Decision Making and Learning in Cybersecurity: Lessons Learned

Coty Gonzalez (Dynamic Decision Making Laboratory, Carnegie Mellon University)

Contributors:

Varun Dutt
Ph.D. Student (2010-2012), Engineering and Public Policy – Now Assistant Professor, IIT, Mandi, India

Young-Suk Ahn

Neeraj Saini
Master student (2012-2013), Information Networking Institute – Now at Carnegie Mellon West

Noam Ben-Asher
Research Objectives

To establish a theoretical model of decision making in cyber-security situations that answers questions such as:

• How do humans recognize and process possible threats?
• How do humans recognize, process and accumulate information to make cyber-defense decisions?
• How do humans’ risk perception and tendencies to perceive rewards and losses influence their decisions in cyber-defense?

To provide a computational cognitive model of human decision making in cyber-security situations that:

• Addresses challenges of cyber-security while accounting for human cognitive limitations
• Provide measures of a human’s decision making and behavior
• Suggest approaches to investigate courses of action and the effectiveness of defense strategies according to the dynamics of cyber-security situations.
Research Approach

- **Laboratory Experiments:**
  - E.g., The “IDS security game”: Study the dynamic process of decisions from experience

- **Cognitive Modeling:**
  - Computational representations of human decision making process, mostly based on Instance-Based Learning Theory (IBLT, Gonzalez et al., 2003)
  - E.g., IBL models of stopping decisions: dynamic accumulation of evidence before an attack is declared

Involves comparison of data from: computational cognitive models and from humans, both performing the same task
A key lesson learned: Scaling up research paradigms, experimentation and modeling

- **Modeling detection with Instance-Based Learning Theory** (Dutt, Ahn, Gonzalez, 2011, 2012)

- **From Individual Decisions from Experience to Behavioral Game Theory: Lessons for Cyber Security** (Gonzalez, 2013)

- **Perspectives from Cognitive Engineering on Cyber Security.** (Cooke et al., 2012).

- **The Cyberwar Simulation Environment and Multi-Agent Models** (Ben-Asher, Rajivan, Cooke & Gonzalez, 2014; Ben-Asher & Gonzalez, in Prep).

Individual (Defender). Cognitive theories, Memory and individual behavior

Pair (Defender and Attacker). Interdependencies, Information, Behavioral Game Theory

Network (Multiple Defenders and Attackers). Behavioral Network Theory; Network science (& topology) Organizational Learning; Group Dynamics; Political and Social Science
Summarize main lessons learned throughout the project about:

1. Cyber security domain and research paradigms
2. Experimental studies involving human decisions, learning, and defense behavior
3. Computational models of decision making in cybersecurity
1. Cybersecurity domain and research paradigms

- Cybersecurity is complex and dynamic, and one of the most challenging domains to study the psychology of decision making
  - Who is the decision maker? Users, Defenders, Attackers, and other identities (e.g., insider threats)? Team and individuals?
  - What are the relevant security indicators for users and defend and attack strategies?
  - What are the motivations of decision makers in cyber security?
  - What type of experience is needed in cybersecurity? How is experience used in this domain?
- Develop relevant decision making research paradigms that abstract most essential elements of the cybersecurity environment that one would like to study
- Use relevant scenarios in decision making studies
Challenges of Cyber Security to decision making theories:

- **Unusually large and highly diverse workload**
  - Human cognitive capacity and processing are limited.

- **Highly uncertain and rapidly changing environments: reliance on experience**
  - Experience depends on memory and human memory often reflects the frequencies, recency, and features of events in the world.
  - Underweighting rare events may result in a belief that the likelihood of an attack is lower than it really is.
  - Lack of concrete and instant gratification in the cyber security context makes learning difficult.
  - Decision biases from Experience: Overconfidence, certainty, confirmation.

- **Dynamic risk on sequential information processing**
  - Risk tolerance and sequential information processing (temporal dependencies)
  - Complex interrelationships of seemingly unrelated events.
Relevant decision making research paradigms at the Individual Level

- The “Cyber Security IDS Game”
  Representation of a simplified IDS that presents network traffic and possible threats to a human analyst

Repeated Decisions from Experience

Score Descr. Wt. Pts
hit hit 1
miss miss -1
false alarm false alarm -5
correct rej. correct rej. 17
Trial Total 13
Cummul. Total 746

Should you protect path A or Path B?

A

B

$-40K

$-30K

$-40K

$-40K

$-40K

$-40K

$-40K

$-30K

$-30K

$-30K

$-30K
### Aggregated Feedback

Only TOTAL score shown

**Trial Total**: 17

**Cummul. Total**: 581

<table>
<thead>
<tr>
<th>Trial #</th>
<th>Is threat</th>
<th>Alert</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>No Alert</td>
<td>A user outside the company sends a packet #B to webserver, but #B fails.</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>#B has signature compromizing the webserver</td>
<td>A user inside the company sends a packet #B to the webserver, and #B succeeds.</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>No Alert</td>
<td>A user outside the company sends a packet #B to webserver, and #B fails.</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>No Alert</td>
<td>A user inside the company sends a packet #B to the webserver, and #B succeeds.</td>
</tr>
<tr>
<td>5</td>
<td>5</td>
<td>#B has signature compromizing the webserver</td>
<td>A user inside the company sends a packet #B to the webserver, and #B succeeds.</td>
</tr>
<tr>
<td>6</td>
<td>6</td>
<td>No Alert</td>
<td>A user outside the company sends a packet #B to webserver, and #B fails.</td>
</tr>
<tr>
<td>7</td>
<td>7</td>
<td>#B has signature compromizing the webserver</td>
<td>A user inside the company sends a packet #B to the webserver, and #B succeeds.</td>
</tr>
<tr>
<td>8</td>
<td>8</td>
<td>#B has signature compromizing the fileserver</td>
<td>A user inside the company sends a packet #B to the fileserver, and #B succeeds.</td>
</tr>
<tr>
<td>9</td>
<td>9</td>
<td>#B has signature compromizing the fileserver</td>
<td>A user inside the company sends a packet #B to the fileserver, and #B succeeds.</td>
</tr>
<tr>
<td>10</td>
<td>10</td>
<td>No Alert</td>
<td>A user outside the company sends a packet #B to webserver, and #B fails.</td>
</tr>
<tr>
<td>11</td>
<td>11</td>
<td>#B has signature compromizing the webserver</td>
<td>A user inside the company sends a packet #B to the webserver, and #B succeeds.</td>
</tr>
<tr>
<td>12</td>
<td>12</td>
<td>No Alert</td>
<td>A user outside the company sends a packet #B to webserver, and #B fails.</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>#B has signature compromizing the webserver</td>
<td>A user inside the company sends a packet #B to the webserver, and #B succeeds.</td>
</tr>
<tr>
<td>14</td>
<td>14</td>
<td>No Alert</td>
<td>A user outside the company sends a packet #B to webserver, and #B fails.</td>
</tr>
<tr>
<td>15</td>
<td>15</td>
<td>#B has signature of a malicious program</td>
<td>A user inside the company sends a packet #B to the fileserver, and #B succeeds.</td>
</tr>
<tr>
<td>16</td>
<td>16</td>
<td>No Alert</td>
<td>A user outside the company sends a packet #B to webserver, and #B fails.</td>
</tr>
<tr>
<td>17</td>
<td>17</td>
<td>File X in directory 'export' is changed</td>
<td>A user inside the company changes binary file X in directory 'export' on fileserver</td>
</tr>
<tr>
<td>18</td>
<td>18</td>
<td>File X in directory 'export' is changed</td>
<td>A user inside the company executes binary file X in directory 'export' on fileserver</td>
</tr>
<tr>
<td>19</td>
<td>19</td>
<td>File Y in directory 'export' is changed</td>
<td>A user outside the company changes binary file Y in directory 'export' on workstation</td>
</tr>
<tr>
<td>20</td>
<td>20</td>
<td>No Alert</td>
<td>A user outside the company changes binary file Y in directory 'export' on workstation</td>
</tr>
<tr>
<td>21</td>
<td>21</td>
<td>File Z has signature that runs a malicious program</td>
<td>A user inside the company changes binary file Z in directory 'export' on fileserver</td>
</tr>
<tr>
<td>22</td>
<td>22</td>
<td>No Alert</td>
<td>A user outside the company changes binary file Z in directory 'export' on fileserver</td>
</tr>
<tr>
<td>23</td>
<td>23</td>
<td>No Alert</td>
<td>A user outside the company changes binary file Z in directory 'export' on fileserver</td>
</tr>
<tr>
<td>24</td>
<td>24</td>
<td>No Alert</td>
<td>A user outside the company changes binary file Z in directory 'export' on fileserver</td>
</tr>
<tr>
<td>25</td>
<td>25</td>
<td>File Y has signature of a malicious program</td>
<td>A user inside the company changes binary file Y in directory 'export' on workstation</td>
</tr>
</tbody>
</table>
Detailed Feedback

Breakdown of scores shown: hits, misses, correct rejections, false alarms

<table>
<thead>
<tr>
<th>Score</th>
<th>Descr.</th>
<th>Wt.</th>
<th>Pts.</th>
<th>ID</th>
<th>Description</th>
<th>Details</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>hit</td>
<td>1</td>
<td>6</td>
<td>16</td>
<td>ftpd has stopped running on web server</td>
<td>The web server is running httpd services. The traffic is 3.3 Mbps between internet and web server, 3.3 Mbps between web server and file server, and 3.3 Mbps between web server and workstation.</td>
</tr>
<tr>
<td></td>
<td>miss</td>
<td>-1</td>
<td>3</td>
<td>17</td>
<td>ftpd has stopped running on web server</td>
<td>The web server is running ftpd and httpd services. The traffic is 3.3 Mbps between internet and web server, 3.3 Mbps between web server and file server, and 3.3 Mbps between web server and workstation.</td>
</tr>
<tr>
<td></td>
<td>false alarm</td>
<td>-1</td>
<td>2</td>
<td>18</td>
<td>ftpd has stopped running on web server</td>
<td>The web server is running ftpd and httpd services. The traffic is 3.3 Mbps between internet and web server, 3.3 Mbps between web server and file server, and 3.3 Mbps between web server and workstation. An ftpd operation has been executed.</td>
</tr>
<tr>
<td></td>
<td>correct rej</td>
<td>1</td>
<td>11</td>
<td>19</td>
<td>There is no traffic between file server and workstation and between file server and web server</td>
<td>The file server has stopped services. The traffic is 0 Mbps between web server and file server, and 0 Mbps between file server and workstation.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>20</td>
<td>There is no traffic between file server and workstation and between web server and workstation</td>
<td>The workstation has stopped services. The traffic is 0 Mbps between file server and workstation, and 0 Mbps between workstation and web server.</td>
</tr>
</tbody>
</table>

The events are colored accordingly
IDS events and alerts: A simple Scenario

- Simple topology
- 2 possible attacks
- But analysts could still get confused
Scenario: Events and Alerts

**Events**

<table>
<thead>
<tr>
<th>Event ID</th>
<th>Event</th>
</tr>
</thead>
<tbody>
<tr>
<td>Event 1</td>
<td>Mallory (i.e., the attacker) sends probing packet #B1 (after TCP 3-way handshake) to port 80 of webServer, but packet #B1 fails.</td>
</tr>
<tr>
<td>Event 2</td>
<td>Good packet #G1 gets into port 80 of webServer.</td>
</tr>
<tr>
<td>Event 3</td>
<td>Good packet #G2 gets into port 80 of webServer.</td>
</tr>
<tr>
<td>Event 4</td>
<td>Mallory sends probing packet #B2 to webServer, but packet #B2 fails.</td>
</tr>
<tr>
<td>Event 5</td>
<td>Good packet #G3 gets into port 80 of webServer.</td>
</tr>
<tr>
<td>Event 6</td>
<td>Good packet #G4 gets into port 80 of webServer.</td>
</tr>
<tr>
<td>Event 7</td>
<td>Good packet #G5 gets into port 80 of webServer.</td>
</tr>
<tr>
<td>Event 8</td>
<td>Mallory sends probing packet #B3 to webServer; packet #B3 succeeds.</td>
</tr>
<tr>
<td>Event 9</td>
<td>Mallory sends probing packet #B4 to the RPC port of fileServer, but packet #B4 fails.</td>
</tr>
<tr>
<td>Event 10</td>
<td>Good packet #G6 gets into the RPC port of fileServer.</td>
</tr>
<tr>
<td>Event 11</td>
<td>Mallory sends probing packet #B5 to the rpc port of fileServer; packet #B5 succeeds. The network is now in the state specified by Node 23.</td>
</tr>
<tr>
<td>Event 12</td>
<td>Good packet #G7 gets into the RPC port of fileServer.</td>
</tr>
<tr>
<td>Event 13</td>
<td>Good packet #G8 gets into the RPC port of fileServer.</td>
</tr>
<tr>
<td>Event 14</td>
<td>Good packet #G9 gets into the RPC port of fileServer.</td>
</tr>
<tr>
<td>Event 15</td>
<td>Binary file X in directory &quot;export&quot; is changed by a good user.</td>
</tr>
<tr>
<td>Event 16</td>
<td>Binary file X in directory &quot;export&quot; is changed by another good user.</td>
</tr>
<tr>
<td>Event 17</td>
<td>Mallory changes file X in directory &quot;export&quot; to install a Trojan horse.</td>
</tr>
<tr>
<td>Event 18</td>
<td>Binary file Y in directory &quot;export&quot; is changed by a good user.</td>
</tr>
<tr>
<td>Event 19</td>
<td>File X, the Trojan horse, is executed by admin. The Trojan horse has root privilege.</td>
</tr>
<tr>
<td>Event 20</td>
<td>Binary file Y in directory &quot;export&quot; is changed by another good user.</td>
</tr>
<tr>
<td>Event 21</td>
<td>File Y is executed by a regular user.</td>
</tr>
<tr>
<td>Event 22</td>
<td>Binary file Z in directory &quot;export&quot; is changed by another good user.</td>
</tr>
<tr>
<td>Event 23</td>
<td>File Z is executed by a regular user.</td>
</tr>
</tbody>
</table>

**Alerts**

IDS generates Alerts from Events. May contain noise (false alarms) and are generally imprecise.
A simplified but realistic network (Lye and Wing, 2005)

- A router routes internet traffic to and from the local network
- Firewall prevents unwanted connections
- The network has 2 subnetworks (one public web server and other private)
- The Web server runs an HTTP server and an FTP server for serving Web pages and data.
- It is accessible by the public through the Internet.
Types of attack and network scenarios

Example of an attack:
Stealing information from a workstation
1. Normal operation
2. Attack httpd service
3. Penetrate through the hacked httpd service
4. Install sniffer to collect passwords
5. Hack into a workstation
6. Steal information
7. Create additional damage by sabotaging the network

4 different types of cyber attacks which represent different intentions of an attacker:
   Denial of Service (DoS)
   Steal Information
   Install Sniffer
   Deface a website

Adapted from Lye & Wing (2005)
The cybersecurity IDS game

<table>
<thead>
<tr>
<th>Events</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>No Alert</td>
</tr>
<tr>
<td>2</td>
<td>#B has signature compromising the webserver</td>
</tr>
<tr>
<td>3</td>
<td>No Alert</td>
</tr>
<tr>
<td>4</td>
<td>No Alert</td>
</tr>
<tr>
<td>5</td>
<td>No Alert</td>
</tr>
<tr>
<td>6</td>
<td>No Alert</td>
</tr>
<tr>
<td>7</td>
<td>#B has signature compromising the webserver</td>
</tr>
<tr>
<td>8</td>
<td>No Alert</td>
</tr>
<tr>
<td>9</td>
<td>No Alert</td>
</tr>
<tr>
<td>10</td>
<td>No Alert</td>
</tr>
<tr>
<td>11</td>
<td>#B has signature compromising the fileserv</td>
</tr>
<tr>
<td>12</td>
<td>No Alert</td>
</tr>
<tr>
<td>13</td>
<td>#B has signature compromising the fileserv</td>
</tr>
<tr>
<td>14</td>
<td>No Alert</td>
</tr>
<tr>
<td>15</td>
<td>File X in directory 'export' is changed</td>
</tr>
<tr>
<td>16</td>
<td>No Alert</td>
</tr>
<tr>
<td>17</td>
<td>#B has signature of a malicious program</td>
</tr>
<tr>
<td>18</td>
<td>File X in directory 'export' is changed</td>
</tr>
<tr>
<td>19</td>
<td>File Y in directory 'export' is changed</td>
</tr>
<tr>
<td>20</td>
<td>File Z has signature that runs a malicious program</td>
</tr>
<tr>
<td>21</td>
<td>No Alert</td>
</tr>
<tr>
<td>22</td>
<td>No Alert</td>
</tr>
<tr>
<td>23</td>
<td>File Y has signature of a malicious program</td>
</tr>
<tr>
<td>24</td>
<td>No Alert</td>
</tr>
</tbody>
</table>
| 25     | No Alert    | A user outside the company executes binary file Y in directory 'export' on fileserv.
2. Experimental studies involving human decisions, learning and defense behavior

1. Detecting Cyber threats: Comparing experts and Novices (Ben-Asher & Gonzalez, 2015)
   a) Cyber security experience is reflected in both, theoretical and practical knowledge.
   b) Most important effect of experience involves identification of attributes used by experts in detecting threats

2. Detecting Cyber threats: Effects of similarity and feedback on detection success (Ben-Asher & Gonzalez, 2013)
   a) Detailed feedback improves detection of threats during training
   b) Detailed feedback is essential to be successful in NOVEL scenarios (different from what they are trained on).

3. Learning and Risk Taking in Information Security: Timing of events (Ben-Asher & Gonzalez, under review)
   a) Prior experience and risk tolerance interact in the successful detection of threats
   b) Experts tend to react on time in fast attacks if they are risk averse (they become risk averse in fast attacks)
   c) but they “over react” in slow attacks regardless of their risk tendency.
3. Computational models of decision making in cybersecurity

- A generic theory of learning and decisions from experience (Instance-Based Learning Theory, IBLT) represents the process of human decisions from experience in cybersecurity.
- These models reproduced human data obtained from experiments.
- An important challenge for the accuracy of predictions of an IBL model is the identification of relevant cues (attributes) in the instances.
- IBL models were extended to represent strategic and non-asymmetric interactions beyond the individual: in 2x2 games and in cyberwar (multiplayer) games.
A generic DDM cognitive process: Recognition, Judgment, Choice, Execution, Feedback

Formalization of representations:
- Instance: tripled: Situation, Decision, Utility (SDU)
- Relies on mathematical mechanisms proposed by ACT-R and on EV formulations from Economics

Computational representations are run in simulations: to provide concrete predictions of human behavior in various task types

Modeling platforms and examples are made available online:
http://www.hss.cmu.edu/departments/sds/ddmlab/downloads.html
1. Each experience combination is created as an instance in memory (e.g. A-10; N-8; A-1; N-5; A-5) when the outcome is experienced.

2. Each instance has a memory “activation” value based on frequency, recency, similarity, etc.

3. The probability of retrieving an instance from memory depends on activation.

4. For each option, memory instances are “blended” to determine next choice by combining value and probability.

5. Choose the option with the maximum blended value.

Example:

IBL model of choice at the individual level

(Gonzalez & Dutt, 2011; Lejarraga et al., 2012)

10
10
8
1
5
...
A formalization of this IBL model

1. **Each Instance** has an Activation: simplification of ACT-R's mechanism (Anderson & Lebiere, 1998):

\[
A_{i,t} = \ln \left( \sum_{t_i \in \{1, \ldots, t-1\}} (t - t_i)^{-d} \right) + \sigma \cdot \ln \left( \frac{1 - y_{i,t}}{y_{i,t}} \right)
\]

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Recency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Free parameters:</td>
<td>Free parameters:</td>
</tr>
<tr>
<td>(d): high (d)-More recency</td>
<td>(\sigma): high (s)-high variability</td>
</tr>
</tbody>
</table>

2. **Each Instance** has a probability of retrieval is a function of memory Activation (A) of that outcome relative to the activation of all the observed outcomes for that option given by:

\[
p_{i,t} = \frac{e^{A_{i,t}/\tau}}{\sum_j e^{A_{j,t}/\tau}}
\]

\[\tau = \sigma \cdot \sqrt{2}\]

3. **Each Option** has a Blended Value that combines the probability of retrieval and outcome of the instances:

\[
V_j = \sum_{i=1}^{n} p_i x_i
\]

4. **Choose** the option with the highest experienced expected value (“blended” value)
Risk tolerance and timing of adversarial threats: Modeling detection with IBLT

(Dutt, Ahn, Gonzalez, 2012)

- Instance structure with the possible values in an scenario

<table>
<thead>
<tr>
<th>type</th>
<th>Attribute name</th>
<th>Possible value</th>
<th>Actual code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Situation</td>
<td>IP (location)</td>
<td>Webserver</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Fileservlet</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Workstation</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Directory</td>
<td>Missing value</td>
<td>-100</td>
</tr>
<tr>
<td></td>
<td></td>
<td>FileX</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Operation</td>
<td>Successful</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Not successful</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>Alert</td>
<td>Present</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Absent</td>
<td></td>
</tr>
<tr>
<td>Decision</td>
<td>Decision</td>
<td>Cyber-attack (calculated indirectly based upon U)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>0</td>
</tr>
<tr>
<td>Utility</td>
<td>Threat</td>
<td>Yes</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>No</td>
<td>0</td>
</tr>
</tbody>
</table>
Current Game-theoretic Technologies don’t account for human characteristics: Theory of mind, trust, fairness, social value orientation.

Modeling detection with Instance-Based Learning Theory (Dutt, Ahn, Gonzalez, 2011, 2012)

From Individual Decisions from Experience to Behavioral Game Theory: Lessons for Cyber Security (Gonzalez, 2013)

Perspectives from Cognitive Engineering on Cyber Security. (Cooke et al., 2012).
**Relevant decision making research paradigms at the pair Level**

### Game Theory 2x2 Games

#### Prisoner’s Dilemma

<table>
<thead>
<tr>
<th>Player 1 Action</th>
<th>D</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-1, -1, 10, -10</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-10, 10, 1, 1</td>
<td></td>
</tr>
</tbody>
</table>

#### Chicken Dilemma

<table>
<thead>
<tr>
<th>Player 1 Action</th>
<th>D</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-10, -10, 10, -1</td>
<td></td>
</tr>
<tr>
<td>C</td>
<td>-1, 10, 1, 1</td>
<td></td>
</tr>
</tbody>
</table>

#### Repeated Decisions from Experience against attack/defend strategies

Repeated decisions and strategies in these games illustrate how players make decisions in strategic interactions, considering the potential outcomes and the strategies of their opponents.

### Simultaneous and Sequential Security Games – Two Person

- **Simultaneous Security Game:**
  - **Defender** has two strategies: `d1` and `nd`.
  - **Attacker** has two strategies: `a` and `na`.
  - Payoffs:
    - `d`: Defender chooses `d1` and Attacker chooses `a` results in $32, $0.
    - Defender chooses `nd` and Attacker chooses `a` results in $32, $30.
    - Defender chooses `d1` and Attacker chooses `na` results in $30, $32.
    - Defender chooses `nd` and Attacker chooses `na` results in $30, $0.

- **Sequential Security Game:**
  - The defender moves first by choosing either `d1` or `nd`.
  - The attacker then observes the defender's move and chooses `a` or `na`.
  - The payoffs are as follows:
    - If the defender chooses `d1` and the attacker chooses `a`, the respective payoffs are $32, $0.
    - If the defender chooses `d1` and the attacker chooses `na`, the payoffs are $30, $32.
    - If the defender chooses `nd` and the attacker chooses `a`, the payoffs are $32, $30.
    - If the defender chooses `nd` and the attacker chooses `na`, the payoffs are $30, $0.

---

*Should you protect path A or Path B?*
1. Game theory prescribes solutions (Nash-equilibria) that are often not in accordance with actual observed human behavior (learning, suboptimal):
   – Our approach: gather data that show actual human behavior over time and compare to Nash-equilibria solutions.

2. Traditional game theory often assume full information and ignore partial, asymmetric and gradual discovery of information from experience
   – Our approach: Create an “information continuum” where we systematically evaluate the effects of different information levels and types

3. Traditional game theory problems are formulated in terms of two individual decision makers and two actions that limit investigation of multiple-strategies and groups
   – Our approach: scaling-up the 2x2 games to study more complex games

4. Traditional game theory often ignore the integration of cognitive (e.g., memory, learning) and social factors (e.g., power, fairness)
   – Our approach: integrate social effects in cognitive process
Example:
IBL model of choice at the pair level

Gonzalez, Ben-Asher, Martin & Dutt, 2014

**Game Theory 2x2**

**Games**

**Prisoner’s Dilemma**

<table>
<thead>
<tr>
<th>Player 1 Action</th>
<th>D</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td>D</td>
<td>-1, -1</td>
<td>10, -10</td>
</tr>
<tr>
<td>C</td>
<td>-10, 10</td>
<td>1, 1</td>
</tr>
</tbody>
</table>

**Players: Defender and Attacker**

**IBL-PD**

- Experiential & Descriptive
  - An instance includes both players’ actions and outcomes 
    
    $[C, D, -10, 10], [C, C, 1, 1], [D, C, 10, -10], \text{ and } [D, D, -1, -1]$ 

- Adding the “other” outcome to the blending equation:
  \[
  V_j = \sum_{i=1}^{n} p_{ij} (x_{ij} + w_{ij})
  \]

- And how do humans weigh the “other” information into their own decisions? ($w=f(t)$)?
  - Dynamic adaptation of expectations \( w_t = 1 - \text{Surprise}_t \)
  - Surprise is a function of the gap between the expected outcome and the outcomes actually received:

\[
\text{Gap}_t = \text{Abs}[Vj - (Xj + 0)]
\]

\[
\text{Mean} (\text{Gap}_t) = \text{Mean}(\text{Gap}_{t-1}) \left(1 - \frac{1}{200}\right) + \text{Gap}(t) \left(\frac{1}{200}\right)
\]

\[
\text{Surprise}_t = \frac{\text{Gap}_t}{[\text{Mean}(\text{Gap}_t) + \text{Gap}_t]}
\]
Fitting the model’s parameters to data in a Prisoner’s Dilemma game
Repeated Decisions from Experience

- **N players** – Each player makes decisions whether to: Attack, Defend, do Nothing against each of the other players.
- Each player is characterized by cognitive attributes (memory, learning) and social attributes (Power) and economic attributes (Assets).
- Decisions are led by the goal of maximizing own assets, while accounting for social and cognitive attributes.
- Learning and decisions from experience are obtained through multi-round games.
- Decisions result in an Outcome (Gain or Loss) which changes the Assets available in the following round.
- Actions have a cost: Cost of attack, cost of defend, cost of doing nothing is zero.
Each active agent evaluates the other active agents, one at a time.

Each active agent is evaluated by calculating the possible outcome from attacking it.

Then the agent evaluates how likely it is to actually obtain that outcome.

Each agent selects to attack the agent that would yield the highest utility of attacking.

Makes a decision whether to attack or not, according to the highest blended value of the two types of actions “attack” or “no attack.”

Example: IBL model at the Network Level

Cyber War: multiple attackers/Defenders
Simulations and Results

• A network with 9 different types agents
  – Power (High, Medium, Low)
  – Asset Value (High, Medium, Low)
• Each network was simulated for 2500 trials.
• 60 simulations with the same network setting.
• Successful attack yields 20% of the opponent's assets
• Downtime - An agent without assets is suspended for 10 trials
• IBL Agents with $d=5$ and $\sigma = 0.25$
Power influenced the dynamics of agents’ state and the network heterogeneity
Conclusions

- Significant progress in the development of theoretical models of decision making in cyber-security situations. Theoretical models evolved from
  - Individual (Instance-Based Learning Theory)
  - Pair-level (Behavioral Game Theory and IBL-Game Theory)
  - Network Level (Network Theory and IBL-Network)
- Development of experimental paradigms that served to collect human data and conclude with behavioral phenomena:
  - IDS tool, Binary choice repeated decisions, Game theory games, CyberWar game
- Development of computational cognitive models based on theoretical developments including
  - IBL model
  - IBL-PD
  - IBL-CyberWar