

Threat Modeling Research and Machine Learning

Dr. Nancy R. Mead

Topics

- SEI Threat Modeling Research
- SEI's Hybrid Threat Modeling Method (hTMM)
- CMU MITS Student Project on Machine Learning
- Machine Learning and hTMM Research

Threat Modeling Research

SEI Threat Modeling Research 2015 – 2016

Threat Modeling Definition(s)

- A **threat modeling method (TMM)** is an approach for creating an abstraction of a software system, aimed at identifying attackers' abilities and goals, and using that abstraction to generate and catalog possible threats that the system must mitigate. (SEI)
- Threat modeling is a methodology and a tool used to identify and classify vulnerabilities which, if exploited, would result in adverse business impact. (Microsoft/Ford Motor Company)
- Threat modeling is repeatable process to find and address all threats to your product. (Microsoft)

SEI Initial Threat Modeling Research

- Focus on early lifecycle activities (e.g., requirements engineering, design), independent of a lifecycle model.
- Evaluate competing TMMs to
 - identify and test principles regarding which ones yield the most efficacy
 - provide evidence about the conditions under which different ones are most effective
 - allow reasoning about the confidence in threat modeling results

TMMs Studied

STRIDE

- Represents the state of the practice
- Developed at Microsoft; “lightweight STRIDE” variant adopted from Ford Motor Company
- Successive decomposition of w/r/t system components, threats



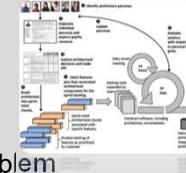
Security Cards

- Design principle: inject more creativity and brainstorming into process; move away from checklist-based approaches
- Developed at the University of Washington
- Physical resources (cards) facilitate brainstorming across several dimensions of threats
- Includes reasoning about attacker motivations, abilities



Persona non Grata (PnG)

- Design principle: make the problem more tractable by giving modelers a specific focus (here: attackers, motivations, abilities)
- Developed at DePaul University based on proven principles in HCI
- Once attackers are modeled, process moves on to targets and likely attack mechanisms



Universal weakness: empirical evaluation in the context of the software development lifecycle

Threat Modeling Research

STRIDE Approach

STRIDE Threat Model

- Invented in 1999 by Kohnfelder & Garg; implemented at Microsoft and widely adopted
- Typical implementation:
 - Model system w/ Data Flow Diagrams (DFD)
 - Map the DFD to Threat Categories
 - Determine the threats (via threat trees)
 - Document the threats and steps for prevention
- Can be implemented manually or through free SDL Threat Modeling Tool
- Considered relatively easy to implement but...time-consuming and prone to different results based on implementer

- STRIDE Threat Categories:



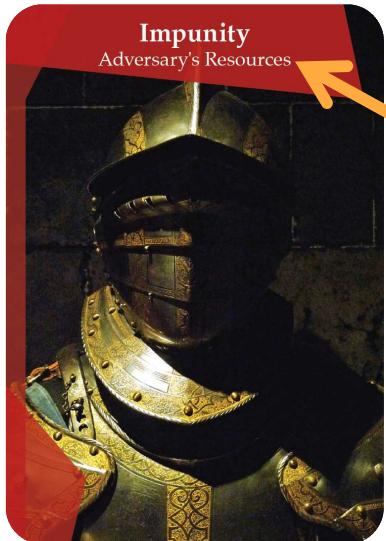
| Threat | Security Property |
|------------------------|-------------------|
| Spoofing | Authentication |
| Tampering | Integrity |
| Repudiation | Non-repudiation |
| Information Disclosure | Confidentiality |
| Denial of Service | Availability |
| Elevation of Privilege | Authorization |

Scandariato et. al., 2015; Hernan et. al., 2006

Threat Modeling Research

Security Cards Approach

Example Card (Front)



Card topic

Card dimension

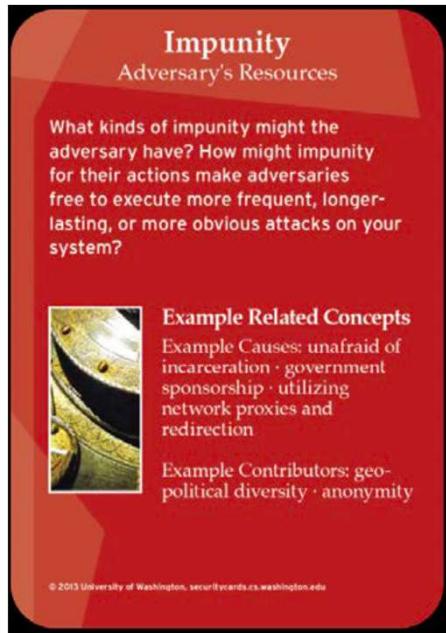


Adversary's Motivations

Adversary's Resources

Adversary's Methods

Example Card (Back)



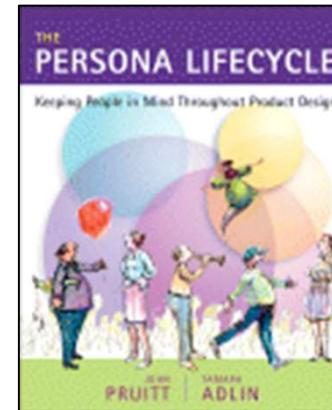
Threat Modeling Research

PnG Approach

What Is a Persona?

- “**Personas** are detailed descriptions of imaginary people constructed out of well-understood, highly specified data about real people.”

● — John Pruitt & Tamara Adlin



J. Pruitt, T. Adlin. *The Persona Lifecycle: Keeping People in Mind Throughout Product Design*. Morgan Kaufman, 2006. (<https://dl.acm.org/citation.cfm?id=1076976>)

Example Persona



- Thomas is a 76-year-old retired accountant who enjoys spending time with his grandchildren. During his retirement, he enjoys reading newspapers, working in his garden, and staying in touch with friends. He is a free spirit and enjoys exploration and technology, but only when it doesn't get in his way.

Developing a PnG

1. **Motivations:** What is the PnG's motivations? Monetary gain? Revenge? Recognition? Laughs?
2. **Goals:** What goals does the PnG have to fulfill its motivation (i.e., what does it want to do and how does it plan to get away with it)?
3. **Skills:** What skills does it have to achieve their goal? What other assets does it have (e.g., access to infrastructure, relationships to those who have skills)?
4. **Misuse Cases:** What are the misuse cases the PnG can follow to achieve their goals?

Example PnG: Mike –1

- **Description:** Mike worked as a contractor installing SCADA radio-controlled sewage equipment for a municipal authority. After leaving the contractor, Mike applied for a job with the municipality but was rebuffed. Feeling bitter and rejected, Mike decides to get even with the municipality and his former employer.
- **Goals:** Cause raw sewage to leak into local parks and rivers and make the events appear as malfunctions. Create a public backlash against the contractor and municipality.



"Mike" is based on the true story of Vitek Boden, who was convicted of causing the release of sewage in Maroochy Shire Council in Queensland, Australia in 2000 after hacking the associated SCADA system. See Abrams & Weiss, *Malicious Control System Cyber Security Attack Case Study—Maroochy Water Services, Australia, 2008*. ([http://www.scadahackr.com/library/Documents/Case_Studies/Case%20Study%20-%20NIST%20-%20Maroochy%20\(presentation\).pdf](http://www.scadahackr.com/library/Documents/Case_Studies/Case%20Study%20-%20NIST%20-%20Maroochy%20(presentation).pdf))

Example PnG: Mike –2

- **Skills:** Extensive knowledge of SCADA equipment, including control computers, relevant programs, and radio communication protocols; access to specialized equipment
- **Misuse Cases:**
 - Steal control computer and radio equipment from his former employer.
 - Using the stolen computer, construct a fake pumping control station from which to send radio signals.
 - Gain remote access to the SCADA system and disable alarms at pumping stations.
 - Issue radio commands (using stolen radio equipment) to instruct pumping stations to release sewage.

SEI Study Methodology

- 250+ subjects
 - novice learners (SW and cyber), returning practitioners, professionals
- All applied TMMs to common testbeds: systems with understandable Concept of Operations, and DoD relevance



UAV (CPS)



Aircraft maintenance app
(IT)

- Within-subjects design: Each team learns and applies one approach on a testbed and then learns the next and applies it on the other testbed.

The threat template, scenarios, and examples were designed to be reusable.

One of Several Results: How Frequently Is a Given Threat Type Reported?

- If we know that a TMM was able to find a given threat, how confident can we be that it would be reported by a team?
 - STRIDE: Great variability
 - Security Cards: Able to find the most threat types, but also substantial variability across teams
 - PnG: Was the most focused TMM, but showed the most consistent behavior across teams
- No single TMM led to teams reporting a majority of the valid threats.

Results: Do the TMMs Help Modelers Find Important Classes of Threats?

- **Primary Measure**

How many of the threat types identified by professionals were found by the student teams?

- **Other Aspects of Effectiveness**

- Some types of threats were never uncovered by teams using some TMMs.
- Some TMMs led to many threat types from outside our expert set. (May be false positives or just unusual.)

Overall Impressions from the Study

- **STRIDE** is intended to be used at a slightly later time in the lifecycle, when the system can be represented using data flow diagrams. It has more of a “cookbook” style than the other methods.
- The **Security Cards** approach encourages thinking outside the box and creativity, with variability in results.
- The **PnG** approach focuses more narrowly but provides consistent results.
- The SEI team believed that a merger of the Security Cards and PnG approaches would produce a more consistent and complete view of threats.

Reading Assignment

Threat Modeling: A Summary of Available Methods;

Nataliya Shevchenko, Timothy A. Chick, Paige O'Riordan, Tom Scanlon, Carol Woody, PhD

<https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=524448>

Threat Modeling Research

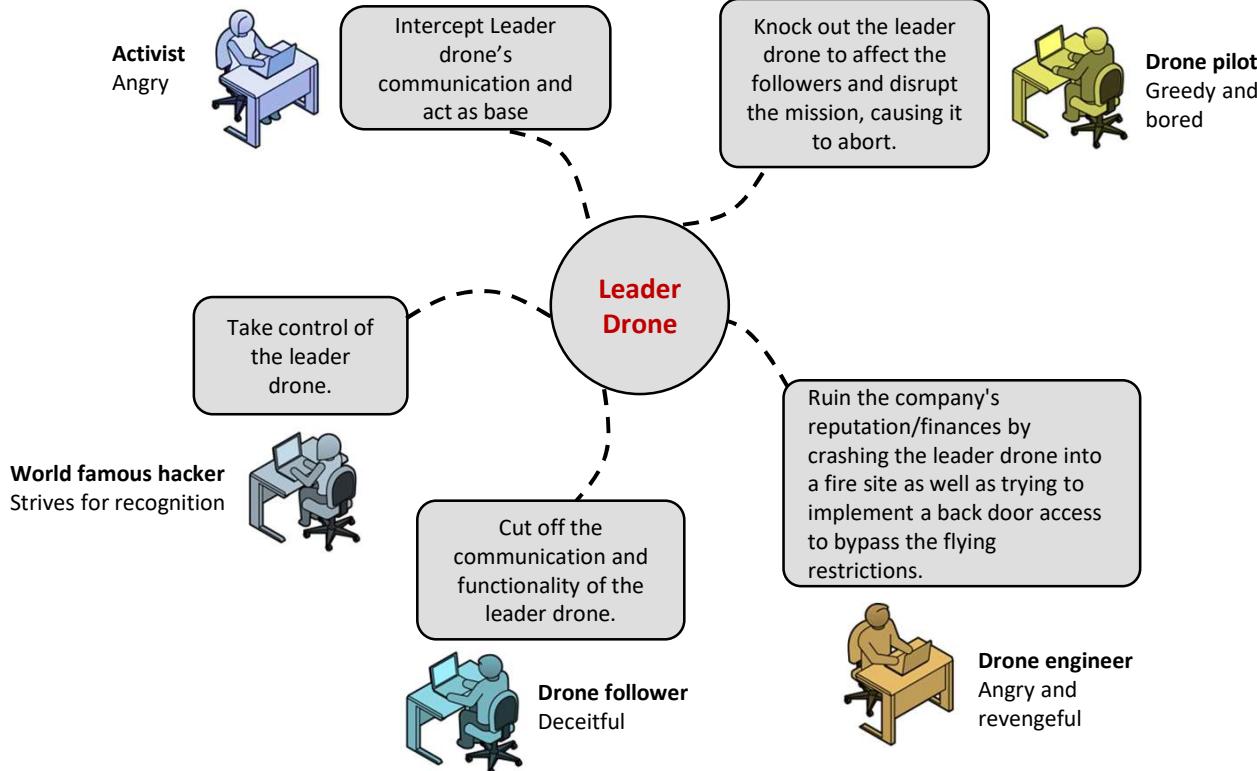
Additional PnG Analysis

PnG Study Methodology



- 108 students in two introductory information security courses (undergrad and graduate)
 - novice learners (SW and cyber), returning practitioners, professionals
 - These are the “crowd”
- All applied PnG to an Unmanned Autonomous Vehicle (UAV) system scenario, in teams of 3-4 people.

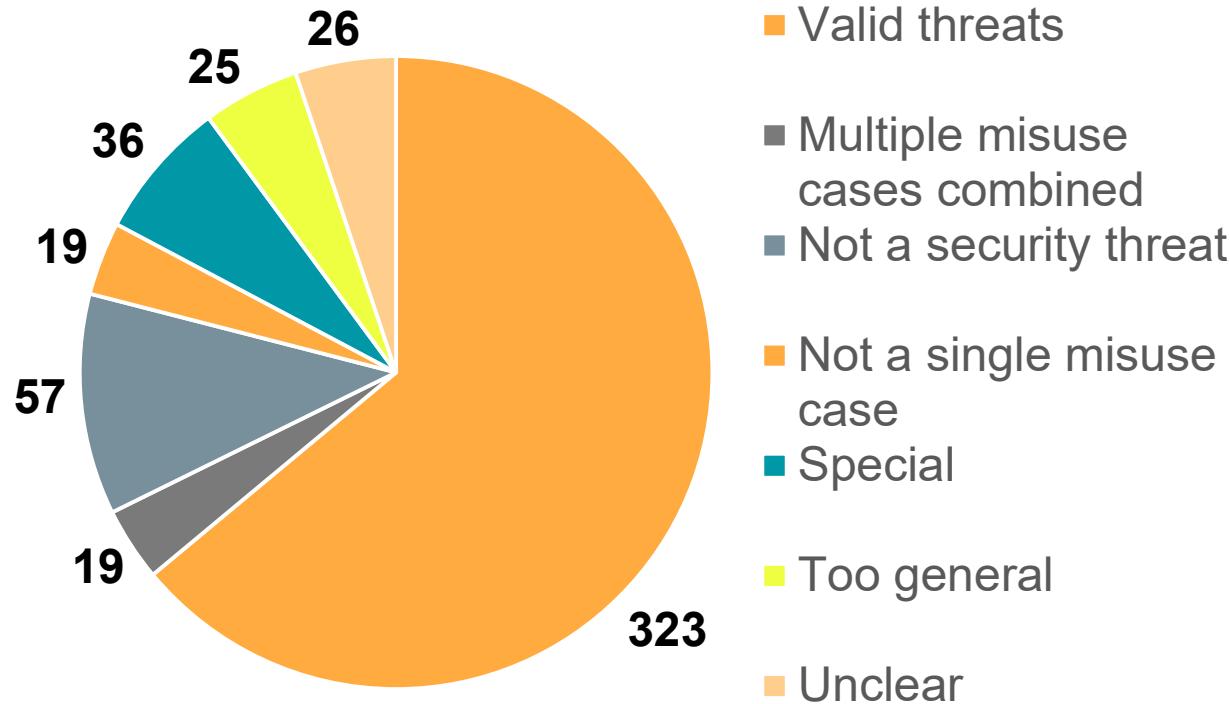
Spider Web View of Threats to Leader Drone



PnG Merging Process

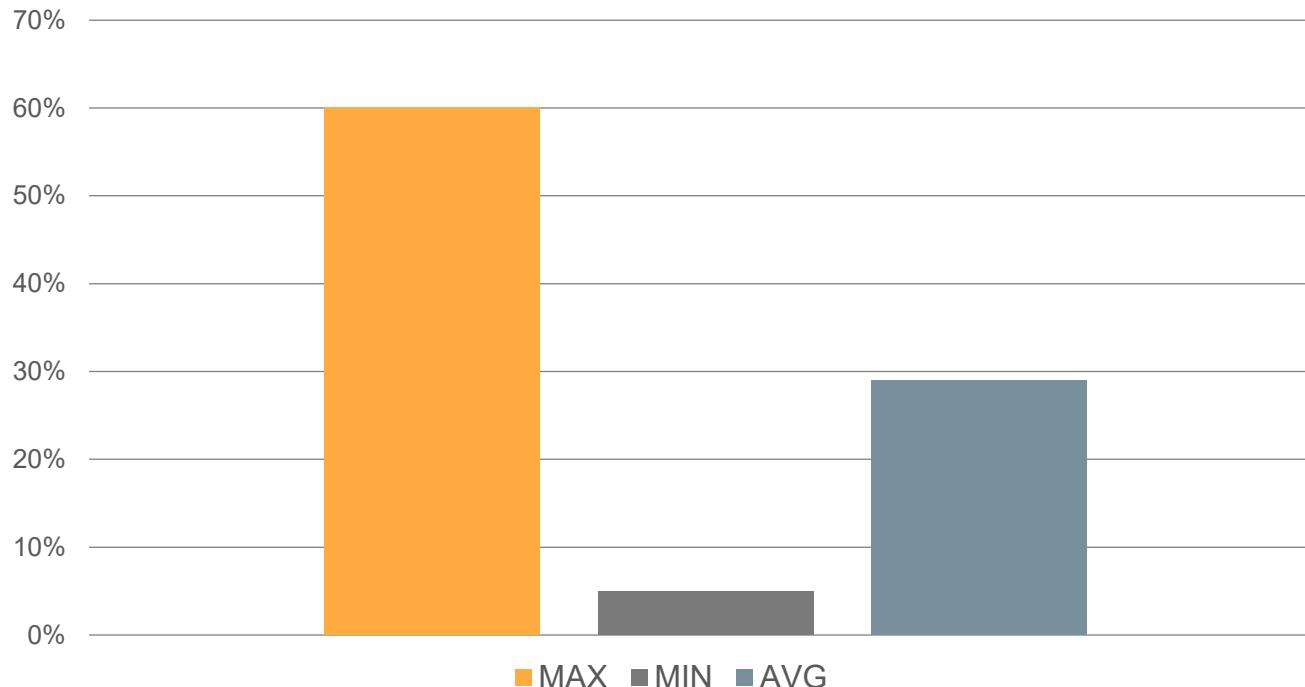
- Step 1: Discover domain-specific concepts
- Step 2: Identify attack targets
- Step 3: Visually display attack mechanisms
- Step 4: Merge individual threats into new PnGs
- Step 5: Check for redundancy

Student PnG Analysis Insights – Overview



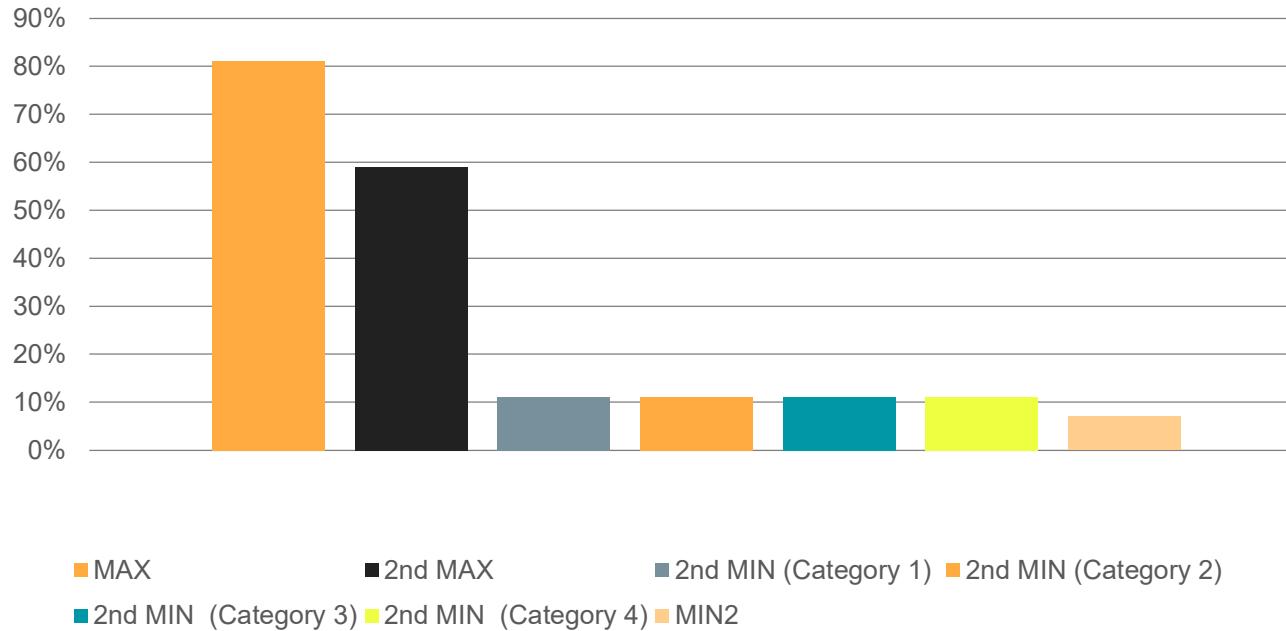
Student PnG Analysis Insights – Valid threats

Teams



Student PnG Analysis Insights – Valid threats

Threat Frequency



Threats to Validity

- Only one case study was explored.
- Crowd was information systems students, not necessarily IT professionals.
- Presented only one example, which was not evaluated quantitatively.

PnG and Machine Learning Conclusion

- Machine Learning could be used to analyze individual PnGs created by a crowd.
- Our approach resulted in PnGs that could serve as input to the requirements process.
- Approach was illustrated in one project domain, but not fully evaluated with users.
- Experimentation in diverse domains and projects, with tools, would be useful.

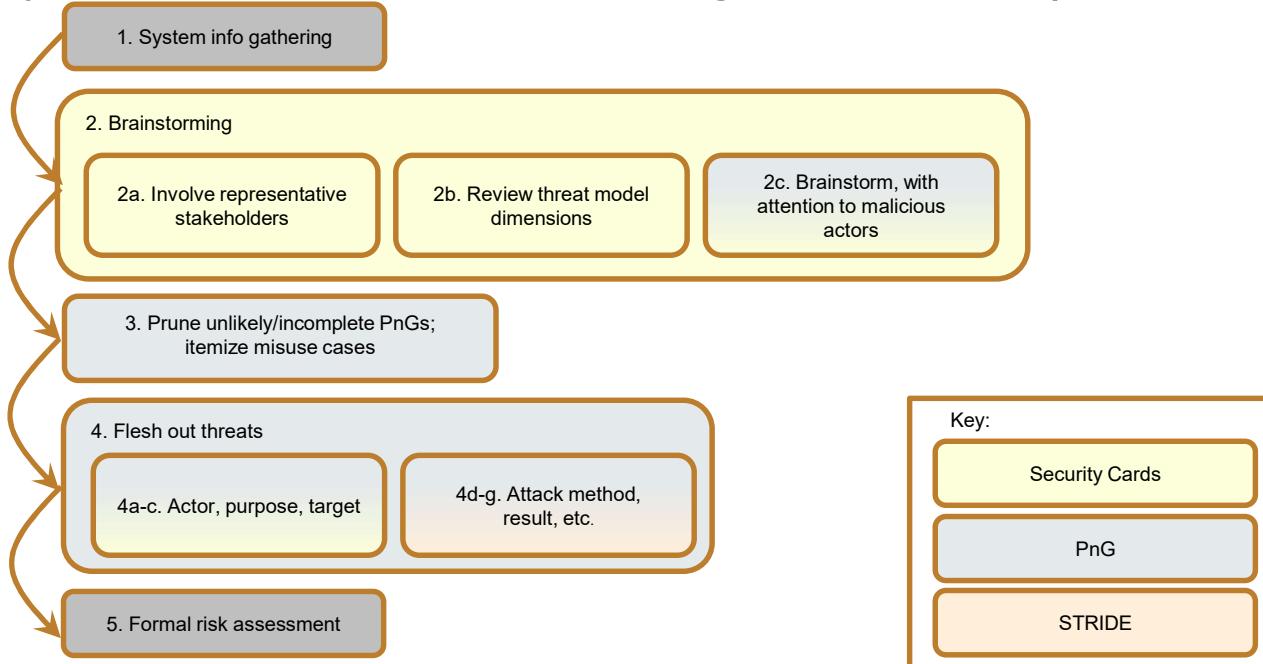
Threat Modeling Research

Hybrid Threat Modeling Method 2017 – 2018

Desirable Threat Modeling Characteristics

- Desirable Characteristics of a Threat Modeling Method
 - minimal false positives
 - minimal overlooked threats
 - consistent results regardless of who is doing the threat modeling
 - cost-effective (doesn't waste time)
 - empirical evidence to support its efficacy
- Other Considerations
 - has tool support
 - suggests a prioritization scheme
 - easy to learn, intuitive

Initial Hybrid Threat Modeling Method (hTMM) –1



Subsequent Activity

- Applied on real-world medium-size project.
- Method was tailored to interests of project managers.
- Feedback on the tailored method was provided, but more general conclusions could not be derived from this case study.

Reading Assignment

- SEI Blog Entry: A Hybrid Threat Modeling Method
https://insights.sei.cmu.edu/sei_blog/2018/04/the-hybrid-threat-modeling-method.html
- On Canvas: Crowd Sourcing the Creation of Personae Non Gratae for Requirements-Phase Threat Modeling; Nancy Mead, Forrest Shull, Janine Spears, Stefan Hiebl, Sam Weber, and Jane Cleland-Huang; IEEE International Requirements Engineering Conference Proceedings, September 2017, pp. 404-409 DOI 10.1109/RE.2017.803633

Additional Reference

- SEI Technical Note: A Hybrid Threat Modeling Method:
<https://resources.sei.cmu.edu/library/asset-view.cfm?assetid=516617>

Threat Modeling Research

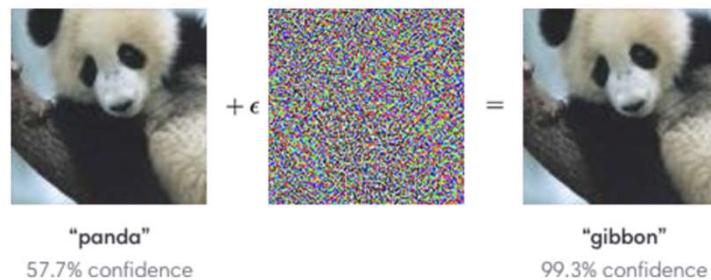
CMU MITS Student Project on Machine Learning 2018 - Advex

Project Goal

- Assess robustness of machine learning models against adversarial examples.

What are Adversarial Examples?

- **Adversarial examples** are inputs to machine learning models that an attacker has intentionally designed to cause the model to make a mistake.



Source: <https://blog.openai.com/adversarial-example-research>

Motivation for the Project

- There are no security frameworks in place to prevent these adversarial examples from attacking a machine learning model.
- Thus, people need to make sure their models are robust enough to be safe from such adversarial examples.
- However, there are only a few resources for people to evaluate their model's robustness against attacks.
- Current resources are not comprehensive enough for users to assess their models thoroughly.

Goal

- To build a website that assesses robustness of deep learning models against adversarial examples on computer vision tasks.
- The website will evaluate and provide feedback about the robustness of models uploaded by users using the Cleverhans library and ImageNet dataset.

Input: Clean Images

Output: Adversarial Examples (Images) generated by each attack method

Highlights

- By using attack methods built with different assumptions, we measure the model's vulnerability against adversaries with different levels of capabilities.
- By controlling the amount of noise introduced in the adversarial examples, we measure the model's vulnerability under different noise tolerance levels.
- By using attack methods that are generalizable across datasets, the assessment results will be reliable.

Features

- Scalable System
- User Friendly Web Interface
- Comprehensive and Reliable Feedback

Functional Requirements & Constraints

Functional Requirements

- Dashboard
- Model Upload Form
- Submission History
- Submission Detail
- Information Page

Business Constraints

- Budget limit for student teams
- Free to users

Technical Constraints

- Based on Cleverhans library
- Deployed on AWS
- Supports Keras only
- Targets Computer Vision models only

Summary of Prior Results

- Initial research showed there is no single “best method” for threat modeling.
- The hTMM was successfully applied to a small example and was then applied to a medium-size system.
- The CMU MITS Project was completed and informed other machine learning research projects

Threat Modeling Research

Machine Learning and hTMM Research Ideas

Future Research Ideas - 1

- Apply multiple threat modeling methods, including hTMM to systems that will employ machine learning, to try to understand which ones are effective
 - Opportunity for collaboration with student teams, practitioners
 - Focus on one of our prior examples, with a machine learning twist
 - Develop new examples to be used threat modeling/machine learning experimentation

Future Research Ideas - 2

- Examine machine learning in a different way
- Can machine learning be used to improve threat modeling?
- Examine existing research on machine learning to improve resistance to attacks.
- Focus on a specific domain or one of our prior examples
 - Power (SCADA) systems
 - Drone example
- Examine whether machine learning can be used in conjunction with hTMM to improve threat modeling

References for the MITS Project

<https://blog.openai.com/adversarial-example-research/>

<https://github.com/tensorflow/cleverhans>

<https://www.youtube.com/watch?v=KJ1zZsia5yQ>

References for ML

- Detecting Stealthy False Data Injection Using Machine Learning in Smart Grid, Mohammad Esmalifalak, Lanchao Liu, Nam Nguyen, Rong Zheng, and Zhu Han
- Deep Learning-Aided Cyber-Attack Detection in Power Transmission Systems, David Wilson, Yufei Tang, Jun Yan, and Zhusuo Lu
- SEI Blog Post: Measuring Resilience in Artificial Intelligence and Machine Learning Systems

<https://insights.sei.cmu.edu/insider-threat/2019/12/measuring-resilience-in-artificial-intelligence-and-machine-learning-systems.html>

- The Top 10 Risks of Machine Learning Security, McGraw et al, IEEE Computer, June 2020
- Threats for Machine Learning, Mark Sherman, SEI Webinar

Homework Assignment

With your project team, use the PnG method to identify PnGs for your project, identifying their motivations, goals, and skills. Does this suggest threats that you might have missed with STRIDE, or does it validate the STRIDE results?

Note: You are not required to develop full misuse cases for this exercise. However, if you were using PnG for threat modelling, you would need the misuse cases in order to develop the corresponding mitigation strategies.