake of a concrete running example, consider the detection of malicious executables, "malware", as they enter a network. Most modern networks have analytics for extracting various features from an arriving executable in order to classify it as benign ("goodware") or as malware. These analytics often employ *supervised machine learning*, which examines training data, "groundtruth" examples of goodware and malware, to generate rules for classifying future unknown executables.

For additional insight, it is often helpful to use the descriptive features to simply cluster executables that are similar, even without *a priori* labeling as goodware or malware. This is *unsupervised machine learning.*

Classic machine learning methods depend on two assumptions violated by adversarial attention[5]:

- First, that the groundtruth is indeed *true*, that it has not been poisoned ahead of time by deceptive samples, or tampered with afterwards by adversarial access. That is, if an adversary knows we are gathering training data to build a malware detector, it behooves them to inject, now, harmless executables which are otherwise nearly identical to the malware they are planning.
- Second, that the eventual real data is *statistically similar* to the training data. Unfortunately, as adversaries will try to learn our classification model and modify their new malware to evade it, future data will necessarily drift in nature from the training data.

Research Directions For Counter-Adversarial Data Analytics We propose three complementary approaches to addressing those violated assumptions: robust, predictive, and dynamic defenses.

Robust: Perhaps the simplest idea is to increase the robustness of the learning algorithm via methods such as regularization, minimization of worst-case loss, and compensation for "concept drift"[4]. However, these techniques do not explicitly account for the adversarial nature of the data discrepancies. Better would be to *exploit* the attacker/defender relationship. For instance, we have previously developed a bipartite graph-based "transfer learning" methodology which enables information concerning previous attacks to be effectively used against *novel* attacks[1]. In this LDRD we will extend and generalize that approach.

In a complementary activity, we can harden machine learning models by improving the quality of their training data, specifically through detecting and addressing mislabeled "truth". We have already demonstrated the ability to detect and correct *accidentally* mislabeled truth[8]. We also have anecdotal results indicating that adversarial attempts to systematically mislabel data can paradoxically draw additional attention to the masked samples. In this LDRD we will solidify and generalize this finding.

Predictive: A more reactive approach is to develop predictive defenses, in which future attack strategies are explicitly anticipated and preemptively countered. Game theory[7] provides the basic concepts and formal mathematics necessary to analyze such repeated attacker/defender interactions. As an example, our recent work combined game theory and machine learning to predict the way adversaries would modify test data in "feature space"[3].

However, standard game-theoretic solutions are often impractical, due to the explosion of the solution space and the resulting computationally infeasible calculations and optimizations. In this LDRD we will investigate two complementary approaches to circumvent this difficulty:

- Rather then applying game-theoretic analysis to the full space of attacker actions, we will instead attempt a novel model of the attacker from within the reduced machine learning feature space, yielding an aggressive yet lossless information abstraction.
- Rather than focus on the predictions generated by the model, we will focus on the operational response. Our approach will be to develop a tractable[6] utility-maximizing model of the attacker *and* metrics to assess the deviations perhaps induced by that model.

Dynamic: Unfortunately, explicitly modeling the attacker, and thereby tuning our defense to that model, inevitably gives a wily attacker the ability to manipulate our defenses. A mitigating and potentially very powerful response is to adopt a dynamic, "moving target" perspective. That is, if instead of a single malware classifier we switch among a dozen such detectors, then an adversary requires twelve times as much work to reverse engineer our defense. If we can easily generate an infinite number of such classifiers, then suddenly the "asymmetric warfare" benefits the defender.

We have previously developed an approach where the coevolutionary relationship between attackers and defenders was leveraged to derive a moving target defense scheduling policy which is optimal or nearly optimal[2]. As a bonus, by explicitly modeling the incomplete information available to the attacker, we demonstrated that a defender who makes uses of that information superiority for the present round of defense may thereby give away information and lose that advantage in future rounds.

One of the difficulties in dynamically switching defenses is that all defenses cannot be equally effective. So one is occasionally weaker than necessary in the moment in order to have an overall stronger defense.

Modern ensemble machine learning methods point to possible mitigations, however. We will pursue:

- Computing the optimal randomization over K classifiers, given that they do not yield the same accuracy.
- Dividing the feature set into subsets when generating the classifiers.
- Generating dynamic, optimally switched weighted voting schemes across all classifiers. Using all classifiers but changing the weights may make an "infinite" set of defenses possible.

Demonstration Applications, and Initial Data So far we have spoken of our data analytics as abstractly divorced from application. That is often a productive stance, but not always. Precisely because the Mission Challenges (MC) are *challenges*, we expect their counter-adversarial data analysis needs to be particularly challenging as well; likely the specifics of the underlying problems cannot be so easily or productively abstracted over.

Mission Challenge Engagement We will therefore start by closely engaging with the stakeholders behind each of the Mission Challenges, to survey the range of data analytics currently used by each, and to consider the possible nature of adversarial tampering with those analytics. The twin goals are to:

- Finish with a deep and broad understanding of Mission Challenge analytics, expressed as a table indicating where a given CADA research thrust might address a Mission Challenge need.
- Select a single Mission Challenge and conduct a trial deployment of CADA methods in its support. This will require collecting data, building an implementation, and defining the success metrics.

We do not start from scratch. The proposal team already has some awareness of or involvement in, for instance: the use of unsupervised machine learning in counterintelligence in support of a "Safe, Secure Stockpile", the data analytics employed by many biosensors likely to be pertinent to any attempt to "Reduce Global Biological Dangers", and the *many* cybersecurity data analysis methods that will help "Enable the US to Operate Effectively in Cyberspace".

Initial Data In the second half of the LDRD we will focus on the Mission Challenge we have selected, based on criteria such as evidence of adversary adaptation, data availability, and impact. To make progress from day one, however, we have already identified and acquired two surrogate data sets suitable for initial focus. We have, from the Sandia Forensic Analysis Repository for Malware (FARM), roughly 100K truthed goodware and malware executables, date-stamped over a three-year period of acquisition, described both statically and behaviorally. We also have transaction data for a popular e-commerce auction site, including roughly 54M transactions over a seven-year period, with a small number of fraudulent sellers identified.

2.2 Key R&D Goals, Objectives, and Project Milestones

2.2.1 Goal and Success Measure

The overall technical goal of the project is to develop generalizable methods for countering adversarial attacks on national security analytics, demonstrated by quantitatively assessed defense of the analytics underlying at least one Mission Challenge. Success would result in methods and assessments publishable in key journals *and* in practical changes in how Sandia conducts its analytics.

2.2.2 Key Objectives and Milestones

Milestone	Completion
Objective: demonstrate quantitative utility to a Mission Challenge	
Circulate survey of MC analytic use as SAND Report	09/30/2013
Select a single MC problem for FY14 focus	09/30/2013
Acquire and validate data in support of MC focus problem	12/31/2013
Define MC test structure and required architecture	12/31/2013
Implement appropriate robust, predictive, dynamic defenses	06/30/2014
Conduct tests of defenses on MC focus problem, document in SAND report	09/30/2014

Objective: advance robust defense	
Extend bi-partite transfer learning approach to malware detection	06/30/2013
Quantify the effects of random and adversarial poisoned "truth"	06/30/2013
Conduct systematic evaluation of semantic proximity methods for poisoned truth detection	09/30/2013
Develop and test "adversary-aware" clustering	12/31/2013
Objective: advance predictive defense	
Implement a model of attacker actions in the machine learning feature space	09/30/2013
Implement a model, and metrics, for a "utility maximizing" attacker	12/31/2013
Test machine learning feature space attacker model against standard competitors	03/31/2014
Objective: advance dynamic defense	
Compute optimal randomization over K heterogeneous-in-accuracy classifiers	09/30/13
Develop feature subset switching defense	12/31/2013
Develop randomly weighted ensemble weighting defense	03/31/2014
Test both defenses against each other and against optimal static methods	06/30/2014
Analyze the value of "information revelation" entropy for imperfect information attackers	06/30/2014

2.3 Technical Risk and Likelihood of Success

- Game theoretic approaches often generate solutions which are impractically intractable.
  - We are investigating aggressive abstraction approaches here via the novel idea of embedding attacker actions in the machine learning feature space.
  - We are investigating a novel and simpler "utility maximizing" attacker model, while also generating assessment metrics to insure that we have not unrealistically simplified the problem.
- In 1.5 years it may be difficult to both advance data science *and* demonstrate practical utility.
  - We have staged our data and development resources to permit both immediate technical progress against initial data and eventual development against real data.

3 Resources

3.1 Key Research Team Members

Name	Org	FTE	Role
Tim Shead	1461	0.40	development and applications (in FY14)
Warren Davis	1461	0.20	machine learning analysis and development
Kristin Glass	5624	0.25	GT/ML research and development
Jeremy Wendt	5632	0.20	machine learning analysis and development
Richard Colbaugh	5635	0.20	GT/ML research and development
Brian Jones	6131	0.20	machine learning analysis and development
Yevgeniy Vorobeychik	8953	0.25	GT/ML research and development
Ken Chiang	8965	0.20	data access and application integration
Philip Kegelmeyer	8900	0.30	PI, machine learning analysis
David Zage	9516	0.35	machine learning analysis and development

Staff FTE levels will change in FY14; in particular, the PI will increase his commitment to 0.5 FTEs.

3.2 Qualifications of the Team to Perform This Work

Philip Kegelmeyer is a Senior Scientist with many years of machine learning experience. His recent work as PI of the Networks Grand Challenge LDRD has acquainted him with Sandia's national security analysts and relevant data analytics. Rich Colbaugh has recently contributed to methods that combine game theory, machine learning, and dynamical systems theory to enable effective analysis in the presence of adversaries. Kristin Glass has expertise in developing analytics for co-evolving systems and has successfully applied these

tools and methods to various specific national security problems. Yevgeniy Vorobeychik has expertise in game theory, particularly in the context of cyber security and for improving machine learning robustness in the face of an adversary. Tim Shead has significant recent experience implementing scalable scientific analysis tools and machine-learning algorithms, and a broad grasp of relevant data science support technologies. The rest of the team all have recent Sandia experience in developing and implementing learning algorithms.

4 Strategic Alignment and Potential Benefit

4.1 Relevance to DOE and National Security Missions

National security analysis must contend with adversarial manipulation of the data environment. National security analytics increasingly rely on open source data, and smart infrastructures such as the power grid are being tuned to their environments. All this necessarily creates vulnerabilities to manipulation of those data environments. Sample challenges for the DOE include network and data security, operational security and counterintelligence, and stability of the power grid. The DHS requires robust counterterrorism and biodetection analytics, and the DoD's dependence on air and ground combat, physical security, overhead sensing, and missile defense all require automated decisions that outwit adversarial algorithms.

4.2 Anticipated Outputs and Outcomes

The anticipated outcome of this R&D is new capabilities for defending Sandia's national security data analytics against adversarial action. Specific outputs will include technical publications, software implementations, and at least one quantitatively assessed deployment in support of a Mission Challenge.

4.3 Programmatic Benefit to Investment Area, if Successful

Data analysis permeates Sandia's many missions. We have a long history of adroitly addressing noise, error, bias and other "unintelligent" data challenges. We have much less experience in dealing with adversarial tampering with our analytics. Success in this project will both increase our ability to defend those analytics and help us understand the realistic limits of that defense. This will have practical benefit for Sandia's Mission Challenges and help ground further work in the Data Sciences Research Challenge with a nuanced understanding of adversarial attacks and how to counter them.

4.4 Communication of Results

We will engage the Mission Challenges through discussions with Mission Challenge stakeholders, a mid-project "survey of Mission Challenge analytics" SAND report, and a final SAND report documenting the test deployment. We will communicate our technical advances through publication, targeting the journal *Machine Learning* and the conferences *KDD* and *NIPS*. If appropriate, we will file Technical Advances to cover any new intellectual property generated.

References

[1] Colbaugh, R., and Glass, K. Proactive defense for evolving cyber threats. In *IEEE International Conference on Intelligence and Security Informatics (ISI)* (2011), IEEE, pp. 125–130.

[2] Colbaugh, R., and Glass, K. Predictability-oriented defense against adaptive adversaries. In *IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (2012), IEEE, pp. 2721–2727.

[3] Colbaugh, R., and Glass, K. Predictive defense against evolving adversaries. In *IEEE International Conference on Intelligence and Security Informatics (ISI)* (2012).

[4] Globerson, A., and Roweis, S. Nightmare at test time: robust learning by feature deletion. In *ICML '06: Proceedings of the 23rd international conference on Machine learning* (New York, NY, USA, 2006), ACM Press, pp. 353–360.

[5] Laskov, P., and Lippmann, R. Machine learning in adversarial environments. *Machine Learning 81*, 2 (11 2010), 115–119.

[6] Letchford, J., and Vorobeychik, Y. Optimal interdiction of attack plans. In *International Conference on Autonomous Agents and Multiagent Systems* (2013). To appear.

[7] Peters, H. *Game theory: A Multi-leveled approach*. Springer, 2008.

[8] Shoemaker, L., Banfield, R. E., Hall, L. O., Bowyer, K., and Kegelmeyer, W. P. Detecting and ordering salient regions. *Data Mining and Knowledge Discovery 22*, 1–2 (2011), 259–290.