Design of Dynamic and Personalized Deception: A Research Framework and New Insights for Cyber Defense

Cleotilde (Coty) Gonzalez

Dynamic Decision Making Laboratory
Social and Decision Sciences Department
Carnegie Mellon University
www.cmu.edu/ddmlab

This research was sponsored by the Army Research Office and accomplished under MURI Grant Number W911NF-17-1-0370.

.

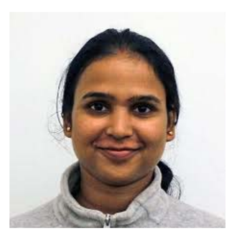
1

Collaborators

Christian Lebiere

Milind Tambe

Drew Cranford



Palvi Aggarwal

2

Emergence of Human Deception in very young children
(Evans & Lee, 2013)

- 65 children 2-3 years old
- Recorded, and asked whether they peeked
- Confessor: If they peeked and admitted peeking
- Lie teller: If they peeked but denied peeking

- 80% peeked (52/65)
- Of 52 peekers, 40% lied about having peeked
- Executive function skills play an important role in lie telling: Kids with higher cognitive capacity lie more
- Follow up studies show that older children lie more than younger children (younger children may lack the executive functioning skills to lie).

3

Deception is a principle of war

Sun Tzu, (Giles, 2005): All warfare is based on deception.
able to attack ➔ appear unable
when active ➔ appear inactive
when near ➔ make enemy believe we are far
…

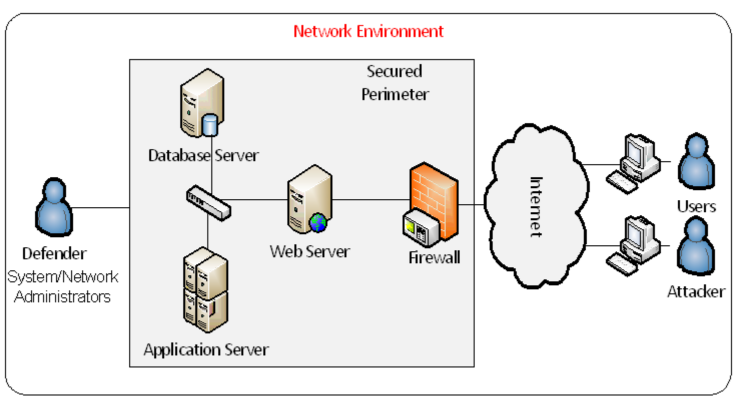
Decoy equipment (inflatable tank) used in WW II

National Archives and Records Administration, (2015) Deception in the cyber world:

The act of intentionally misleading through the strategic use of information (by inducing and suppressing signals) to cause behavioral changes on an agent that benefit the deceiver.

5

Deception in the cyber world



- If we are so good at deception why are we so trusting in cyber world? And why we cannot successfully deceive the attacker?
- Identities, actions, and intentions are easier to conceal in the cyberworld.

6

Deception-Based attack strategies

1. **Strategic manipulation of information**.
   a) Attention-catching strategies: high value targets; positive and negative values
   b) Use nudges: emergency, urgency, opportunity
   e.g., draws the phishing victim's attention away from the identity of the sender.
2. **Influence of trust, familiarity, similarity**
   a) We tend to trust things/people that are more familiar or similar to ourselves, share our own opinions.
   e.g. Spear phishing: impersonating someone familiar to us and we trust.
3. **Human cognitive experiential biases and context**.
   a) Framing effects (e.g., negative frames incite risk taking)
   b) Confirmation bias, gamblers' fallacy, misperception of randomness
   e.g., Search information that confirm our expectations.

7

Deception-Based cyber defense strategies

- Deception-based mechanism are also common for cyber defense (e.g., honeypots).
- Honeypots are used for detection to catch illicit interactions; in prevention, to assist in slowing attackers down; and many other defense possibilities.
- However, the effectiveness of honeypot techniques is questionable, as they often rely on static allocations that can often be easily discovered by attackers.
- Most of our cyber defenses remain static today. Attackers know it.
  - They can afford the time to engineer reliable exploits and plan their attacks because the targets do not change.
  - They can persist after a success inside a compromised network because the network does not change!

8

Goal: design dynamic and personalized effective defense strategies

By enhancing:

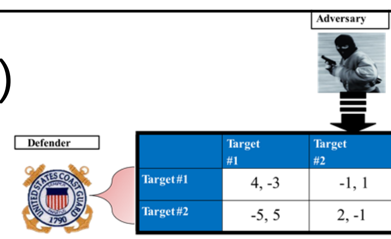
Game-theoretic approaches (Stackelberg Security Games) and algorithms for the optimization of limited resources of defense

With:

Behavioral laboratory experiments that elicit human attack and defend decisions and cognitive models that represent human behavior computationally.

9

Stackelberg Security Games (SSGs)



- In a SSG, there is a set of targets T ={t1; t2; : : : ; tn} which the defender protects by allocating K < n resources over them.
- A pure defense strategy is an allocation of the resources, with a mixed strategy being a randomization over these pure strategies. A mixed strategy represented as coverage probabilities over the targets, z = $\{Z_t\}$
- The attacker is aware of z (but not the pure strategy) and chooses a target t to attack accordingly.
- If the defender is protecting t, the attacker incurs a penalty and the defender is rewarded; If t is unprotected, the attacker gets a reward and the defender gets a penalty

Successful applications of the Strong Stackelberg Equilibrim (SSE): Optimize allocation of limited defense resources (Tambe's group)



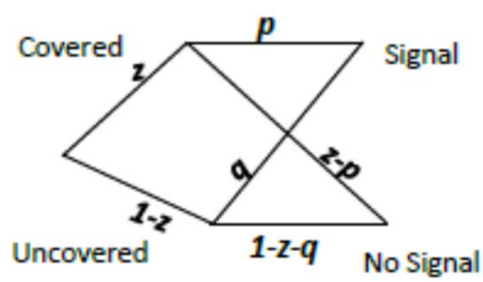



SSE with Persuasion (peSSE): (Xu et al., 2015)

A round of the two-stage SSG plays out as follows:

1. The defender allocates her resources, covering a random subset of the targets based on her mixed strategy z.
2. Aware of the defender's mixed strategy, the attacker chooses a target, t, to attack accordingly.

3. The defender sends a (possibly deceptive) signal to the attacker regarding the current protection status of t. Signaling scheme consists of probabilities (p & q) given coverage or not.
4. Based on the information given in the signal, the attacker chooses to either (1) continue attacking or (2) withdraw the attack yielding payoffs of zero for both players.



The optimal ("perfectly rational") act for the attacker is to always withdraw given a signal

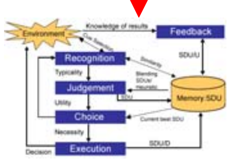
- **Our premise:**
  - These technical solutions may be more effective if they take advantage of the attacker's cognitive weaknesses (e.g., attacker's cognitive biases)
  - The "right balance" of deceptive and truthful signals depends directly on the **human attacker's beliefs**
  - To adjust the signal dynamically, we need a computational representation of the evolution of human beliefs.
- Our research program aims at advancing our understanding of how deceptive signals can be designed and presented to attackers in order to maximize their effectiveness, and how to develop computational models that predict human beliefs rather than relying on the assumption of perfect rationality.

13

A Research Framework for the Design of Adaptive and Personalized Deception

Attacker

- **Human participant** in learning (multiple rounds) experiments Making "attack" decisions



- **Instance-Based Learning Theory (IBLT):** Build IBL models of human experiential choice

Experimental Games

Defender

- **Stackelberg Security Games (SSGs):** Optimal Defender Resource Allocation Strategy
- **Signaling Theory (PeSSE):** Generating Deceptive Signals
- **Attacker's belief state predictions** (e.g., Confirmation Biases)

14

Experimental Games and Human Experiments

- To apply the game-theoretical solutions, we need to choose the right abstractions that isolate exactly the strategic issues of interest in cyber security.
- Insights on human behavior by studying "would-be" attackers in laboratory experiments.

Advantages and disadvantages

- Simplicity in modeling facilitates reasoning and allows a model to cover a broad class of relevant scenarios.
- But stylized models may be too generic and difficult to apply to particular solutions in cybersecurity.

15

Increasing complexity and realism of experimental games

The Box Game

Insider Attack Game

HackIT Simulation

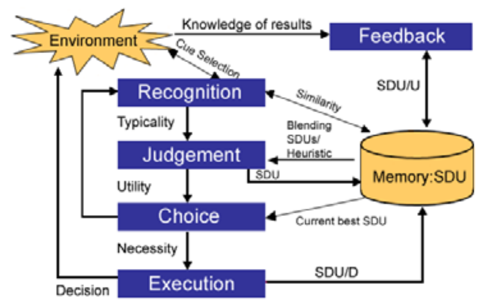
ExploitIT in CyberVAN

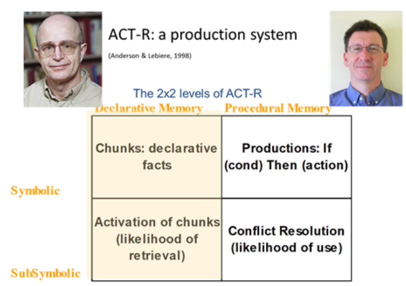
Complexity and Realism (increasing semantics)

Cognitive models of human dynamic decision making

(Gonzalez, Lerch, & Lebiere, 2003)

- The dynamics of human choice are captured by Instance-Based Learning Theory (IBLT): cognitive processes of Recognition, Judgment, Choice, and Feedback.
- IBLT relies on ACT-R's mathematical formulations of human memory processing.





17

Feedback

Experienced Utility

$$A_i = \ln \sum_{j=1}^{n} (t - t_j)^{-d} + MP * \sum_{k} Sim(v_k, c_k) + \varepsilon_i$$

$$P_i = \frac{e^{A_i/s}}{\sum_j e^{A_j/s}}$$

$$BV = \sum_i P_i \cdot V_i$$

Judgment

Expected Utility

Choice Alternative Context

Road	Smooth	High Traffic Probability	Action	Utility
A	-10	0.40	Drive	7.5

Recognition

partial match / partial match / partial match / partial match

Blended Value (BV)

Road	Smooth	High Traffic Probability	Action	Utility
A	-9	0.36	Drive	8

Choice

Road	Smooth	High Traffic Probability	Action	Utility
B	-2	0.25	Drive	9

Instances in Memory

Situation Decision Utility

Memory Instances (Unique combinations)

Road	Smooth	High Traffic Probability	Action	Utility
A	-9	0.36	Drive	8

Road	Smooth	High Traffic Probability	Action	Utility
A	-12	0.50	Drive	6

Road	Smooth	High Traffic Probability	Action	Utility
A	-9	0.60	Drive	7

Road	Smooth	High Traffic Probability	Action	Utility
B	-2	0.25	Drive	9

Example1 – The Box Game
Work in progress

Questions:

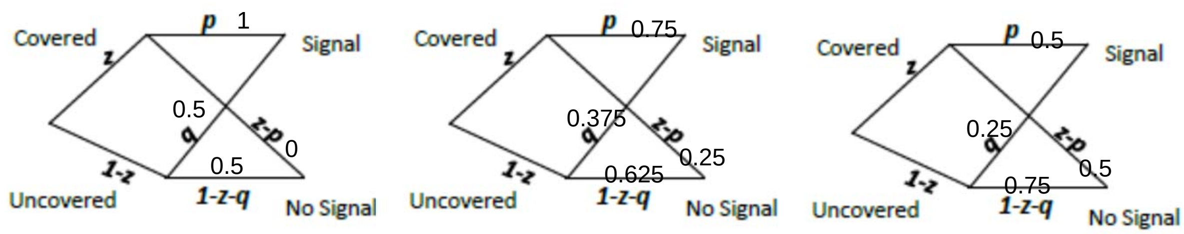
1) What effect has the frequency of signals on attack decisions?
2) What is the effect of the type of signal used (e.g., framing of the signal)?

The box game experiment

- **Player's Goal:** Find a "treasure" hidden in 2 identical boxes during each of 50 rounds.
- The **Defender (algorithm)** aims at protecting the boxes and prevent you from finding the treasure. But the **Defender** can **only protect one of the two boxes at a time.**
- The **Defender** sends **signals** stating whether the box is protected or not. *Sometimes may say that the box is protected when it is not; sometimes it may say that the box is not protected when in fact it is; and sometimes it may say nothing*.
- When you select a box you may receive a **signal**. Then, you can choose whether or not you want to actually select the box. If you choose:
  - **A box that is NOT protected**, you found the treasure! (win: 100 points);
  - **A box that IS protected**, you got caught (lose: -50 points);
  - **Not to attack** the box (0 points).

Experimental Methods

PeSSE – 1-way | PeSSE – 2-way- 0.75 | PeSSE – 2-way- 0.5



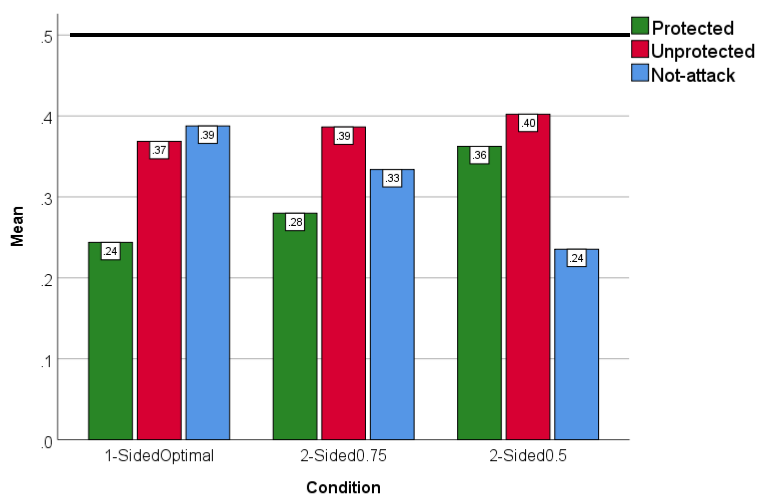
(Signal) Signal 1: This server may be protected by the Defender.
If you proceed, you may lose 50 points.

Signal 2: This server may **NOT** be protected by the Defender.
If you proceed, you may win 100 points.

3 experimental conditions of manipulation of the probability of sending a truthful signal when protected: 1-way, 2-way(0.75), and 2-way(0.50) (100 participants in each condition). All conditions use optimal allocation of defense resources (50% protection probability), both nodes are of same value (+100/-50). EV(Signal)=0

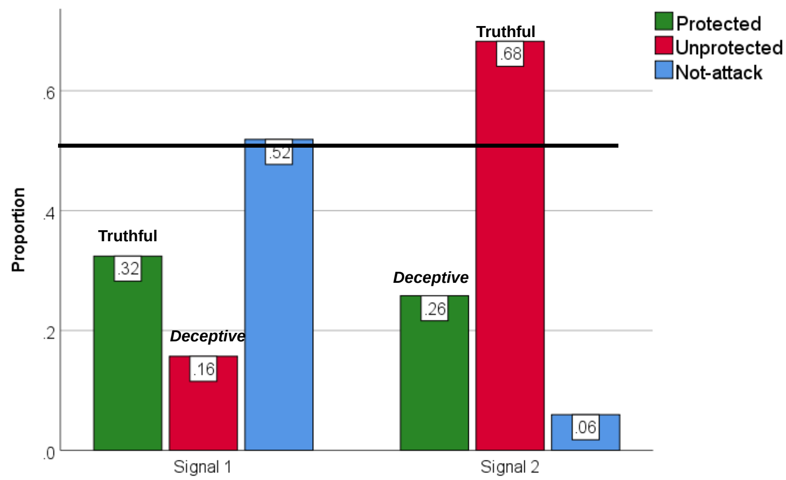
21

Signaling Works and Frequency of Signaling Matters



1. The proportion of attacks was reduced (from 0.5) in all conditions through signaling.
2. The signaling frequency has a significant effect on attacks to protected nodes: reducing the frequency of signal increases attacks to protected nodes
3. And it decreases no-attacks decisions

Type of signal matters



Signal 1: This server may be protected by the Defender.
If you proceed, you may lose 50 points.

Signal 2: This server may **NOT** be protected by the Defender.
If you proceed, you may win 100 points.

1. In general, a *deceptive* signal generates less attacks than a truthful one.
2. But Signal 2 (i.e., "hopeful" signal) produces largely more attacks when truthful than deceptive compared to Signal 1 (i.e., "suspicious" signal).
3. Signal 1 causes significantly more deterrence than Signal 2.

Example2 – Insider Attack Game

Questions:

1) How do humans react to deceptive signals?
2) What is the right "balance" of signal frequency?
3) Can we use cognitive models of human behavior to develop more effective signaling schemes?

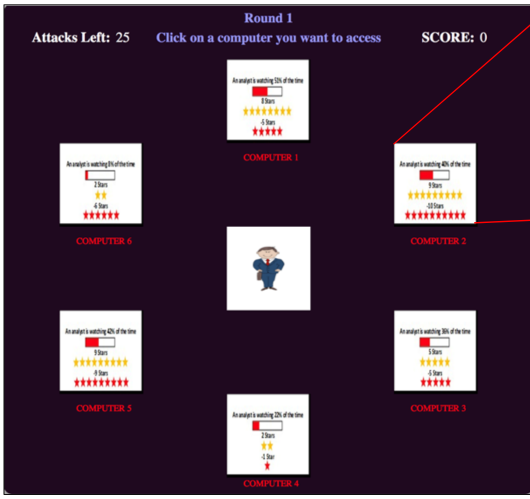
Cooney, S., Wang, K. Bondi, E., Nguyen, T., Vayanos, P., Winetrobe, H., Cranford, E. A., Gonzalez, C., Lebiere, C., Tambe, M. (2019). Learning to Signal in the Goldilocks Zone: Improving Adversary Compliance in Security Games. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2019). September 16-20, 2019, Würzburg, Germany.

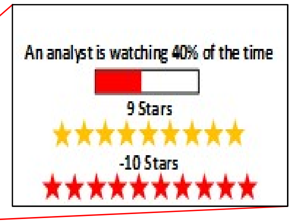
Cranford, E. A., Gonzalez, C., Aggarwal, P., Cooney, S., Tambe, M., Lebiere, C. (2019). Towards personalized deceptive signaling for cyber defense using cognitive models. In Proceedings of the 17th Annual Meeting of the International Conference on Cognitive Modelling. Montreal, CA.

Cranford, E A., Lebiere, C., Gonzalez, C., Cooney, S., Vayanos, P., & Tambe, M. (2018). Learning about Cyber Deception through Simulations: Predictions of Human Decision Making with Deceptive Signals in Stackelberg Security Games. 40th Annual Meeting of the Cognitive Science Society (CogSci 2018). July 25-28, 2018, Madison, WI.

24

Insider Attack Game – PeSSE 1-way deception

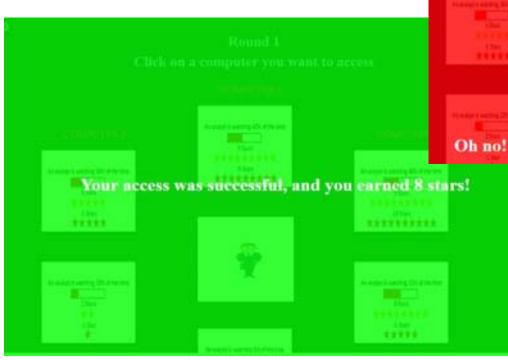




25

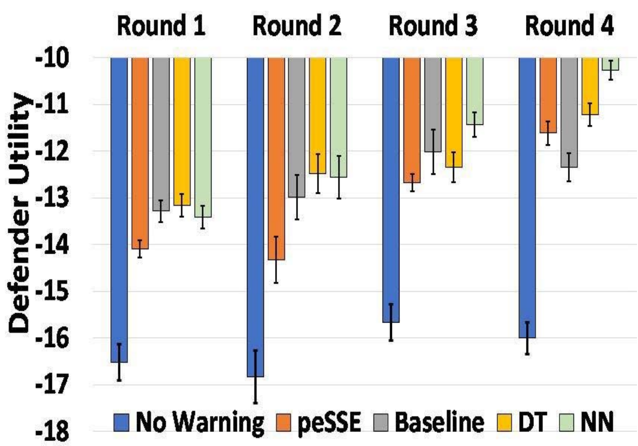
PeSSE 2-way deception

Truth…



or Deception

Signaling Works and Frequency of Signaling Matters



There is a significant benefit to the defender when using signaling against boundedly rational attackers compared to using no signaling, or when using the peSSE algorithm.

All three 2-way signaling schemes outperformed the peSSE algorithm: reducing the frequency of signaling improves performance against boundedly rational attackers.

A ***Goldilocks Zone***: lowering the signaling frequency can increase compliance with regard to signals, but must be carefully balanced so that instances in which no signal is shown do not offset the gain to the defender.

Cooney, S., Wang, K. Bondi, E., Nguyen, T., Vayanos, P., Winetrobe, H., Cranford, E. A., Gonzalez, C., Lebiere, C., Tambe, M. (2019). Learning to Signal in the Goldilocks Zone: Improving Adversary Compliance in Security Games. The European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases (ECML PKDD 2019). September 16-20, 2019, Würzburg, Germany.

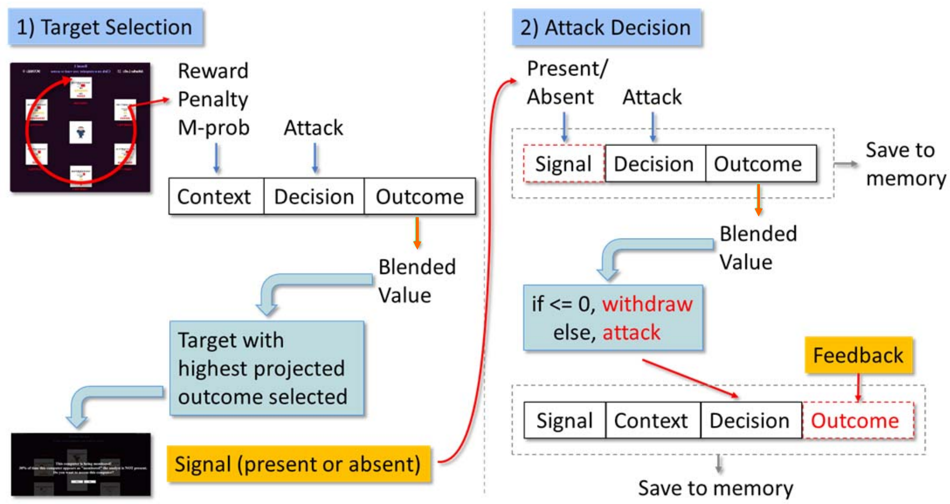
27

Using the cognitive model to inform the signal rate for a particular individual

The “right balance” of deceptive and truthful signals depends directly on the **human attacker’s beliefs**

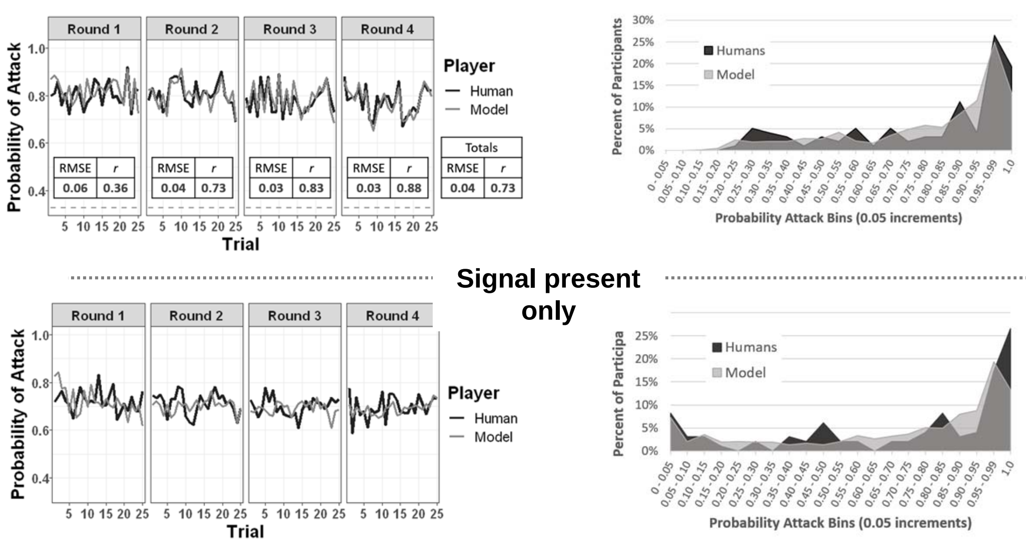
28

IBL Model Procedure



29

Mean Probability of Attack



30

Confirmation Bias/Hot-Stove effect

- Human tendency to seek evidence that confirms one's beliefs
  - People do not test their beliefs about the world by trying to disconfirm them, but rather, by trying to confirm them
- Hot-Stove effect produces a "win-stay"/"lose-shift" behavior.

- Experiences of rewards when a signal is present increases the probability of attacking in the future, while experiences of penalties given a (deceptive) signal reduces the probability of attacking in the future.
- Eliminating deceptive signals restores belief in the signal.
- The goal for the cognitive signaling scheme is to induce, and preserve, the belief that attacking given a signal will result in a loss.

31

New Cognitive Signaling Scheme

- Relying on individualized memory instances estimate the expected utility (through Blending) of attacking given a signal E(A|S) and not, E(A|$\bar{S}$).

If selected target is **covered**:

If $E(A|S) > E(A|\bar{S})$ → Signal

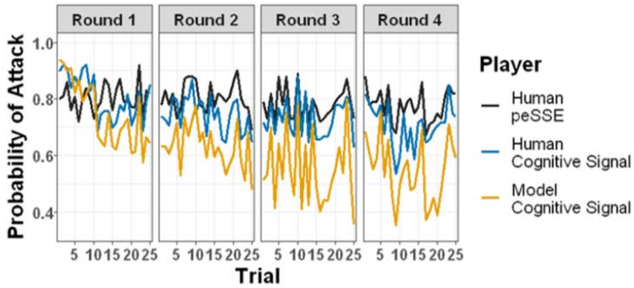
Else → No Signal

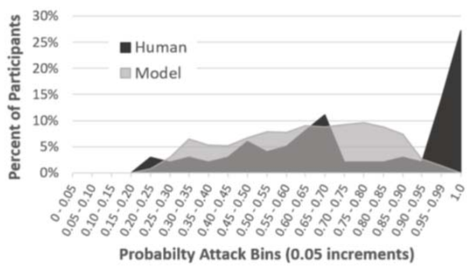
If selected target is **not covered**:

If $E(A|S) > E(A|\bar{S})$ → No Signal

Else → Signal

32

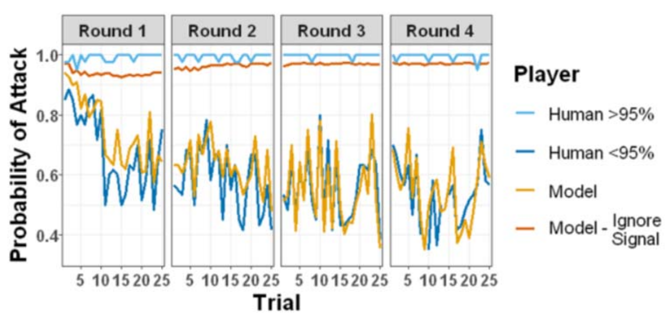




Compared to PeSSE, cognitive signaling reduces the probability of attack.
But the model predicts sharper reductions in the probability of attack than what humans actually do

The model fails to account for approximately 44% of participants that attacked at a rate of 95% or more.

What is going on?



the model is highly accurate at predicting performance of the approximately 56% of participants that attack at a rate less than 95%.

post-experiment survey: a majority of participants that attacked more than 95% responded that they ignored the signal.

we created a version of the cognitive model that does not consider the signal when generating an expected outcome of attacking the selected target.
Blending of instances IGNORES the signal.

The model attacks on 96.0% of trials (*SD* = 15.1%), matching well to the distribution.

Example 3 – HackIt

Questions:

1) What is the effect of honeypot allocation/distribution on the attacker's performance?
2) What reconnaissance strategies do attacker's follow?

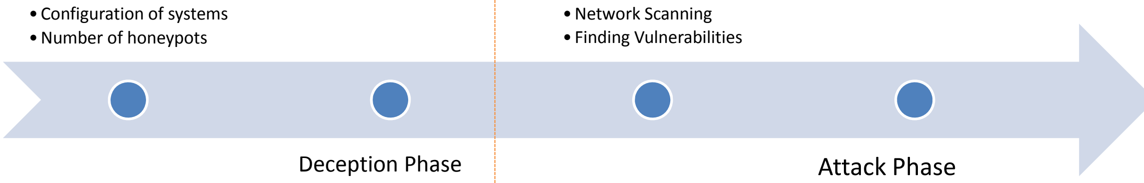
Aggarwal, P., Gautam, A., Agarwal, V., Gonzalez, C., & Dutt, V. (2019). HackIT: A Human-in-the-loop Simulation Tool for Realistic Cyber Deception Experiments. *10h International Conference on Applied Human Factors and Ergonomics*, Orlando, Florida, USA

Defender

Attacker

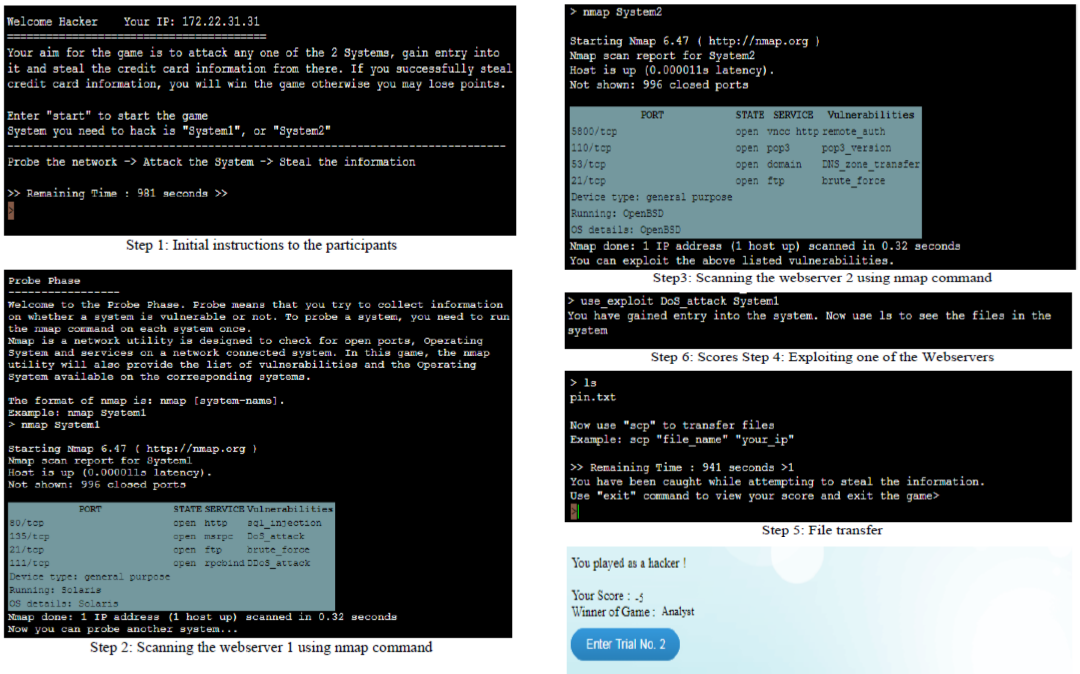
Configuration Phase
- Network size
- Topology
- Configuration of systems
- Number of honeypots

Probe Phase
- Network Scanning
- Finding Vulnerabilities



Deception Phase
- Definition of honeypot
  - Content
  - Configuration
- Timing of deception
- Amount of deception

Attack Phase
- Exploiting systems
- Gaining Access
- Stealing information
- Destroying infrastructure



Step 1: Initial instructions to the participants

Step 2: Scanning the webserver 1 using nmap command

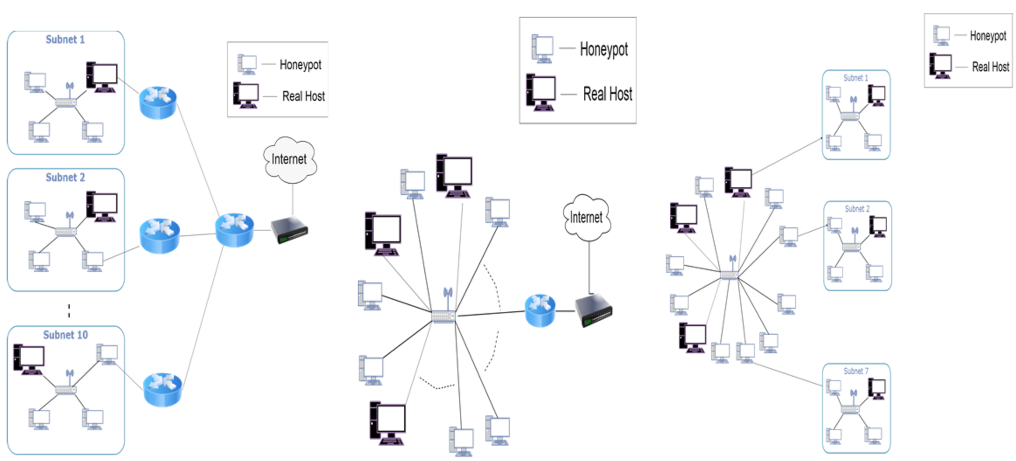
Step3: Scanning the webserver 2 using nmap command

Step 6: Scores Step 4: Exploiting one of the Webservers

Step 5: File transfer

Experimental Conditions

Reconnaissance Deceptive Server (RDS) non-RDS mixed configuration

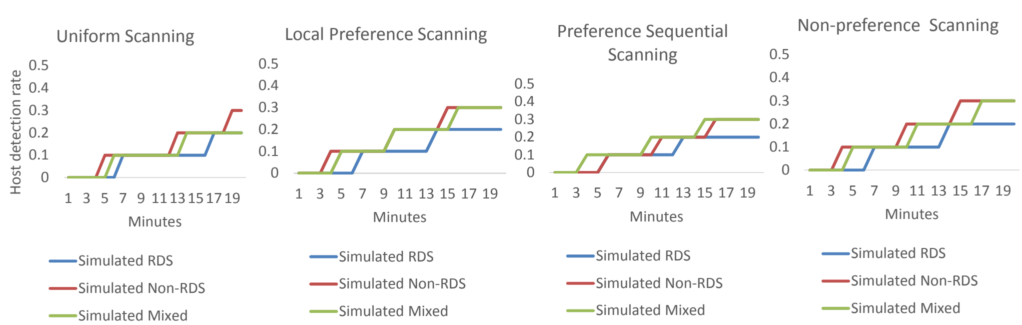


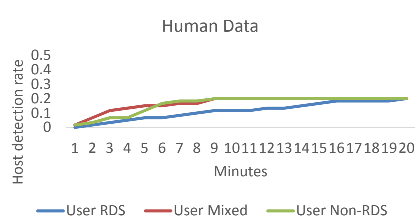
Reconnaissance Strategies

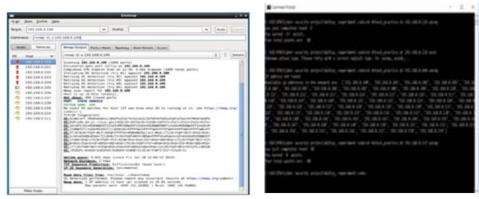
- Achleitner et al. (2016) simulated following ***reconnaissance strategies*** in deceptive and non-deceptive networks:
  - Uniform Scanning
  - Local Preference Scanning
  - Preference Sequential Scanning
  - Non-Preference Sequential Scanning
  - Preference Parallel Scanning

Achleitner, S., La Porta, T., McDaniel, P., Sugrim, S., Krishnamurthy, S. V., & Chadha, R. (2016, October). Cyber deception: Virtual networks to defend insider reconnaissance. In *Proceedings of the 8th ACM CCS international workshop on managing insider security threats* (pp. 57-68). ACM.

Results: RDS has lower detection rate – but no difference in reconnaissance strategies



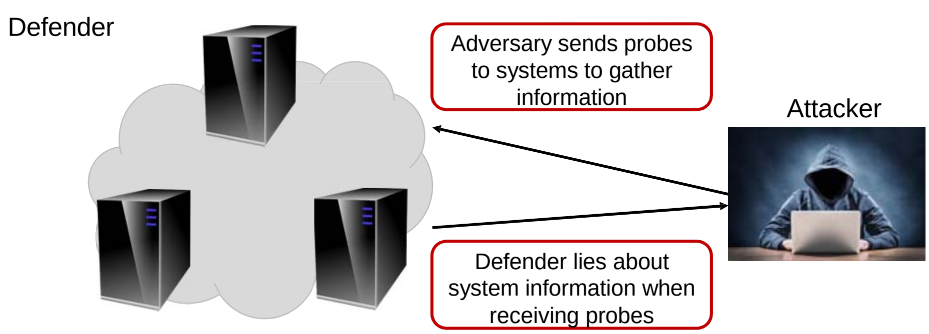




Example 4 – CyberVan
Work in Progress

Questions:

1) What is the effect of an optimal masking strategy?
2) What reconnaissance strategies do attacker's follow?



- Sets a <u>True Configuration</u> (TC)
- <u>And Observable Configurations</u> (OC)
- Choose the OC for each TC, masking constraints and
- Cost function for masking TC with an OC

- Views OC of systems with scanning; observes state of network
- Attacks systems according to OC

	OC			
TC	*freeBSD*	*Win2008*	*Openwrt*	*Ubuntu8*
avayagw	3	0	0	0
Ubuntu8	2	0	0	0
Win7pro	0	2	0	0
Win7ent	0	2	0	0
WinXP	0	2	0	0
Slackware	0	0	0	1

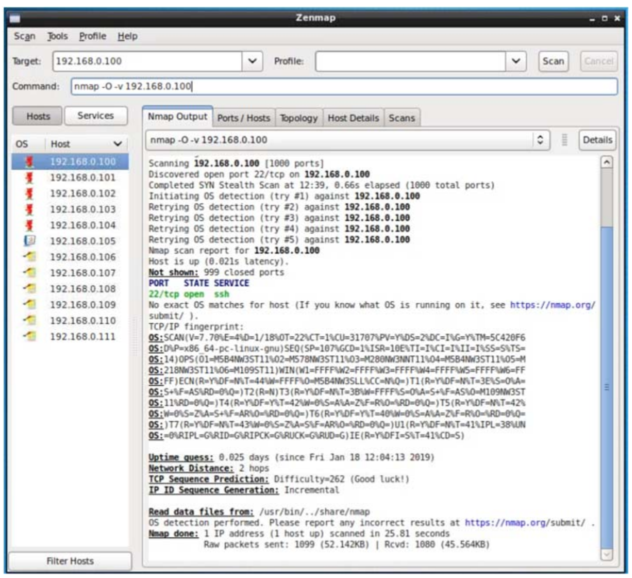
There are 4 Observable Configurations (OCs) and 6 True Configurations (TCs)

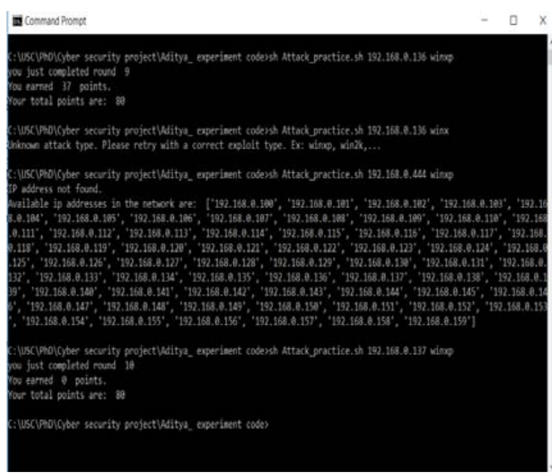
TCs are mapped to Ocs:

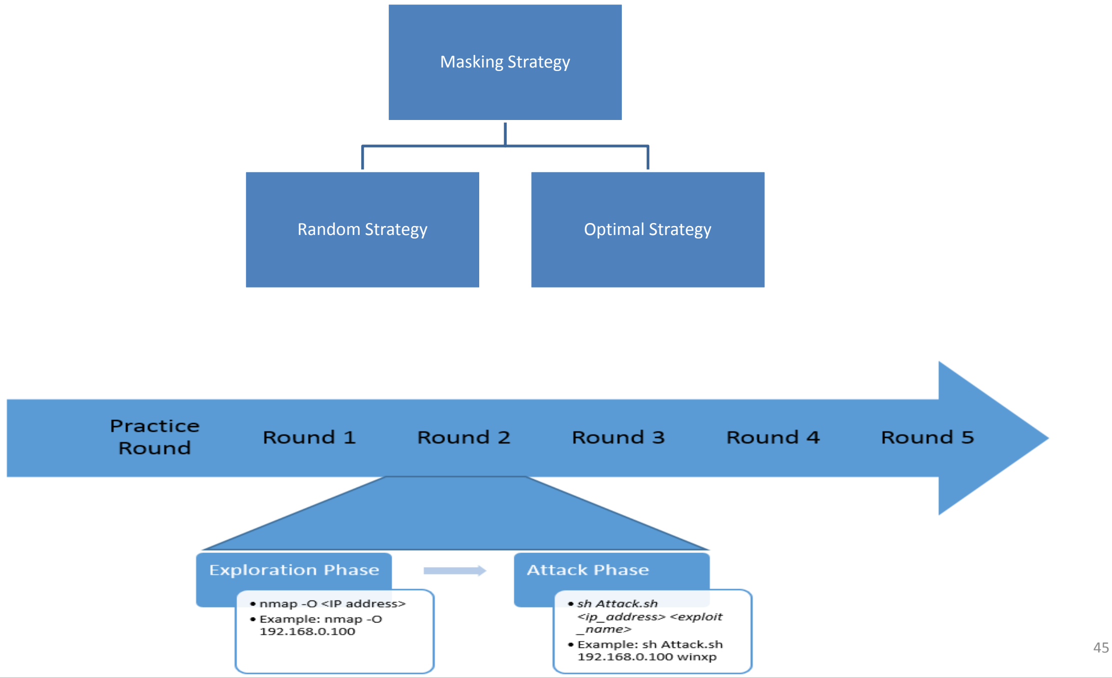
- 5 machines are shown as freebsd, out of which 3 are actually avayagw and 2 are ubuntu8
- 6 machines are shown as win2008, out of which 2 are win7pro, 2 are win7ent, and 2 are winxp
- 1 machine is shown as ubuntu8, which is actually slackware

Based on this information attacker may decide which machine to attack.

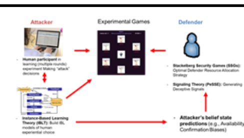
Task in CyberVAN –Perspecta Labs







Conclusions



- Our research program contributes to SSGs research by providing:
  - insights from human experiments regarding human trust to truthful or deceptive signals
  - creating cognitive models that emulate attacker's behavior
- These models help in the design of dynamic and personalized deception strategies
- Across levels of complexity in interactive security games and using the insights of cognitive models of attacker behavior, we find that:
  1. signaling algorithms optimized for perfectly rational attackers improve defense compared to no signaling at all;
  2. humans behave far differently than predicted under the assumption of perfect rationality
  3. humans exhibit boundedly rational behaviors that result in cognitive biases (e.g., confirmation bias)
  4. new adaptive and personalized theories that increase attacker's compliance are possible through cognitive modeling and human-in-the-loop experiments
  5. Model fits average behavior and individual distribution of actions.
- Extending our cognitive models to accommodate greater complexity will enable the models to capture the richness of realistic cyber-security situations.

Thank you!

Questions?

47