AFRL-AFOSR-VA-TR-2019-0217

SCIENCE OF CYBER SECURITY MODELING, COMPOSITION, AND MEASUREMENT

John Mitchell
LELAND STANFORD JUNIOR UNIVERSITY

07/25/2019
Final Report

Air Force Research Laboratory
AF Office Of Scientific Research (AFOSR)/ RTA2
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
11/05/2018	Final Technical Report	11/30/2011 - 06/30/2017

4. TITLE AND SUBTITLE
"(MURI-10) Science of Cyber Security: Modeling, Composition, and Measurement"

5a. CONTRACT NUMBER
FA9550-12-1-0040

5b. GRANT NUMBER
NA

5c. PROGRAM ELEMENT NUMBER
NA

6. AUTHOR(S)
Mitchell, John C.

5d. PROJECT NUMBER
NA

5e. TASK NUMBER
NA

5f. WORK UNIT NUMBER
NA

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
(1) Stanford University, 3160 Porter Dr., Ste. 100, Palo Alto, CA, 94304. (2) Carnegie Mellon University, Pittsburgh, PA 15213. (3) Cornell University, PO Box 22, Ithaca, NY 14851. (4) UC Berkeley, 2195 Hearst Ave., Berkeley, CA 94720. (5) University of Pennsylvania, 3451 Walnut Street, Philadelphia, PA 19104.

8. PERFORMING ORGANIZATION REPORT NUMBER
NA

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
USAF, AFRL
AF OFFICE OF SCIENTIFIC RESEARCH
875 NORTH RANDOLPH STREET, RM 3112
ARLINGTON VA 22203

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/AFOSR

11. SPONSOR/MONITOR'S REPORT NUMBER(S)
NA

12. DISTRIBUTION/AVAILABILITY STATEMENT
Distribution A, approved for public release.

13. SUPPLEMENTARY NOTES
N/A

14. ABSTRACT
This project has advanced the science base for trustworthiness by developing concepts, relationships, and laws with predictive value. Focusing on pressing problem areas amenable to rigorous treatment and generalizable solutions, the project was organized around three thrust areas: Security Modeling. Secure Composition, and Security Measurement.

15. SUBJECT TERMS
Cybersecurity, Science, Foundations, Security Modeling, Secure Composition, Security Measurement.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Tristan Nguyen (AFOSR)
					19b. TELEPHONE NUMBER *(Include area code)*
None	None	None	UU	1	(703) 696-7796

Table of Contents

1. Summary 1

2. Introduction 1

3. Methods, Assumptions, and Procedures 2

4. Results and Discussion 2

4.1 Modeling 2

a. Science of Usable Security 2

b. Rational Resource-Bounded Models for Human Agents 2

c. Exact Bounds for Cryptographic Protocols 3

d. Space-bounded Adversary Models 3

4.2 Composition 4

a. Information Flow Analysis of Big Data Systems 4

b. Small Model Theorems and Model Validation for Scalable System Security 4

c. Language and Hardware-Supported Timing Channel Mitigation 5

d. Sound Static Analysis for Vulnerability Discovery in Distributed Systems 5

4.3 Measurement 5

a. Worth-based Information Theory 5

b. Security Metrics 5

c. Software Security Measurement 5

d. Scientific Evaluation of CAPTCHA Strength 6

e. Usable Security Measurement 7

5. Conclusions 7

1. Summary

Our nation's increasing dependence on computing systems that are not trustworthy puts individuals, commercial enterprises, the public sector, and our military at risk. This project has advanced the science base for trustworthiness by developing concepts, relationships, and laws with predictive value. Focusing on pressing problem areas amenable to rigorous treatment and generalizable solutions, the project was organized around three thrust areas: Security Modeling. Secure Composition, and Security Measurement. A uniform approach to security modeling allows systematic approaches to be developed and applied to a broad range of richly connected systems, supporting analysis of resilience against graduated classes of clearly defined threat models. Principles of secure composition support systematic and modular construction of trustworthy systems, relative to security properties that can be verified and validated through theoretical proof and/or experiment. New security measurement concepts support comparing the relative strengths of defense mechanisms, determining whether security improves from one version of a system to another, and when additional security mechanisms are warranted given incentives associated with system attackers and defenders. Together, advances in these three complementary thrusts support a science base for future systems that proactively resist attacks through secure design, development, and implementation based on principled foundations.

2. Introduction

This project advanced the science base for trustworthiness by developing concepts, relationships, and laws with predictive value. The multidisciplinary team focused on problems that are amenable to rigorous treatment and that have generalizable solutions. Recognizing the target areas of greatest need and scientific opportunity, the project was organized around three complementary and interrelated thrusts

Security modeling. We developed aspects of a general security-modeling framework that focuses on characterizing system behavior, possible actions of an attacker, and properties that must be preserved in the face of attacks. Our uniform approach to security modeling supports analysis of a broad range of components and richly connected systems, supporting analysis of resilience against graduated classes of clearly defined threat models. Logics enable us to give precise definitions and leverage them for verifying the effectiveness of cyber security policies and mechanisms. Further, game-theoretic concepts are used to model incentives for the defender and disincentive mechanisms for the attacker, supporting analysisthat incorporates rational defense strategies and decisions.

Secure Composition. Conventional wisdom—that security properties do not compose—is not supported by science. We draw on our previous work that gave evidence to the contrary in specific topic areas, ranging from network security to information flow. From this starting point, we developed principles for explaining when security schemes do, or do not, compose, and how to achieve compositionality. For instance, even if the security property we care about arenot fully compositional, there may be stronger enforcement mechanisms that do compose and that suffice to ensure the security property is achieved. We developed theoretical frameworks for secure composition to identify what types of reasoning compose, and then apply this theory and corresponding principles to a wide range of important contemporary systems and security

properties. This thrust focused on composition, as appropriate to modular system design and construction, as well as refinements that preserve information flow, integrity, and that incorporate assumptions about what components can be trusted and what must be trusted.

Security Measurement. Metrics for comparing the security of alternative system designs and defensive mechanisms have been widely recognized as key goals for a science base. Developments in this area necessarily require precise security models, leveraging the modeling effort of the project described above. Our work in quantitative measures for information leakage and for integrity degradation provided an initial basis for further success. We also developed measures for determining relative strengths of certain kinds of defense mechanisms, design improvements, and determining how additional mechanisms may serve goals associated with various kinds of system attackers and defenders.

3. Methods, Assumptions, and Procedures

This university research combined direct study of fundamental topics supporting a general science of cyber security with continued investigation of particular areas of recognized importance. Just as classical physics, for example, has separable theories of kinematics, optics, and thermodynamics, we currently have separate subareas of cyber security concerned with networks, operating systems, software isolation, and programming language security. We worked to develop a broadly applicable science base for trustworthiness that involves both an explicit emphasis on the science of cyber security and continued efforts to both bring general principles to bear on specific problems and extract general understanding from case studies and specific efforts.

4. Results and Discussion

4.1 Modeling

a. Science of Usable Security

While usability has been the subject of much interesting research in the computer security community in the last decade, a significant challenge that has not been addressed is how to formalize usability. PI Datta and colleagues at CMU (Blocki, Blum) break new ground towards this important scientific goal by focusing on the concrete problem of formalizing and constructing usable password schemes. Our naturally rehearsal password schemes cleverly share elements of passwords across multiple sites to enable natural rehearsals (provably improving password usability) while ensuring security even if some sites are compromised. Our GOTCHA scheme requires users to perform an additional task of matching inkblots with descriptions they provide themselves while quantifiably increasing the cost of offline attacks on hashed passwords. These results have also received significant press, including articles in Scientist American and the MIT Technology Review.

b. Rational Resource-Bounded Models for Human Agents

PI Halpern (Cornell) and colleagues have produced a set of fundamental models of human decision-making. These results can serve as a foundation for further work on science of security

in settings where decisions made by humans (users, system administrators etc) have security consequences.

I. Showed that rather than viewing people as systematically irrational, it may be possible to understand their actions as the outcome of being perfectly rational, but resource bounded. Specifically, we showed that the optimal automaton for solving various problems of interest makes the same systematic deviations from rationality that people do, no matter how many states it has. This is discussed in the paper "I'm doing as well as I can: modeling people as rational finite automata".

II. Showed that by taking seriously the idea that a player's preferences are expressed in some language we can capture many interesting phenomena in game theory. Of particular interest is the role of coarse languages that cannot express all the distinctions that an agent may be able to perceive. These issues are discussed in the paper "Language-based games".

III. We show that an epsilon-Nash equilibrium in repeated games can be found in polynomial time. In contrast, earlier work by Borgs et al. suggests that this problem is intractable. We get our result by making small and arguably natural and well-justified changes to the Borgs et al. model---we model the players as polynomial-time Turing machines that maintain state (rather than stateless polynomial-time Turing machines) and assuming that deviators must themselves be polymomial time--and making some standard cryptographic hardness assumptions (the existence of public-key encryption). This result is discussed in the paper "The truth behind the myth of the folk theorem".

c. Exact Bounds for Cryptographic Protocols

Prior foundational work on systematic and formalized reasoning about network security is based on either a highly idealized syntactic “Dolev-Yao” model, or form of asymptotic complexity theory that does not help determine the required value of important parameters such as the length of cryptographic keys. In order to advance the science of network protocol security, we have taken substantial steps in developing a formal logic for quantitative reasoning about security properties of network protocols. The system allows us to derive exact security bounds that can be used to choose key lengths or other concrete security parameters. The system includes axioms for digital signatures and random nonces, with concrete security properties based on concrete security of signature schemes and pseudorandom number generators (PRG). The formal logic supports first-order reasoning and reasoning about protocol invariants, taking exact security bounds into account. Proofs constructed in our logic also provide conventional asymptotic security guarantees because of the way that exact bounds accumulate in proofs. As an illustrative example producing exact bounds, we use the formal logic to prove an authentication property with exact bounds of a signature-based challenge-response protocol.

d. Space-bounded Adversary Models

PI Scedrov (Penn) and coauthors considered bounded memory protocols and bounded memory Dolev-Yao adversaries. Like the standard Dolev-Yao adversary in the security protocol analysis literature, whose memory is unbounded, the bounded adversaries may also act as the network, intercepting, sending, and composing messages, but their memories are bounded. Scedrov and

coauthors showed that such bounded memory adversary, when given enough memory, can carry out many known security protocol anomalies. This led them to the question whether it is possible to compute an upper-bound on the memory required by the standard Dolev-Yao adversary to carry out an anomaly from the memory restrictions on the bounded protocol. Scedrov and coauthors proved that the answer was negative, that is, even for bounded memory protocols, the standard Dolev-Yao adversary cannot be computably approximated by a sequence of bounded memory Dolev-Yao adversaries.

4.2 Composition

a. Information Flow Analysis of Big Data Systems

With the rapid increase in cloud services collecting and using user data to offer personalized experiences, ensuring that these services comply with their privacy policies has become a business imperative for building user trust. However, most compliance efforts in industry today rely on manual review processes and audits designed to safeguard user data, and therefore are resource intensive and lack coverage. We designed and implemented a system to automate privacy policy compliance checking in Bing. Central to the design of the system are (a) LEGALEASE —a language that allows specification of privacy policies that impose restrictions on how user data is handled; and (b) GROK —a data inventory for Map-Reduce-like big data systems that tracks how user data flows among programs. GROK maps code-level schema elements to datatypes in LEGALEASE , in essence, annotating existing programs with information flow types with minimal human input. Compliance checking is thus reduced to information flow analysis of big data systems. The system, bootstrapped by a small team, checks compliance daily of millions of lines of ever-changing source code written by several thousand developers. This result is reported in one of the highest ranked papers at the 2014 IEEE Symposium on Security and Privacy. PI Datta (CMU) worked jointly with a team at Microsoft Research on this effort.

b. Small Model Theorems and Model Validation for Scalable System Security

PI Wagner (Berkeley) and colleagues at Berkeley and Intel have developed a set of scientific techniques for improving the scalability of secure system analysis. The first result below builds on prior work by PI Datta's team at CMU.

I. We developed new techniques for reasoning about the security of systems composed out of smaller components. State-space explosion poses a fundamental challenge in this area that limits our ability to verify the security of complex systems built by composing many smaller components. We focus specifically on table-oriented systems, where the state-space of the full space contains arrays or tables, and where the state-space for each entry in the array or table is limited. Our experience is that table-driven systems are common in hardware and low-level systems software, but standard approaches to verification (such as model checking) tend to fail for these systems; we show how to extend them to handle this case. We then analyze a number of case studies to show that our approach is effective in practice. This work is described in "Verification with Small and Short Worlds".

II. We extended our work on verification by examining the model validation problem. Often, security verification is done by manually constructing a mathematical model of the system and

then verifying that the model meets our security goals. However, this leaves open the question of whether the mathematical model accurately reflects the behavior of the actual system. We developed automated techniques to check whether the model matches the system. This work is described in “Symbolic Software Model Validation”.

c. Language and Hardware-Supported Timing Channel Mitigation

PI Myers and colleagues at Cornell have developed new methods for provably mitigating timing channels at the language and hardware levels. Much prior work on timing channel mitigation has focused on cryptographic techniques like blinding and run-time enforcement techniques like injecting worst-case delays.

d. Sound Static Analysis for Vulnerability Discovery in Distributed Systems

PI Myers and colleagues at Cornell have developed new methods for finding security flaws in distributed systems. In particular, they:

I. Introduced new methods for automatically and soundly identifying and diagnosing static information flow errors in program code.

II. Identified and formally specified new classes of security vulnerabilities in distributed systems equipped with remote references, and introduced a new static analysis for preventing these vulnerabilities.

4.3 Measurement

a. Worth-based Information Theory

PI Scedrov (Penn) with his postdoc Alvim (who was supported by this MURI) and PI Schneider (Cornell) have proposed a framework for quantitative information flow in which leaks that involve a given number of bits may not all be equally harmful. In this approach secrets are defined in terms of fields, which are combined to form structures. A worth assignment is introduced to associate each structure with a worth, perhaps in proportion to the harm that would result from its disclosure. Alvim, Scedrov and Schneider show how they can capture inter-dependence among structures within a secret and how they can model secret-sharing, information-theoretic predictors, and computational guarantees for security. Using non-trivial worth assignments, Alvim, Scedrov, and Schneider generalize Shannon entropy, guessing entropy, and probability of guessing. For deterministic systems, they consider lattice of information in order to provide an underlying algebraic structure for the composition of attacks.

b. Security Metrics

PI Schneider (Cornell) developed a theory of security metrics and proved that sound and complete metrics are undecidable.

c. Software Security Measurement

PI Wagner (Berkeley) and colleagues conducted a set of experiments aimed at understanding and improving software security.

I. We studied what factors predict how effective developers are at detecting security vulnerabilities through source code review. We hired 30 developers with some knowledge of security and asked them to review a small software system for security vulnerabilities, and

measured their effectiveness at finding the known vulnerabilities in the system. We also gathered data from them on their experience, education, and other information. Surprisingly, we found that none of the following factors had any statistically significant correlation with effectiveness at finding security vulnerabilities: educational level, amount of education or training on security, experience at code reviews, experience with computer security, experience with the particular programming language, experience in software development, self-reported confidence in their review, self-reported comprehension of the application code, self-reported confidence as a developer, self-reported confidence as a security expert. This is surprising, because it suggests that many indicators that might be used for selecting or hiring software developers are not effective at predicting task effectiveness for the specific task of security code review. We also found that there was widespread variance in the effectiveness of these developers, but that most developers found only a small fraction of the vulnerabilities. This work has implications for selection and hiring on software projects where security is important. It also advances work into the emerging area of empirical, evidence-based study of what factors and practices lead to secure systems, and what doesn't. This work is reported in “An Empirical Study on the Effectiveness of Security Code Review.”

II. We studied the effectiveness of vulnerability rewards programs, sometimes called “security bug bounty” programs. We conducted an empirical analysis of vulnerability rewards programs used by Mozilla (Firefox) and Google (Chrome). Our work found evidence that these programs are highly effective: they find many vulnerabilities, at a modest cost. It also suggested some factors that influence the effectiveness of these programs. Ultimately, our conclusion was that vulnerability rewards programs are a valuable part of the software development lifecycle. This work has proved to be timely; since our work was published, over a dozen companies have introduced vulnerability reward programs. The work was published in “An Empirical Study of Vulnerability Rewards Programs.”

d. Scientific Evaluation of CAPTCHA Strength

CAPTCHAs, which are automated tests intended to distinguish humans from programs, are used on many web sites to prevent bot-based account creation and spam. To avoid imposing undue user friction, CAPTCHAs must be easy for humans and difficult for machines. However, the scientific basis for successful CAPTCHA design was minimal prior to this MURI project. We contributed to the basic science by devising a systematic parameterized characterization of CAPTCHAs and evaluating their difficulty for humans and automated algorithms, as a function of these parameters. This effort also involved fundamental work in machine learning, as part of the evaluation.

In one study, we carried out a systematic study of existing visual CAPTCHAs based on distorted characters that are augmented with anti-segmentation techniques. Applying a systematic evaluation methodology to 15 current CAPTCHA schemes from popular web sites, we find that 13 are vulnerable to automated attacks. Based on this evaluation, we identify a series of recommendations for CAPTCHA designers and attackers, and possible future directions for producing more reliable human/computer distinguishers.

In another study, we present a large scale evaluation of CAPTCHAs from the human perspective, with the goal of assessing how much friction CAPTCHAs present to the average user. For the purpose of this study we have asked workers from Amazon's Mechanical Turk and an underground CAPTCHA-breaking service to solve more than 318,000 captchas issued from the 21 most popular CAPTCHA schemes (13 images schemes and 8 audio scheme). Analysis of the resulting data reveals that CAPTCHAs are often difficult for humans, with audio captchas being particularly problematic. We also find some demographic trends indicating, for example, that non-native speakers of English are slower in general and less accurate on English-centric CAPTCHA schemes. Evidence from a week's worth of eBay CAPTCHAs (14,000,000 samples) suggests that the solving accuracies found in our study are close to real-world values, and that improving audio CAPTCHAs should become a priority, as nearly 1% of all CAPTCHAs are delivered as audio rather than images. Finally our study also reveals that it is more effective for an attacker to use Mechanical Turk to solve CAPTCHAs than an underground service.

e. Usable Security Measurement

PI Wagner (Berkeley) and his team studied users' understanding of, if an application on the user's smartphone misbehaves, how to identify which application was responsible for the malicious behavior. Our work found that many users would have difficulty identifying the application that is responsible in many situations, and that applications running in the background pose a particular challenge for attribution; users tend to blame whatever application happens to be running in the foreground at the moment the misbehavior becomes apparent, which may be inaccurate if another application is running in the background. Our online and lab studies gave new insights into how users attribute the source of misbehavior. Based upon this, we were able to identify new mechanisms that could be used by smartphone operating systems to help users do better at attribution, thus improving our ability to deter malicious or overly-aggressive applications. This work is described in “When It’s Better to Ask Forgiveness than Get Permission: Attribution Mechanisms for Smartphone Resources.”

5. Conclusions

This project has advanced the science base for trustworthiness by developing concepts, relationships, and laws with predictive value. Focusing on pressing problem areas amenable to rigorous treatment and generalizable solutions, the project was organized around three thrust areas: Security Modeling. Secure Composition, and Security Measurement. A uniform approach to security modeling allows systematic approaches to be developed and applied to a broad range of richly connected systems, supporting analysis of resilience against graduated classes of clearly defined threat models. Principles of secure composition support systematic and modular construction of trustworthy systems, relative to security properties that can be verified and validated through theoretical proof and/or experiment. New security measurement concepts support comparing the relative strengths of defense mechanisms, determining whether security improves from one version of a system to another, and when additional security mechanisms are warranted given incentives associated with system attackers and defenders. Together, advances in these three complementary thrusts support a science base for future systems that proactively

resist attacks through secure design, development, and implementation based on principled foundations.

This project ran 11/30/2011 - 06/30/2017 and produced results as summarized above. The team of multidisciplinary university researchers is grateful to AFOSR for the opportunity to do this work and for the support of the program managers as well as others in DoD and other supportive offices of the US Government.