Report from Dagstuhl Perpectives Workshop 12371

Machine Learning Methods for Computer Security

Edited by

Anthony D. Joseph[1], Pavel Laskov[2], Fabio Roli[3], J. Doug Tygar[4], and Blaine Nelson[5]

1 **Intel Berkeley, US,** adj@eecs.berkeley.edu
2 **Universität Tübingen, DE,** pavel.laskov@uni-tuebingen.de
3 **Università di Cagliari, IT,** roli@diee.unica.it
4 **University of California – Berkeley, US,** tygar@cs.berkeley.edu
5 **Universität Tübingen, DE,** blaine.nelson@wsii.uni-tuebingen.de

Abstract

The study of learning in adversarial environments is an emerging discipline at the juncture between machine learning and computer security that raises new questions within both fields. The interest in learning-based methods for security and system design applications comes from the high degree of complexity of phenomena underlying the security and reliability of computer systems. As it becomes increasingly difficult to reach the desired properties by design alone, learning methods are being used to obtain a better understanding of various data collected from these complex systems. However, learning approaches can be co-opted or evaded by adversaries, who change to counter them. To-date, there has been limited research into learning techniques that are resilient to attacks with provable robustness guarantees making the task of designing secure learning-based systems a lucrative open research area with many challenges.

The Perspectives Workshop, "Machine Learning Methods for Computer Security" was convened to bring together interested researchers from both the computer security and machine learning communities to discuss techniques, challenges, and future research directions for secure learning and learning-based security applications. This workshop featured twenty-two invited talks from leading researchers within the secure learning community covering topics in adversarial learning, game-theoretic learning, collective classification, privacy-preserving learning, security evaluation metrics, digital forensics, authorship identification, adversarial advertisement detection, learning for offensive security, and data sanitization. The workshop also featured workgroup sessions organized into three topic: machine learning for computer security, secure learning, and future applications of secure learning.

Seminar 09.–14. September, 2012 – www.dagstuhl.de/12371

1998 ACM Subject Classification C.2.0 Computer-Communication Networks (General): Security and Protection (e.g., firewalls); D.4.6 Operating Systems (Security and Protection); I.2.6 Artificial Intelligence (Learning); I.2.7 Artificial Intelligence (Natural Language Processing); I.2.8 Artificial Intelligence (Problem Solving, Control Methods, and Search); K.4.1 Computers and Society (Public Policy Issues): Privacy; K.6.5 Management of Computing and Information Systems (Security and Protection)

Keywords and phrases Adversarial Learning, Computer Security, Robust Statistical Learning, Online Learning with Experts, Game Theory, Learning Theory

Digital Object Identifier 10.4230/DagRep.2.9.109

Machine Learning Methods for Computer Security, *Dagstuhl Reports*, Vol. 2, Issue 9, pp. 109–130
Editors: Anthony D. Joseph, Pavel Laskov, Fabio Roli, J. Doug Tygar, and Blaine Nelson
DAGSTUHL REPORTS Dagstuhl Reports
Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl Publishing, Germany

1 Executive Summary

Anthony D. Joseph
Pavel Laskov
Blaine Nelson
Fabio Roli
Doug Tygar

Arising organically from a variety of independent research projects in both computer security and machine learning, the topic of machine learning methods for computer security is emerging as a major direction of research that offers new challenges to both communities. Learning approaches are particularly advantageous for security applications designed to counter sophisticated and evolving adversaries because they are designed to cope with large data tasks that are too complex for hand-crafted solutions or need to dynamically evolve. However, in adversarial settings, the assets of learning can potentially be subverted by malicious manipulation of the learner's environment. This exposes applications that use learning techniques to a new type of security vulnerability in which an adversary can adapt to counter learning-based methods. Thus, unlike most application domains, computer security applications present a unique data domain that requires careful consideration of its adversarial nature to provide adequate learning-based solutions—a challenge requiring novel learning methods and domain-specific application design and analysis. The Perspectives Workshop, "Machine Learning Methods for Computer Security", brought together prominent researchers from the computer security and machine learning communities interested in furthering the state-of-the-art for this fusion research to discuss open problems, foster new research directions, and promote further collaboration between the two communities.

This workshop focused on tasks in three main topics: the role of learning in computer security applications, the paradigm of secure learning, and the future applications for secure learning. In the first group, participants discussed the current usage of learning approaches by security practitioners. The second group focused of the current approaches and challenges for learning in security-sensitive adversarial domains. Finally, the third group sought to identify future application domains, which would benefit from secure learning technologies.

Within this emerging field several recurrent themes arose throughout the workshop. A major concern that was discussed throughout the workshop was an uneasiness with machine learning and a reluctance to use learning within security applications and, to address this problem, participants identified the need for learning methods to provide better transparency, interpretability, and trust. Further, many workshop attendees raised the issue of how human operators could be incorporated into the learning process to guide it, interpret its results, and prevent unintended consequences, thus reinforcing the need for transparency and interpretability of these methods. On the learning side, researchers discussed how an adversary should be properly incorporated into a learning framework and how the algorithms can be designed in a game-theoretic manner to provide security guarantees. Finally, participants also identified the need for a proper characterization of a security objective for learning and for benchmarks for assessing an algorithm's security.

This document summarizes the presentations and working groups held at the 2012 "Machine Learning Methods for Computer Security" Dagstuhl Perspectives Workshop. Sections 3, 4 and 5 summarize the invited presentations held by the workshop's participants. Section 6 then provides a short summary of the topics discussed by each of the workshop's

three workgroups. Finally, the open problems discussed during the workshop are summarized in Section 7 and are followed by our acknowledgments and a list of the workshops attendees.

2 Contents

Executive Summary

Talks I: Overview of Adversarial Machine Learning

Talks II: Adversarial Learning and Game-theoretic Approaches

Talks III: Secure Learning Applications, Evaluation, and Practices

Working Groups

3 Talks I: Overview of Adversarial Machine Learning

3.1 Adversarial Attacks Against Machine Learning and Effective Defenses

Richard P. Lippmann (MIT – Lexington, US)

Machine learning is widely used to solve difficult security problems by adaptively training on large databases. Examples include computer spam detection, antivirus software, computer intrusion detection, automated Internet search engines such as Google, credit-card fraud detection, talker identification by voice, and video surveillance. Many of these systems face active adversaries with strong financial incentives to defeat accurate performance. Just as humans are susceptible to fraud and misdirection, many of these new learning systems are susceptible to adversarial attacks. This presentation provides a taxonomy of the types of adversarial attacks that can be launched against learning systems and also a summary of effective defenses that can be used to counter these attacks. This analysis is meant to raise the awareness of weaknesses in many widely deployed learning systems, of successful defenses to counter adversarial attacks, and of the arms race this interaction engenders.

3.2 Adversarial Machine Learning, Part I

Doug Tygar (University of California – Berkeley, US)

Joint work of Tygar, J. D.; Barreno, Marco; Huang, Ling; Joseph, Anthony D.; Nelson, Blaine; Rao, Satish; Rubinstein, Benjamin I. P.
Main reference L. Huang, A.D. Joseph, B. Nelson, B.I.P. Rubinstein, J.D. Tygar, "Adversarial machine learning," in Proc. of the 4th ACM Workshop on Security and Artificial Intelligence (AISec'11), pp. 43–58, ACM.
URL http://dx.doi.org/10.1145/2046684.2046692

This talk is the first part of a two-part presentation on research on adversarial machine learning research at UC Berkeley: the study of effective machine learning techniques against an adversarial opponent. In this talk, we give a taxonomy for classifying attacks against online machine learning algorithms; discuss application-specific factors that limit an adversary's capabilities; introduce two models for modeling an adversary's capabilities; explore the limits of an adversary's knowledge about the algorithm, feature space, training, and input data; explore vulnerabilities in machine learning algorithms; discuss countermeasures against attacks; introduce the evasion challenge; and discuss privacy-preserving learning techniques. This talk focuses particularly on the taxonomy for classifying attacks and the technology of Reject on Negative Impact.

3.3 Adversarial Machine Learning, Part II

Anthony D. Joseph (University of California – Berkeley, US)

Joint work of Marco Barreno, Ling Huang, Alex Kantchelian, Blaine Nelson, Justin Ma, Satish Rao, Benjamin I.P. Rubinstein, and J. Doug Tygar

This talk is the second part of a two-part presentation on research on adversarial machine learning research at UC Berkeley: the study of effective machine learning techniques against an adversarial opponent. In this talk, we discuss attacks against a network anomaly detector and an approach to making the detector more robust against such attacks; present approaches to the near-optimal evasion problem for convex-inducing classifiers; and introduce the problem of spam detection for social networks and a content complexity-based approach to detecting spam in such environments.

3.4 Pattern recognition systems under attack: issues learned in Cagliari

Fabio Roli (Università di Cagliari, IT)

In this talk, I discuss the research experience of the PRA Lab (prag.diee.unica.it) of the University of Cagliari on adversarial pattern recognition. I start with our early work, at the end of the 1990s, on multiple classifiers for intrusion detection in computer networks, then I move to our work on evade-hard multiple classifiers and the exciting digression on image-spam filtering, up to our current activity on countermeasures against spoofing attacks in biometric systems. A few issues that I learned in the context of real applications are the focus of my talk. I do not provide the details of the partial solutions we proposed, as the talk's goal is most of all sharing some issues that, I hope, can stimulate discussion. This talk is coordinated with the other three participants coming from my Lab.

3.5 Attack Detection in Networks and Applications: lessons learned in Cagliari

Giorgio Giacinto (Università di Cagliari, IT)

Joint work of Giacinto, Giorgio; Ariu, Davide; Corona, Igino; Roli, Fabio

The investigation of Machine Learning paradigms for detecting attacks against networked computers was a response to the weaknesses of attack signatures. As a matter of fact, signatures usually capture just some characteristics of the attack, thus leaving room for the attacker to produce the same effects by applying slight variations in the way the attack is crafted.

The generalization capability of machine learning algorithms has encouraged many researchers to investigate the possibility of detecting variations of known attacks. While machine learning succeeded in achieving this goal in a number of security scenarios, it was also a source of large volumes of false alarms. We learned that to attain the trade-off between detection rate and false alarm rate was not only a matter of the selection of the learning paradigm, but it was largely dependent on the problem statement. Source data selection, feature definition, and model selection have to be carefully crafted to attain the best trade-off between detection accuracy, generalization capability, and false alarm generation. These issues have been outlined by referring to the detection of attacks against web applications.

References

1 I. Corona, R. Tronci, G. Giacinto, *SuStorID: A Pattern Recognition System to the Protection of Web Services*, In the Proceedings of the 21st International Conference on Pattern Recognition, Japan, Nov 11-15, 2012 (in press).

2 R. Perdisci, I. Corona, G. Giacinto, *Early Detection of Malicious Flux Networks via Large-Scale Passive DNS Traffic Analysis*, IEEE Trans. on Dependable and Secure Computing, 9(5), 2012, pp. 714–726

3 D. Ariu, R. Tronci, G. Giacinto, *HMMPayl: an Intrusion Detection System based on Hidden Markov Models*, Computers & Security, 30, 2011, pp. 221–241

4 I. Corona, G. Giacinto, C. Mazzariello, F. Roli, C. Sansone, *Information fusion for computer security: State of the art and open issues*, Information Fusion, 10, 2009, pp. 274–284

5 G. Giacinto, R. Perdisci, M. Del Rio, F. Roli, *Intrusion detection in computer networks by a modular ensemble of one-class classifiers*, Information Fusion, (1), 2008, pp. 69–82

6 G. Giacinto, F. Roli and L. Didaci, *Fusion of multiple classifiers for intrusion detection in computer networks*, Pattern Recognition Letters, 24(12), 2003, pp. 1795–1803

3.6 Security evaluation of pattern classifiers: lessons learned in Cagliari

Battista Biggio (Università di Cagliari, IT)

Joint work of Biggio, Battista; Fumera, Giorgio; Roli, Fabio

Pattern recognition systems are increasingly being used in adversarial environments like biometric authentication, network intrusion detection and spam filtering tasks, in which data can be adversely manipulated by humans to undermine the outcomes of an automatic analysis. Current pattern recognition theory and design methods do not explicitly consider the intrinsic, adversarial nature of these problems. Consequently, pattern recognition systems exhibit vulnerabilities which can be exploited by an adversary to make them ineffective. This may limit their widespread adoption as potentially useful tools in many applications. Extending pattern recognition theory and design methods to adversarial settings is thus a very relevant research direction, which has not yet been pursued in a systematic way.

In this talk I discuss a general framework that addresses one of the main open issues in the field of adversarial machine learning, namely, the security evaluation of pattern classifiers. The goal of such analysis is to give a more complete view of the classifier performance in adversarial environments, by assessing the performance degradation that may be incurred under different, potential attacks. Depending on the application, this may lead to different design choices; for instance, the selection of a different classification model, or parameter setting. Our framework is based on an explicit model of adversary and data distribution, and encompasses, in a coherent and unifying way, different ideas, models and methods proposed in the adversarial classification literature thus far. It can also be exploited to design more secure classifiers. Some application examples and research directions will also be discussed.

This talk is coordinated with the other three participants coming from the PRA Lab (prag.diee.unica.it) of the University of Cagliari.

3.7 Evade-hard multiple classifiers: lessons learned in Cagliari

Giorgio Fumera (Università di Cagliari, IT)

Joint work of Fumera, Giorgio; Biggio, Battista; Roli, Fabio

Multiple classifier systems (MCSs) have been considered for decades in the pattern recognition and machine learning fields as a technique for improving classification accuracy, with respect to the traditional approach based on the design of a single classifier. Recent theoretical results on adversarial classification, as well as tools used in real adversarial environments, like intrusion detection, spam filtering, and biometric identity recognition, lead our research group to investigate whether MCSs can also be useful to increase the "hardness of evasion". Besides providing some (partial) answers to this question, our analysis pointed out that the "hardness of evasion" of pattern classifiers must be defined and evaluated taking into account the characteristics and constraints of the specific application at hand, as well as using a proper adversary's model. This talk is coordinated with the other three participants coming from my Lab (Fabio Roli, Giorgio Giacinto and Battista Biggio).

4 Talks II: Adversarial Learning and Game-theoretic Approaches

4.1 Machine Learning in the Presence of an Adversary

Tobias Scheffer (Universität Potsdam, DE)

Joint work of Brückner, Michael; Kanzow, Christian; Scheffer, Tobias
Main reference M. Brückner, C. Kanzow, T. Scheffer, "Static Prediction Games for Adversarial Learning Problems," in Journal of Machine Learning Research 13:2617–2654, 2012.
URL http://jmlr.csail.mit.edu/papers/volume13/brueckner12a/brueckner12a.pdf

Machine learning algorithms are commonly based on the assumption that data at training and application time are governed by identical distributions. This assumption is violated when the test data are generated by an adversary in response to the presence of a predictive model. A number of robust learning models have been studied that are based on the worst-case assumption that the adversary will afflicts such changes to the data at application time that achieve the greatest possible adverse effect. This assumption, however, is in many cases overly pessimistic and does not necessarily lead to the ideal outcome if the adversary pursues a goal that is in conflict with, but not necessarily directly antagonistic to, the goal of the learner. We model this interaction as a non-zero-sum, non-cooperative game between learner and data generator. The game-theoretic framework enables us to explicitly model the players' interests, their possible actions, their level of knowledge about each other, and the order at which they commit to their actions. We first assume that both player choose their actions simultaneously, without the knowledge of their opponent's decision. We identify conditions under which this Nash prediction game has a unique Nash equilibrium, and derive algorithms that find the equilibrial prediction model. As a second case, we consider a data generator who is potentially fully informed about the move of the learner. This setting establishes a Stackelberg competition. We derive a relaxed optimization criterion to determine the solution of this game and show that this Stackelberg prediction game generalizes existing prediction models. In case studies on email spam filtering, we empirically explore properties of all derived models as well as several existing baseline methods.

References

1 Michael Brückner, Christian Kanzow, and Tobias Scheffer. *Static Prediction Games for Adversarial Learning Problems*. In the Journal of Machine Learning Research 13:2617–2654, 2012.

2 Michael Brückner and Tobias Scheffer. *Stackelberg games for adversarial prediction problems*. In KDD 2011: Proceedings of the 17th ACM SIGKDD International Conference on Data Mining and Knowledge Discovery, 2011.

4.2 Sparse reward processes and multi-task security games

Christos Dimitrakakis (EPFL – Lausanne, CH)

Joint work of Dimitrakakis, Christos; Auer, Peter
Main reference C. Dimitrakakis, "Sparse Reward Processes," arXiv:1201.2555v2 [cs.LG] , 2012.
URL http://arxiv.org/abs/1201.2555v2

In many security applications, the importance of different resources to be protected is unknown, or arbitrarily changing. In this case, the agent must automatically adapt to the new goals, while continuing to gather information about the environment. The main problem is how much should the agent focus its attention to the task at hand, and how much he should gather information about parts of the environment which may not be directly related to the current task, but which may be relevant for future tasks.

We formalize this setting by introducing a class of learning problems where the agent is presented with a series of tasks. Intuitively, if there is a relation among those tasks, then the information gained during execution of one task has value for the execution of another task. Consequently, the agent might not have to explore its environment beyond the degree necessary to solve the current task.

This paper develops a decision theoretic setting that generalizes standard reinforcement learning tasks and captures these intuitions. More precisely, we introduce sparse reward processes, as a type of multi- stage stochastic game between a learning agent and an opponent. The agent acts in an unknown environment, according to a utility that is arbitrarily selected by the opponent. In the case of security games, the opponent is in fact our client, and the utility is related to the value of our resources and the costs of protection.

Apart from formally describing the setting, we link it to bandit problems, bandits with covariates and factored MDPs. Finally, we examine the behavior of a number of learning algorithms in such a setting, both experimentally and theoretically.

4.3 Near-Optimal Node Blacklisting in Adversarial Networks

Aikaterini Mitrokotsa (EPFL – Lausanne, CH)

Joint work of Dimitrakakis, Christos; Mitrokotsa, Aikaterini
Main reference C. Dimitrakakis, A. Mitrokotsa, "Near-Optimal Node Blacklisting in Adversarial Networks," arXiv:1208.5641 [cs.CR], 2012.
URL http://arxiv.org/abs/1208.5641v1

Many communication networks contain nodes which may misbehave, thus incurring a cost to the network operator. We consider the problem of how to manage the nodes when the operator receives a payoff for every moment that a node stays within the network, but where each malicious node incurs a hidden cost. The operator only has some statistical information about each node's type, and never observes the cost. We consider the case when there are two possible actions: removing a node from a network permanently, or keeping it for at least one more time-step in order to obtain more information. Consequently, the problem can be seen as a special type of intrusion response problem, where the only available response is blacklisting. We first examine a simple algorithm (HiPER) which has provably good performance compared to an oracle that knows the type (honest or malicious) of each node. We then derive three other approximate algorithms by modeling the problem as a Markov decision process. To the best of our knowledge, these algorithms have not been employed before in network management and intrusion response problems. Through experiments on various network conditions, we conclude that HiPER performs almost as well as the best of these approaches, while requiring significantly less computation.

4.4 Formalizing Secure Learning: Quantifying the Security of a Learning Algorithm as a Basis for Assessing and Designing Classifiers

Blaine Nelson (Universität Tübingen, DE)

Joint work of Nelson, Blaine; Biggio, Battista; Laskov, Pavel
Main reference B. Nelson, B. Biggio, P. Laskov., "Understanding the Risk Factors of Learning in Adversarial Environments," in Proc. of the 4th ACM Wworkshop on Security and Artificial Intelligence (AISec '11). pp. 87-92, ACM.
URL http://dx.doi.org/10.1145/2046684.2046698

Machine learning algorithms are rapidly emerging as a vital tool for data analysis and autonomic systems because learners can infer hidden patterns in large complicated datasets, adapt to new behaviors, and provide statistical soundness to decision-making processes. This makes them useful for many emerging tasks in security, networking, and large-scale systems applications. Unfortunately, learning techniques also expose these systems to a new class of security vulnerabilities—learners themselves can be susceptible to attacks. Many common learning algorithms were developed under the assumption that training data is from a natural or well-behaved distribution. However, these assumptions are perilous in a security sensitive setting. With financial incentives encouraging ever more sophisticated adversaries, attacks increasingly target these learners (a prime example is how spammers have adapted their messages to thwart the newest spam detectors). An intelligent adversary can alter his approach based on knowledge of the learner's shortcomings or mislead it by cleverly crafting data to corrupt the learning process.

For this reason, security analysis is a crucial element for designing a practical learning system and for providing it with a sound foundation [1, 2]. However, to properly assess a system's security, the community needs appropriate notions for measuring and benchmarking the security of a learner. Part of this task has already be accomplished: qualitative assessments of security threats have been developed and measures from areas like robust statistics [3, 4] and game-theoretic learning [5] provide a basis for assessing security. However, a comprehensive measure of a learner's security (akin to differential privacy in privacy-preserving learning) has yet to be fully developed or widely accepted as a criteria for assessing classifiers.

Measures of a learner's security are essential to gain a better understanding of the security properties of learning and, ultimately, for their successful deployment in a multitude of new domains and will form the core of a more formal approach to secure learning. In this talk, we motivate the need for security measures of learning as a basis for systematically advancing secure learning. We overview the prior work for assessing a learner's stability and what needs to be done to formalize the notion of secure learning. We also introduce a notion of adversarial corruption that is directly incorporated into a learning framework and derive from it a new criteria for classifier robustness to adversarial contamination.

References

1 Marco Barreno, Blaine Nelson, Anthony D. Joseph, and J. D. Tygar. *The Security of Machine Learning*. Machine Learning, 81(2):121–148, 2010.

2 Pavel Laskov and Marius Kloft. *A Framework for Quantitative Security Analysis of Machine Learning*. In Proceedings of the 2nd ACM Workshop on Security and Artificial Intelligence (AISec), pages 1–4, 2009.

3 Frank R. Hampel, Elvezio M. Ronchetti, Peter J. Rousseeuw, and Werner A. Stahel. *Robust Statistics: The Approach Based on Influence Functions*. John Wiley and Sons, New York, NY, USA, 1986.

4 Peter Huber. *Robust Statistics*. John Wiley & Sons, New York, NY, USA, 1981.

5 Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006.

4.5 Convex Adversarial Collective Classification

Daniel Lowd (University of Oregon, US)

Joint work of Torkamani, Mohamadali; Lowd, Daniel

Many real-world domains, such as web spam, auction fraud, and counter-terrorism, are both relational and adversarial. Previous work in adversarial machine learning has assumed that instances are independent from each other, both when manipulated by an adversary and labeled by a classifier. Relational domains violate this assumption, since object labels depend on the labels of related objects as well as their own attributes.

In this talk, I present a novel method for robustly performing collective classification in the presence of a malicious adversary that can modify up to a fixed number of binary-valued attributes. This method is formulated as a convex quadratic program that guarantees optimal weights against a worst-case adversary in polynomial time. In addition to increased robustness against active adversaries, this kind of adversarial regularization can also lead to improved generalization even when no adversary is present. In experiments on real and simulated data, our method consistently outperforms both non-adversarial and non-relational baselines.

4.6 Data Privacy and Machine Learning

Benjamin I. P. Rubinstein (Microsoft Research – Mountain View, US)

Joint work of Rubinstein, Benjamin I. P.; Bartlett, Peter; Huang, Ling; Narayanan, Arvind; Shi, Elaine; Taft, Nina
Main reference B.I.P. Rubinstein, P.L. Bartlett, L. Huang, N. Taft, “Learning in a Large Function Space: Privacy-Preserving Mechanisms for SVM Learning,” in Special Issue on Statistical and Learning-Theoretic Challenges in Data Privacy, Journal of Privacy and Confidentiality, 4(1), pp. 65–100, 2012.
URL http://repository.cmu.edu/jpc/vol4/iss1/4/

The ubiquitous need for analyzing privacy-sensitive information—including health records, personal communications, product ratings and social network data—is driving significant interest in privacy-preserving data analysis across several research communities. This talk describes two projects related to data privacy and machine learning.

The first (theoretical) part explores the release of Support Vector Machine (SVM) classifiers while preserving the differential privacy of training data. We present efficient mechanisms for finite-dimensional feature mappings and for (potentially infinite-dimensional) mappings with translation- invariant kernels. In the latter case, our mechanism borrows a technique from large-scale learning to learn in a finite-dimensional feature space whose inner-product uniformly approximates the desired feature space inner-product (the desired kernel) with high probability. Differential privacy is established using algorithmic stability, a property used in learning theory to bound generalization error. Utility—when the private classifier is pointwise close to the non-private classifier with high probability—is proven using smoothness of regularized empirical risk minimization with respect to small perturbations to the feature mapping. We conclude the part with lower bounds on the differential privacy of any mechanism approximating the SVM.

The second (experimental) part of the talk describes a winning entry to the IJCNN 2011 Social Network Challenge run by Kaggle.com which is a crowd- sourcing platform for machine learning tasks. The goal of the contest was to promote research on real-world link prediction, and the dataset was a graph obtained by crawling the popular Flickr social photo sharing website, with user identities scrubbed. By applying de-anonymizing attacks on much of the competition test set using our own Flickr crawl, we were able to effectively game the competition. Our attack represents a new application of de-anonymization to gaming machine learning contests, suggesting changes in how future machine learning competitions should be run.

5 Talks III: Secure Learning Applications, Evaluation, and Practices

5.1 Does My Computer (Really) Need My Password to Recognize Me?

Saša Mrdović (University of Sarajevo, SEU)

Basic idea: Computers (and Web sites) we use have enough information on us to enable them to authenticate us without need for (classic) authentication information based on shared secret that needs to be remembered and kept by us and the system. User authentication is, usually, a one time process executed before each interactive session between a user and a machine (OS, Web application, . . .). The user provides little piece(s) of information which confirms that she is entitled to take certain identity that system recognizes. From that moment on, the subject that has been authenticated is bound to her given identity. Authentication is complete with all the rights assigned to given identity and permanent for the duration of the session. A question is if the system should relay on authentication information only to give full user rights and for the whole duration of the session. Continuous authentication has been proposed as a mechanism that continuously re-confirms user's identity during a session. It is usually based on biometric information such as keystroke dynamics or visual data on users obtained through the camera. A limitation of user rights has been implemented through re-authentication before tasks that require elevated privileges can be executed (sudo, UAC). Forensic investigations that I have performed on computers and Web applications showed me that the mentioned systems know things about users that nobody but the user himself should know. Stored data could be used for creation of authentication information that user already knows (does not have to remember) and that is, at the same time, hard for anybody else to know. This is a version of dynamic knowledge based authentication (KBA). Machine learning would be an excellent tool to enable user authentication based on data already stored about user's actions. It could enable continuous authentication as well as authentication for the elevation of privileges. It uses existing data and does not raise (additional) privacy issues. This presentation quickly covers issues with current authentication systems, existing continuous authentication research work and knowledge-based authentication implementations. It gives initial ideas about how a working system could be implemented with its pros and cons.

5.2 Evaluating Classifiers in Adversarial Environments

Alvaro Cárdenas (Fujitsu Labs of America Inc. – Sunnyvale, US)

In machine learning, classifiers are traditionally evaluated based on a testing dataset containing examples of the negative (normal) class and the positive (attack) class. However, in adversarial environments there are many practical situations where we cannot obtain examples of the attack class a priori. There are two main reasons for this: (1) by definition, we cannot obtain examples of zero-day attacks, and (2) using attack examples which are generated independently of the classifier implicitly assumes that the attacker is not adaptive and will not try to evade our detection mechanism.

We argue that instead of using a set of attack samples for evaluating classifiers, we need to find the worst possible attack for each classifier and evaluate the classifier by considering the costs of this worst-case attack.

As a result we can obtain a new trade-off curve as an alternative to ROC curves. The x-axis is still the false positive rate, which can be computed by a dataset of negative examples (which are generally easy to obtain). However, instead of estimating the true positive rate for the y-axis (as in traditional ROC curves) we compute the "cost" of the worst undetected attack by crafting worst-case (in terms of defender cost) attacks. We use the cost of these undetected attacks in the y-axis.

A new area where these types of attacks are easy to create is in the protection of cyber-physical systems and other critical infrastructures, where the state of the system can be associated with a monetary cost. Examples of this new evaluation method are given in the context of electricity theft in the smart metering infrastructure [1] and a chemical reactor [2].

References

1 Daisuke Mashima and Alvaro A. Cardenas. *Evaluating Electricity Theft Detectors in Smart Grid Networks*. In Proceedings of the 15th International Symposium on Research in Attacks, Intrusions, and Defenses (RAID), pp 210-229. Amsterdam, The Netherlands, September 12-14, 2012.

2 Alvaro A. Cardenas and Saurabh Amin and Zong-Syun Lin and Yu-Lun Huang and Chi-Yen Huang and Shankar Sastry. *Attacks against process control systems: risk assessment, detection, and response*. In Proceedings of the 6th ACM Symposium on Information, Computer, and Communications Security (ASIACCS). pp 355-366. Hong Kong, March 22-24, 2011.

5.3 Detection of Malicious PDF Files Based on Hierarchical Document Structure

Nedim Šrndić (Universität Tübingen, DE)

Joint work of Šrndić, Nedim; Laskov, Pavel

Malicious PDF files remain a real threat, in practice, to masses of computer users even after several high-profile security incidents and various security patches issued by Adobe and other vendors. This is a result of widely-used outdated client software and of the expressiveness of the PDF format which enables attackers to evade detection with little effort. Apart from traditional antivirus products, which are always a step behind the attackers, few methods have been proposed which can be deployed for the protection of end-user systems. This talk introduces a highly performant static method for detection of malicious PDF documents which, instead of performing analysis of JavaScript or any other content for detection, relies on essential differences in the structural properties of malicious and benign PDF files. The effectiveness of the method was evaluated on a data corpus containing about 210,000 real-world malicious and benign PDF files and it outperforms each of the 43 antivirus engines at VirusTotal and other specialized detection methods. Additionally, a comparative evaluation of several learning setups with regard to resistance against adversarial evasion will be presented which shows that our method is almost completely immune to sophisticated attack scenarios.

5.4 Detecting Adversarial Advertisements in the Wild

Nathan Ratliff (Google – Pittsburgh, US)

Joint work of Sculley, D.; Otey, Matthew; Pohl, Michael; Spitznagel, Bridget; Hainsworth, John; Zhou, Yunkai
Main reference D. Sculley, M.E. Otey, M. Pohl, B. Spitznagel, J. Hainsworth, Y. Zhou, "Detecting Adversarial Advertisements in the Wild," in Proc. of the Int'l Conf. on Data Mining and Knowledge Discovery (KDD'11).
URL http://www.eecs.tufts.edu/ dsculley/papers/adversarial-ads.pdf

Online advertising is a growing multi-billion dollar industry. Google has an extensive set of ad networks, and wants to provide the best possible advertising experience for both users and advertisers. By providing relevant, high-quality ads to users, customers are more likely to trust the quality of Google's ad networks and make purchases. However, some malicious advertisers attempt to exploit this trust and use questionable means to take users' money. In this talk, we will present the work of Google's Landing Page Quality team, whose purpose it is to remove these adversarial advertisers from Google's ad network. We will give an overview of the systems and methods we use to detect and remove such adversarial advertisers.

References

1 D. Sculley, Matthew Eric Otey, Michael Pohl, Bridget Spitznagel, John Hainsworth, and Yunkai Zhou. *Detecting Adversarial Advertisements in the Wild.* In KDD 2011: Proceedings of the 17th ACM SIGKDD International Conference on Data Mining and Knowledge Discovery. August, 2011.

5.5 Deceiving Authorship Detection

Rachel Greenstadt (Drexel University – Philadelphia, US)

Joint work of Greenstadt, Rachel; Afroz, Sadia; Brennan, Michael; McDonald, Andrew; Caliskan, Aylin; Stolerman, Ariel
Main reference S. Afroz, M. Brennan, R. Greenstadt, "Detecting Hoaxes, Frauds, and Deception in Writing Style Online," IEEE Security and Privacy, 2012
URL https://www.cs.drexel.edu/ sa499/papers/oakland-deception.pdf

In digital forensics, questions often arise about the authors of documents: their identity, demographic background, and whether they can be linked to other documents. The field of stylometry uses linguistic features and machine learning techniques to answer these questions. While stylometry techniques can identify authors with high accuracy in non-adversarial scenarios, their accuracy is reduced to random guessing when faced with authors who intentionally obfuscate their writing style or attempt to imitate that of another author.

In this talk, I will discuss my lab's work in the emerging field of adversarial stylometry. We will discuss our results detecting deception in writing style that may indicate a modified document, demonstrating up to 86% accuracy in detecting the presence of deceptive writing styles. I will also discuss our efforts to aid individuals in obfuscating their writing style in order to maintain anonymity against multiple forms of machine learning based authorship recognition techniques and end with some work in progress extending our research to multilingual data sets and investigating stylometry as a means of authentication.

5.6 Vulnerability Extrapolation: Machine Learning and Offensive Security

Konrad Rieck (Universität Göttingen, DE)

Joint work of Yamaguchi, Fabian; Lottmann, Markus; Rieck, Konrad
Main reference Fabian Yamaguchi, Markus Lottmann, and Konrad Rieck. Generalized Vulnerability Extrapolation using Abstract Syntax Trees. Annual Computer Security Applications Conference (ACSAC), December 2012

Machine learning has been traditionally used as a defensive tool in computer security. Only a little research has studied whether and how learning methods can be applied in an offensive setting. In this talk, we explore this research direction and present a learning-based approach for discovering vulnerabilities in source code. We discuss challenges and difficulties of this setting as well as recent results of our approach for "vulnerability extrapolation".

5.7 Using Machine Learning in Digital Forensics

Felix C. Freiling (Universität Erlangen-Nürnberg, DE)

Joint work of Freiling, Felix C.; Dewald, Andreas; Kälber, Sven; Rieck, Konrad; Ziehe, Andreas

We report on our first experiences of using machine learning for digital event reconstruction in the context of digital forensics. The idea is to learn typical patterns of system/user activities in file system metadata and use this information to infer or classify given file system images with respect to such activities.

5.8 Kernel Mode API Spectroscopy

Viviane Zwanger (Universität Erlangen-Nürnberg, DE)

Joint work of Zwanger, Viviane; Freiling, Felix
Main reference V. Zwanger, F.C. Freiling, "Kernel Mode API Spectroscopy for Incident Response and Digital Forensics," PPREV 2013 (ACM SIGPLAN Program Protection and Reverse Engineering Workshop), Rome.

The new generation of rootkits is heading towards well-designed and carefully crafted attacks to industrial and safety-critical systems with the objective of collecting specific data from a certain target. This creates a need for detecting and analyzing this kind of malicious behavior. We present a simple and surprisingly effective technique which we call "API spectroscopy" to quickly assess the nature of binary code in memory in an automated way. We apply API spectroscopy to the problem of analyzing Windows kernel drivers and kernel mode rootkits. Our method scans the binary code for calls to kernel mode API functions and outputs a histogram of these calls with respect to different semantic classes. We call the result the API spectrum of the driver and the method API spectroscopy. When API calls are grouped into functional classes, an API spectrum can give a compact insight into the possible functionality of an unknown piece of code and therefore is useful in IT incident response and digital forensics. Examples for such semantic classes are “networking”,“filesystem access”, “process management” or “DMA/Port I/O”. We present the design and implementation of an API spectroscope for the Windows operating system. We tested different legitimate drivers from the categories “networking”, “filesystem” and “Hardware” as well as different malicious kernel mode rootkits found in the wild: Duqu, the German Police rootkit software, a facebook-account stealer rootkit and a newly discovered yet unknown rootkit. The resulting API spectra were found to reflect the behavior of the analyzed driver quite well and might be used as a fingerprint system to detect new rootkits.

5.9 Training Data Sanitization in Adversarial Learning: Open Issues

Patrick Pak Kei Chan (South China University of Technology, CN)

The goal of pattern classification is to generalize the knowledge from the training data to the unseen samples. Unfortunately, in many real applications, the training set is influenced by an attacker. The influenced dataset misleads the learner and downgrades its generalization ability. One method to solve this problem is data sanitization, which identifies and eliminates the attack data from the training set before learning. In this talk, the noisy and attack data sanitization concepts are compared. The preliminary idea of sanitizing the training data in adversarial learning is presented.

6 Working Groups

6.1 Machine Learning for Computer Security

Battista Biggio (Università di Cagliari – Cagliari, Italy)
Nedim Šrndić (Universität Tübingen – Tübingen, Germany)

This workgroup explored the current role of machine learning in security research and the success and failures in using learning methods within security applications. This group identified open issues and research priorities including the need for transparency, interpretability and trust for secure learning approaches, the need for preventive measures and the potential for using learning in penetration testing, and the need for scalable procedures. They also identified future applications for secure learning including detecting advanced persistent threats, dynamic authentication, autonomous monitoring, and crime prediction.

6.2 Secure Learning: Theory and Methods

Daniel Lowd (University of Oregon – Eugene, OR, USA)
Rachel Greenstadt (Drexel University – Philadelphia, PA, USA)

The secure learning workgroup confronted the challenges that face learning researchers, who design learning algorithms for adversarial environments. They discussed how security can be formulated as an objective for designing learning procedures; the need for security metrics; the need for developing security-driven benchmarks and adversarial data simulations for evaluating learning approaches; and general techniques, constraints and challenges for developing secure learning technologies. Finally this group identified a set of key open questions including what should be achieved by secure learning, how can we know the adversary, and what is the appropriate role of secure learning within a secure system?

6.3 Future Applications of Secure Learning

Nathan Ratliff (Google – Pittsburgh, PA, USA)
Alvaro Cárdenas (Fujitsu Labs of America Inc. – Sunnyvale, CA, USA)
Fabio Roli (Università di Cagliari – Cagliari, Italy)

The final workgroup addressed the task of identifying future applications, which could benefit from these secure learning technologies. This group examined domains including intrusion detection, malware analysis, spam filtering, online advertisement, social media spam, plagiarism detection / authorship identification, captcha cracking, face detection, copyright enforcement, and sentiment analysis. The group also compiled a list of major research priorities including the need to address poisoning attacks and the need for benchmarks and penetration testing. They also highlighted privacy, non-stationarity of data, and the lack of ground truth as the major hindrances to the production of adequate benchmark datasets for secure learning.

7 Overview of Open Problems

In the "Machine Learning Methods for Computer Security" Perspectives Workshop, attendees identified the following general issues that need to be addressed by the community:

The need for learning approaches to provide a greater degree of trust for seamless integration of these approaches in security applications

There is generally an apprehension within the security community toward black-box learning procedures that can be addressed by providing greater transparency, interpretability and trust. There is, further, a need for preventive and corrective learning procedures and for using machine learning in an offensive role such as using learning to perform penetration testing to better validate existing techniques.

Design problems for incorporating secure learning technologies into security applications

There is also a general question of identifying the appropriate role of secure learning in secure systems and developing a methodology for designing such systems. Among the challenges for this task are finding scalable learning procedures that can meet the security objectives and incorporating human operators to help provide security safeguards.

The need for a more methodical approach to secure learning

Currently, the notion of secure learning largely remains ill-defined and domain-specific and lacks proper evaluation. There is a need to incorporate a realistic model of the adversaries, to formalize secure learning, and to identify which applications can benefit from these technologies. Stronger connections between secure and private learning would be desirable and perhaps result in unifying these currently disparate fields.

Construction of benchmarks and case studies for secure learning

Several groups identified a need for and current lack of adequate benchmarks and case-studies for evaluating secure learning. Unfortunately, real-world security data remains scarce for reasons including the concern of many security-sensitive systems for safeguarding their users' privacy, the ever-evolving nature of the data in these typically non-stationary domains, and the lack of ground truth due to the fact that many attacks may never be revealed or discovered.

Acknowledgments

We thank Dr. Roswitha Bardohl, Susanne Bach-Bernhard, Dr. Marc Herbstritt, and Jutka Gasiorowski for their help in organizing this workshop and preparing this report.

Participants

- Battista Biggio
Università di Cagliari, IT
- Christian Bockermann
TU Dortmund, DE
- Michael Brückner
SoundCloud Ltd., DE
- Alvaro Cárdenas Mora
Fujitsu Labs of America Inc. – Sunnyvale, US
- Christos Dimitrakakis
EPFL – Lausanne, CH
- Felix C. Freiling
Univ. Erlangen-Nürnberg, DE
- Giorgio Fumera
Università di Cagliari, IT
- Giorgio Giacinto
Università di Cagliari, IT
- Rachel Greenstadt
Drexel Univ. – Philadelphia, US
- Anthony D. Joseph
University of California – Berkeley, US
- Robert Krawczyk
BSI – Bonn, DE
- Pavel Laskov
Universität Tübingen, DE
- Richard P. Lippmann
MIT – Lexington, US)
- Daniel Lowd
University of Oregon, US)
- Aikaterini Mitrokotsa
EPFL – Lausanne, CH
- Sasa Mrdovic
University of Sarajevo, SEU
- Blaine Nelson
Universität Tübingen, DE
- Patrick Pak Kei Chan
South China University of Technology, CN
- Massimiliano Raciti
Linköping University, SE
- Nathan Ratliff
Google – Pittsburgh, US
- Konrad Rieck
Universität Göttingen, DE
- Fabio Roli
Università di Cagliari, IT
- Benjamin I. P. Rubinstein
Microsoft – Mountain View, US
- Tobias Scheffer
Universität Potsdam, DE
- Galina Schwartz
University of California – Berkeley, US
- Nedim Srndic
Universität Tübingen, DE
- Radu State
University of Luxembourg, LU
- Doug Tygar
University of California - Berkeley, US
- Viviane Zwanger
Univ. Erlangen-Nürnberg, DE