An Artificial Coevolutionary Framework for Adversarial AI

Jamal Toutouh
toutouh@mit.edu
ALFA
http://groups.csail.mit.edu/ALFA
unamay@csail.mit.edu
Agenda

- Adversarial Engagements and Arms Races
- Network Security Arms Races
  - RIVALS framework
    - RIVALS: Robustness vs Denial
    - AVAIL: Isolation vs Contagion
    - DARK Horse and ADHD: Deception vs reconnaissance
  - Acknowledgments:
    - Funding
      - Member companies supporting Cybersecurity@CSAIL
      - DARPA XD3
      - MIT Lincoln Labs
    - Work
      - Members of the ALFA group and collaborators
LEARNING

GOALS

STRATEGY

TACTICS

TECHNIQUES

PROCEDURES

TOOLS

IMPLEMENTATION

LEARNING
RIVALS helps the defense \textit{anticipate} the attack strategies given a defensive configuration (and mission)

RIVALS helps the defense consider arms races and Design effective courses of action for the network to be resilient
Advanced Persistent Threat Kill Chain

- **Intelligence Gathering**
  - Conduct background research.

- **Initial Exploitation**
  - Execute initial attack.
  - Establish foothold.

- **Command and Control**
  - Enable persistence.
  - Conduct enterprise reconnaissance.

- **Privilege Escalation**
  - Move laterally to new systems.
  - Escalate privileges.

- **Data Exfiltration**
  - Gather and encrypt data of interest.
  - Exfiltrate data from victim systems.
  - Maintain persistent presence.

Deceptive Defense With Honeypots

Fig. 5

DeceptionSDN

SDN Controller

Flow Rule Generator
Analyze Flow Statistics

Network view generator

Configures Deceptive Virtual Network

Deception Server

Manipulate Virtual Network Traffic
Simulate Virtual Network Resources

Achleitner et al.
Engagement

CONFLICTING OBJECTIVES

MEASUREMENTS

d: time for defense to detect a scan (sec.)
t: time to run the scan (sec.)
n: number of scan detections by defender
h/H: ratio of real nodes that were discovered to total real nodes.

Evaluate using Mininet
Adversarial Behaviors

<table>
<thead>
<tr>
<th>Actor</th>
<th>Configuration</th>
<th>Range</th>
</tr>
</thead>
<tbody>
<tr>
<td>Attacker</td>
<td>NMAP IP scan batch size</td>
<td>[10, 50, 100, 200]</td>
</tr>
<tr>
<td></td>
<td>Total number of IPs to scan</td>
<td>[200, 300, 400]</td>
</tr>
<tr>
<td></td>
<td>IP address visit order</td>
<td>random, local, sequential, local-seq.</td>
</tr>
<tr>
<td>Defender</td>
<td>Number of real nodes ($H$)</td>
<td>[20, ... , 40]</td>
</tr>
<tr>
<td></td>
<td>Min honeypots per subnet</td>
<td>[1, ... , 10]</td>
</tr>
<tr>
<td></td>
<td>Max honeypots per subnet</td>
<td>[11, ... , 20]</td>
</tr>
<tr>
<td></td>
<td>Number of subnets</td>
<td>[3, ... , 6]</td>
</tr>
<tr>
<td></td>
<td>Real node distribution in subnets</td>
<td>random, even, crowding</td>
</tr>
</tbody>
</table>

**SDN DECEPTION SYSTEM PARMS**

- SDN Controller
  - Flow Rule Generator
  - Analyze Flow Statistics
- Network View Generator
  - Configures Deceptive Virtual Network
- Deception Server
  - Manipulate Virtual Network Traffic
  - Simulate Virtual Network Resources
Defensive Learning

APT SCANNING TACTICS

SDN DECEPTION SYSTEM PARMS

SDN Controller
- Flow Rule Generator
- Analyze Flow Statistics

Network view generator
- Configures Deceptive Virtual Network

Deception Server
- Manipulate Virtual Network Traffic
- Simulate Virtual Network Resources

MACHINE LEARNING

ALFA
ANYSCALE LEARNING FOR ALL
Static Attack – Optimized Defense

**Hypothesis:** Good defense has more honeypots, subnets and real hosts with even distribution

**Results:**
- More difficult to detect smaller NMAP batch sizes
  - Fitness function rewards discovering more real host less than the penalty of being detected: smaller scans do better
- Defense against an attacker that scans with local preference is the most difficult
- Expected real behavior of attackers is to start scan their local subnet

**Possible recommendation:** create subnets for DHCP leases where real hosts are in a different subnet

<table>
<thead>
<tr>
<th>Visit Order</th>
<th>Batch Size</th>
<th>Num. IPs</th>
<th>N Real Nodes</th>
<th>Real Node Dist</th>
<th>Subnets</th>
<th>Min-Max HP</th>
<th>Nodes Disc.</th>
<th>Detected Scans</th>
<th>1st Detection(s)</th>
<th>HPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>random</td>
<td>5</td>
<td>400</td>
<td>19</td>
<td>crowding</td>
<td>6</td>
<td>7-18</td>
<td>9</td>
<td>9</td>
<td>34.62</td>
<td>96</td>
</tr>
<tr>
<td>local</td>
<td>5</td>
<td>400</td>
<td>18</td>
<td>random</td>
<td>5</td>
<td>8-13</td>
<td>4</td>
<td>10</td>
<td>32.60</td>
<td>77</td>
</tr>
<tr>
<td>sequential</td>
<td>5</td>
<td>400</td>
<td>11</td>
<td>crowding</td>
<td>6</td>
<td>3-18</td>
<td>3</td>
<td>8</td>
<td>42.55</td>
<td>66</td>
</tr>
<tr>
<td>local-seq.</td>
<td>5</td>
<td>400</td>
<td>15</td>
<td>even</td>
<td>6</td>
<td>4-17</td>
<td>3</td>
<td>4</td>
<td>59.04</td>
<td>78</td>
</tr>
<tr>
<td>random</td>
<td>5</td>
<td>200</td>
<td>13</td>
<td>random</td>
<td>6</td>
<td>4-16</td>
<td>6</td>
<td>7</td>
<td>28.61</td>
<td>60</td>
</tr>
<tr>
<td>local</td>
<td>5</td>
<td>200</td>
<td>19</td>
<td>crowding</td>
<td>5</td>
<td>5-18</td>
<td>11</td>
<td>8</td>
<td>30.53</td>
<td>85</td>
</tr>
<tr>
<td>sequential</td>
<td>5</td>
<td>200</td>
<td>17</td>
<td>random</td>
<td>5</td>
<td>8-20</td>
<td>2</td>
<td>7</td>
<td>46.64</td>
<td>78</td>
</tr>
<tr>
<td>local-seq.</td>
<td>5</td>
<td>200</td>
<td>17</td>
<td>crowding</td>
<td>6</td>
<td>1-20</td>
<td>9</td>
<td>6</td>
<td>18.40</td>
<td>42</td>
</tr>
<tr>
<td>random</td>
<td>10</td>
<td>200</td>
<td>15</td>
<td>random</td>
<td>5</td>
<td>7-19</td>
<td>7</td>
<td>5</td>
<td>34.63</td>
<td>61</td>
</tr>
<tr>
<td>local</td>
<td>10</td>
<td>200</td>
<td>13</td>
<td>even</td>
<td>4</td>
<td>6-14</td>
<td>4</td>
<td>8</td>
<td>34.57</td>
<td>37</td>
</tr>
<tr>
<td>sequential</td>
<td>10</td>
<td>200</td>
<td>17</td>
<td>even</td>
<td>6</td>
<td>8-20</td>
<td>4</td>
<td>12</td>
<td>14.37</td>
<td>87</td>
</tr>
<tr>
<td>local-seq.</td>
<td>10</td>
<td>200</td>
<td>15</td>
<td>crowding</td>
<td>5</td>
<td>7-13</td>
<td>11</td>
<td>8</td>
<td>26.45</td>
<td>60</td>
</tr>
</tbody>
</table>
## Results:
- Difficult to attack many honeypots and subnets.
- Easier with crowded distribution of real hosts, large reward when that subnet is scanned (similar for defender when avoiding) Points to adopting high risk- and high reward tactic

<table>
<thead>
<tr>
<th>N Real Nodes</th>
<th>Real Node Dist</th>
<th>Subnets</th>
<th>Min-Max HP</th>
<th>Visit Order</th>
<th>Batch Size</th>
<th>Num. IPs</th>
<th>Nodes Disc.</th>
<th>Detected Scans</th>
<th>1st Detection(s)</th>
<th>HPs</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>even</td>
<td>4</td>
<td>1-10</td>
<td>local</td>
<td>25</td>
<td>300</td>
<td>8</td>
<td>11</td>
<td>22.62</td>
<td>26</td>
</tr>
<tr>
<td>20</td>
<td>crowding</td>
<td>4</td>
<td>1-10</td>
<td>sequential</td>
<td>5</td>
<td>400</td>
<td>12</td>
<td>2</td>
<td>42.74</td>
<td>14</td>
</tr>
<tr>
<td>20</td>
<td>random</td>
<td>4</td>
<td>1-10</td>
<td>local</td>
<td>10</td>
<td>300</td>
<td>4</td>
<td>3</td>
<td>26.61</td>
<td>19</td>
</tr>
<tr>
<td>20</td>
<td>even</td>
<td>4</td>
<td>10-20</td>
<td>local</td>
<td>25</td>
<td>200</td>
<td>5</td>
<td>19</td>
<td>22.54</td>
<td>86</td>
</tr>
<tr>
<td>20</td>
<td>crowding</td>
<td>4</td>
<td>10-20</td>
<td>sequential</td>
<td>50</td>
<td>300</td>
<td>20</td>
<td>19</td>
<td>16.47</td>
<td>73</td>
</tr>
<tr>
<td>20</td>
<td>random</td>
<td>4</td>
<td>10-20</td>
<td>sequential</td>
<td>10</td>
<td>400</td>
<td>8</td>
<td>12</td>
<td>22.43</td>
<td>76</td>
</tr>
<tr>
<td>20</td>
<td>even</td>
<td>10</td>
<td>1-10</td>
<td>sequential</td>
<td>5</td>
<td>400</td>
<td>3</td>
<td>2</td>
<td>24.52</td>
<td>54</td>
</tr>
<tr>
<td>20</td>
<td>crowding</td>
<td>10</td>
<td>1-10</td>
<td>sequential</td>
<td>50</td>
<td>400</td>
<td>20</td>
<td>11</td>
<td>20.50</td>
<td>68</td>
</tr>
<tr>
<td>20</td>
<td>random</td>
<td>10</td>
<td>1-10</td>
<td>local</td>
<td>5</td>
<td>200</td>
<td>2</td>
<td>1</td>
<td>85.35</td>
<td>54</td>
</tr>
<tr>
<td>20</td>
<td>even</td>
<td>10</td>
<td>10-20</td>
<td>local</td>
<td>5</td>
<td>200</td>
<td>1</td>
<td>9</td>
<td>40.72</td>
<td>185</td>
</tr>
<tr>
<td>20</td>
<td>crowding</td>
<td>10</td>
<td>10-20</td>
<td>sequential</td>
<td>50</td>
<td>300</td>
<td>20</td>
<td>22</td>
<td>23.06</td>
<td>214</td>
</tr>
<tr>
<td>20</td>
<td>random</td>
<td>10</td>
<td>10-20</td>
<td>local</td>
<td>5</td>
<td>200</td>
<td>5</td>
<td>9</td>
<td>40.64</td>
<td>170</td>
</tr>
<tr>
<td>20</td>
<td>even</td>
<td>10</td>
<td>40-80</td>
<td>sequential</td>
<td>50</td>
<td>200</td>
<td>2</td>
<td>4</td>
<td>22.13</td>
<td>983</td>
</tr>
<tr>
<td>20</td>
<td>crowding</td>
<td>10</td>
<td>40-80</td>
<td>sequential</td>
<td>50</td>
<td>300</td>
<td>11</td>
<td>18</td>
<td>18.64</td>
<td>683</td>
</tr>
<tr>
<td>20</td>
<td>random</td>
<td>10</td>
<td>40-80</td>
<td>sequential</td>
<td>5</td>
<td>200</td>
<td>2</td>
<td>12</td>
<td>30.84</td>
<td>998</td>
</tr>
<tr>
<td>50</td>
<td>even</td>
<td>10</td>
<td>1-10</td>
<td>local-seq.</td>
<td>25</td>
<td>200</td>
<td>9</td>
<td>13</td>
<td>30.58</td>
<td>36</td>
</tr>
<tr>
<td>50</td>
<td>crowding</td>
<td>10</td>
<td>1-10</td>
<td>sequential</td>
<td>5</td>
<td>300</td>
<td>15</td>
<td>2</td>
<td>68.95</td>
<td>52</td>
</tr>
<tr>
<td>50</td>
<td>random</td>
<td>10</td>
<td>1-10</td>
<td>sequential</td>
<td>5</td>
<td>400</td>
<td>3</td>
<td>2</td>
<td>97.45</td>
<td>48</td>
</tr>
<tr>
<td>50</td>
<td>even</td>
<td>10</td>
<td>10-20</td>
<td>sequential</td>
<td>25</td>
<td>200</td>
<td>7</td>
<td>16</td>
<td>20.42</td>
<td>205</td>
</tr>
<tr>
<td>50</td>
<td>crowding</td>
<td>10</td>
<td>10-20</td>
<td>local-seq.</td>
<td>5</td>
<td>400</td>
<td>20</td>
<td>7</td>
<td>30.76</td>
<td>231</td>
</tr>
<tr>
<td>50</td>
<td>random</td>
<td>10</td>
<td>10-20</td>
<td>sequential</td>
<td>50</td>
<td>200</td>
<td>2</td>
<td>14</td>
<td>22.63</td>
<td>183</td>
</tr>
<tr>
<td>20</td>
<td>even</td>
<td>20</td>
<td>10-20</td>
<td>sequential</td>
<td>25</td>
<td>200</td>
<td>2</td>
<td>17</td>
<td>16.67</td>
<td>379</td>
</tr>
<tr>
<td>20</td>
<td>crowding</td>
<td>20</td>
<td>10-20</td>
<td>local</td>
<td>5</td>
<td>300</td>
<td>12</td>
<td>4</td>
<td>24.81</td>
<td>347</td>
</tr>
<tr>
<td>20</td>
<td>random</td>
<td>20</td>
<td>10-20</td>
<td>local-seq.</td>
<td>25</td>
<td>300</td>
<td>1</td>
<td>18</td>
<td>16.58</td>
<td>389</td>
</tr>
</tbody>
</table>
Coevolutionary Arms Race

APT SCANNING TACTICS

SDN DECEPTION SYSTEM PARMS

Machine Learning

MACHINE LEARNING

AlFA

ANYSCALE LEARNING FOR ALL

RIVALS

Deception SDN

Flow Rule Generator

Analyze Flow Statistics

Configures Deceptive Virtual Network

Deception Server

Manipulate Virtual Network Traffic

Simulate Virtual Network Resources

Analyze Flow Statistics

Configures Deceptive Virtual Network

Machine Learning

MACHINE LEARNING
Coevolution of Scanning and Deception

Best Fitness Lockstep Coevolution

Defender does significantly worse vs an evolving attacker, thus beware of static assumptions.
Evolved for defense & attack
"From Biological Coevolution Towards Adversarial AI Via Artificial Coevolution"

- Biological arms races can provide adaptation
- Can coevolution help to improve robustness in other adversarial settings?
  - Multiple comparisons can aid robustness and improve diversity
  - Help to anticipate
  - Replay the arms-race
Adversarial AI Framework Concept

Learning Variants
- Coevolution
- Non-negative matrix factorization
- Gaussian processes
- Ensembles
DDoS Network Defense
RIVALS: Network Routing Problem

Defender Objective: maximize mission completion time and hops
Attacker Objective: maximize mission disruption attacks in number and duration

\[ f_a^T = \frac{1 - \text{mission}\_\text{success}}{(n_{\text{attacks}} \cdot \text{total\_duration}) + n_{\text{attacks}}} \]

\[ f_d^L = \frac{\text{mission}\_\text{success}}{\text{overall\_time} \cdot n_{\text{hops}}} \]

**ATTACKER ACTIONS**
- node, start time, end time
- complete loss of node

**DEFENDER ACTIONS**
- link flooding
- shortest path
- CHORD

**STRATEGY ADAPTATION & COEVOLUTION OF DDOS DEFENSE & ATTACKS**

**DOS Attack Campaign**
- Defense Strategy
- Mission Progress
- Network Metrics

**PEER TO PEER OVERLAY NETWORK TESTBED**
ALFA Sim of P2P

Fig. 9: Results from a Coevrun for network routing on logical topology on network topology 0. Top: Median and best fitness results for attacker population over 20 generations. Bottom: Median and best fitness results for defender population over 20 generations.

Defender finds an optimal solution
Network Segmentation

The Definitive Guide to Micro-Segmentation, John Friedman, CyberEdge Group
AVAIL: Enclaves vs Contagion

**FITNESS**
- mission delay
- budget remaining

**ATTACKER ACTIONS**
set strength and duration of attack for each enclave

**DEFENDER ACTIONS**
set tap sensitivity and size for each enclave

**OBJECTIVE:** minmax

**STRAEGY ADAPTATION & COEVOLUTION OF DEFENSE & ATTACKS**

Attack: strength, duration
Defense: per enclave, #devices, tap sensitivity
Mission delay
Budget use

**CONTAGION & DETECTION SIMULATOR**

Realistic parameters from a validated low level emulator
Evolve Defense

Defender learns sound network segmentation practices over generations.
Compendium Analysis

Parameters

Attacker Behavior

Fitness Evaluation

Engagement Environment

Measurements

Threat Environment

Objective

Objectives

Use Case

Attacker Population

Defender Population

MACHINE LEARNING

MACHINE LEARNING

Rivals

Dark Horse

AVAIL

Use Case

Compendium Analysis

Cache

SUBSET SELECTION
Then
ALL VS ALL ENGAGEMENTS

COMPRENDIUM
WITHIN SIDE COMPARISON

SOLUTIONS SELECTION

VISUALIZATIONS

ALFA

ANYSSCALE LEARNING FOR ALL

MIT CSAIL
Attack Campaign Performance Comparison
Different metrics and Ranking Schemes

(a) Attacker CS ranking scheme.

(b) Attacker PF ranking scheme.

(c) Attacker AF ranking scheme.

(d) Attacker MF ranking scheme.

Coev
IPC A
MinMax
rIPC A
Compendium
Same Run
Unseen
(a) Attacker pairwise distance. Black arrow shows attacker selected by AF, PF, and CS ranking scheme. The pink arrow shows the attacker selected by the MF ranking scheme.

(b) Defender pairwise distance. Black arrow shows defender selected by all ranking schemes.
Summary & Future Work

• Adversarial Engagements and Arms Races
• Network Security Arms Races
  – RIVALS Adversarial AI framework
    » RIVALS: Robustness vs Denial
    » AVAIL: Isolation vs Contagion
    » DARK Horse and ADHD: Deception vs reconnaissance

• Future
  – Validate, refine, and extend use cases