Adversarial Classification

July 8, 2015
IPAM Summer School
Games and Contracts for Cyber-Physical Security

John Musacchio
University of California, Santa Cruz
johnm@soe.ucsc.edu

Work with :

Lemonia Dritsoula and Braden Soper
University of California, Santa Cruz

Patrick Loiseau
Eurecom, Sophia Antipolis, France

Strategic classification applications

- Spam detection: spammer changes
- frequency of words to evade spam filters e.g. ‘medical marijuana’

- Fraud detection: credit card thief changes
- the expenses/day with stolen credit card

- Click fraud (advertising): botnet might click on random ads on other websites
- …

Thief or fox?

precious goats

animal thief

OR?

cheaper chickens

fox

“poor” shepherd

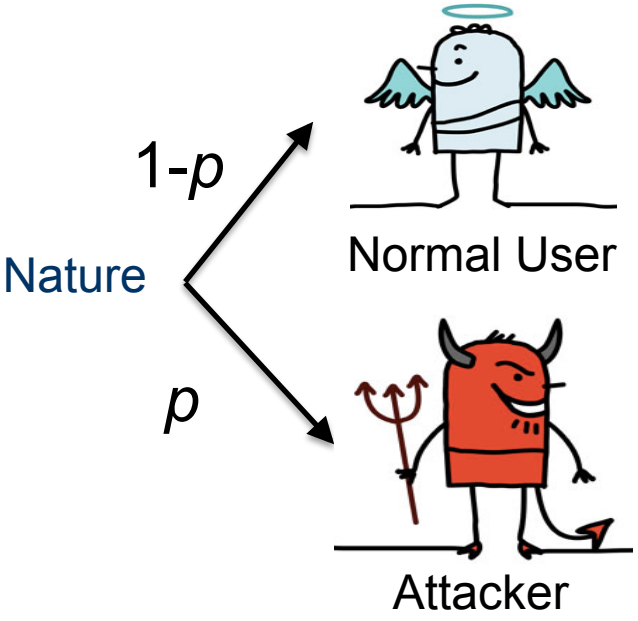
UC SANTA CRUZ

Overview

- **Part I: Adversarial classification [Dritsoula, Loiseau, Musacchio]**
  - Single-feature threshold-based classification [CDC 2012]
  - Generalized payoff structure [GameSec 2012]
  - Optimal defense and attack strategies [to be submitted to TIFS 2015]

- **Part 2: Botnet detection games [Soper, Musacchio]**
  - Homogeneous Agents [Allerton 14]
  - Heterogeneous Agents [NetGCooP 14]

Game Setup

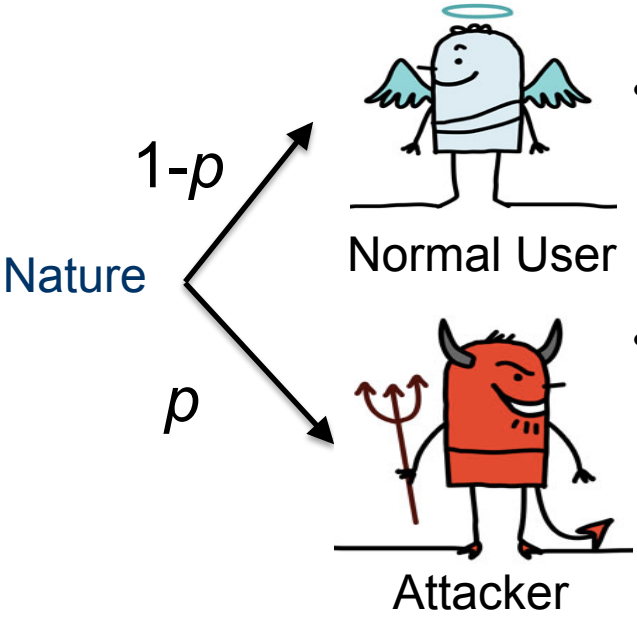


- Non Strategic
- Behaves Randomly
- Generates feature vector $v \in V$, has prob. distribution P_N

- Strategic.
- Chooses feature vector $v \in V$

- Chooses classifier $c \in C$
- Each classifier c maps feature vectors to binary alarm decision ($V \rightarrow \{0,1\}$)

Payoff Functions



- Generates feature vector $v \in \mathsf{V}$, has prob. distribution P_N
- Chooses $v \in \mathsf{V}$

indicator that attack detected

$$U^A(v,c) = -c_d 1_{c(v)=1} + R(v)$$

cost of detection

attack reward

- Chooses $c \in \mathsf{C}$

Defender

$$U^D(v,c) = p \cdot (c_d 1_{c(v)=1} - R(v))$$

payoff if attacker present

$$-(1-p)c_{fa} \sum_{v' \in \mathcal{V}} P_N(v') 1_{c(v')=1}$$

expected false alarm cost

Some related work: adversarial classification

- **[Sommer&Paxson, 2010]:** identifies reasons why Machine Learning algorithms do not work well in adversarial settings

- **[Dalvi et al., KDD '04]**: attackers know classifier
  try to minimize the cost from deviating to defeat classifier
  - (Best response iteration)

- **[Globerson, Roweis et al. 2006,2011]**: focus on worst-case optimization which is over pessimistic

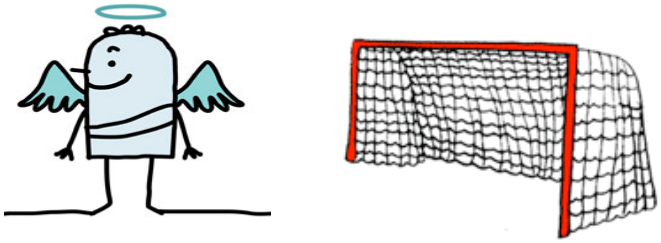
- **[Vorobeychik 2014]**: Identify advantage of randomized classification; study defender choosing between two classifiers.

- **[Brueckner & Scheffer, 2009,11]**: adversarial learning problem and identify conditions on which unique *pure* NE exist. Different setting

- **[Patcha et al., 2006]:** Signaling, multi-stage game (Bayes rule updates)

Mixed Strategies

- Need to turn to mixed strategies to find equilibrium
  - Attacker (usually) “wants” to choose feature vectors that don’t get classified
  - Defender (usually) “wants” to map what attacker does to “detect” and other things to “not-detect”



Defender Strategy Space Reduction

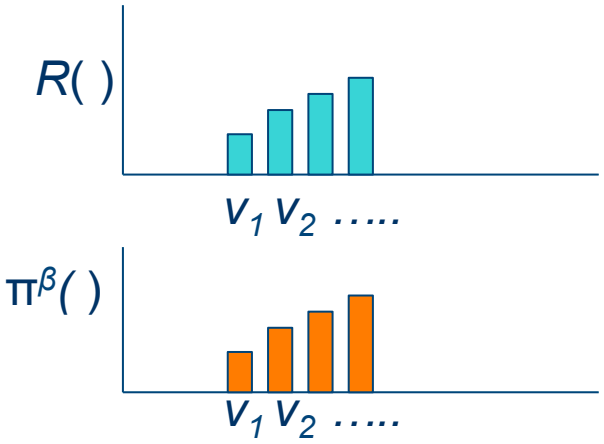
- Defender's strategy space
  - Defender picks β, a probability distribution on C
    - $2^{|V|}$ dimensional real vector
  - Payoffs of both players depend on β only through the alarm prob. for each feature vector

$$\pi^{\beta} : V \rightarrow [0,1]$$

    - $|V|$ dimensional vector
- Can prove, $\pi^{\beta}( )$ can be made to be any arbitrary function $V \rightarrow [0,1]$ with some choice of β
- Thus, can think of defender choosing π^{β} directly

Monotonicity in attacker reward

- Fact: in NE, for all v_1, v_2 in the support of the attacker with $R(v_1) \leq R(v_2)$, $\pi^\beta(v_1) \leq \pi^\beta(v_2)$
  - Suppose not , i.e. $\pi^\beta(v_1) > \pi^\beta(v_2)$
    - Then attacker could shift all weight from v_1 to v_2 improve reward, and reduce detection cost
    - Defender would then want to reduce $\pi^\beta(v_1)$ to reduce false alarm cost…



Defender – Mixed Thresholds

- Any monotone π^{β} can be achieved by mixing on threshold strategies
- Thus, we can restrict our attention to mixed threshold strategies for the defender

Threshold Re-Formulation

- Attacker picks attack vector index

$$I \in \{1, \ldots, N\}$$

  - where w.l.o.g. $R(v_1)$, $R(v_2)$ is an increasing sequence

- Defender picks a threshold

$$T \in \{R(v_1), \ldots, R(v_N), \infty\}$$

- Attacker reward:

$$U^A = c_d 1_{T \le I} - R(v_I)$$

- Defender reward:

$$\tilde{U}^D = p\,(c_d 1_{T \le I} - R(v_I)) - (1-p)\, c_{fa}\, P_N(R(\mathbf{v}) \ge T)$$

Random obs. -under normal user

⇩ Rescale

$$U^D = c_d 1_{T \le I} - R(v_I) - (p^{-1}-1)\, c_{fa}\, P_N(R(\mathbf{v}) \ge T)$$

Vector Matrix Formulation

- α : p.m.f. of attacker on feature vectors
  - (listed in increasing $R(v)$ order)
- β : p.m.f. of defender on thresholds $\{R(v_1), \ldots, R(v_N)\}$
- Attacker's cost

  $\alpha^T \Lambda \beta$
- Defender's Reward

  $\alpha^T \Lambda \beta - \mu^T \beta$

$$\mu = \left(\frac{1}{p} \; -1 \right) c_{\text{fa}} \left[\; \mathrm{P}\left(R(\boldsymbol{v}) > R(v_1)\right) \quad \cdots \quad \mathrm{P}\left(R(\boldsymbol{v}) > R(v_N)\right) \quad 0 \; \right]^T$$

$$\Lambda = c_d \begin{bmatrix} 1 & \cdots & 0 & 0 \\ 1 & \ddots & 0 & 0 \\ 1 & \cdots & 1 & 0 \end{bmatrix} - \begin{bmatrix} R(v_1) \\ \vdots \\ R(v_N) \end{bmatrix} 1_{N+1}^T$$

Zero-Sum Strategic Equivalence

Attacker Minimizes

$\alpha^T \Lambda \beta \quad - \mu^T \beta$

Defender Maximizes

$\alpha^T \Lambda \beta - \mu^T \beta$

- Including a term that attacker can't control doesn't effect his best responses
- Thus, this modified, zero-sum game is strategically equivalent

Linear Program

$$\begin{aligned} \max_{\beta}\min_{\alpha} \quad & \alpha^T \Lambda \beta - \mu^T \beta \\ \text{subject to} \quad & \mathbf{1}^T \alpha = 1,\ \mathbf{1}^T \beta = 1 \\ & \alpha \geq 0,\ \beta \geq 0 \end{aligned}$$

$$\begin{aligned} \max_{\beta} \quad & \min\left[\Lambda \beta\right] - \mu^T \beta \\ \text{subject to} \quad & \mathbf{1}^T \beta = 1,\ \beta \geq 0 \end{aligned}$$

$$\begin{aligned} \max_{\beta} \quad & z - \mu^T \beta \\ \text{subject to} \quad & z\,\mathbf{1} \leq \Lambda \beta \\ & \mathbf{1}^T \beta = 1,\ \beta \geq 0 \end{aligned}$$

Nash Equilibria Analysis

- Specific instances can be numerically solved easily with LP
- Can do further analysis to get more structural insight

Some Nash Equilibrium Insights

- Defender's support contiguous from some min. threshold up to "max attack" ($R(v_N)$)
  - Might also include ∞ threshold (i.e. never alarm)
- Attacker's support starts within one of defender's up to "max attack"
- Proof technique:
  - LP solutions are convex combos of extreme points
  - Manipulate inequalities describing feasible region

Some Nash Equilibrium Insights (2)

- For adjacent indices in defender support, attacker needs to make defender "indifferent"
  - Defender increasing threshold from *i* to *i*+1 decreases false alarm penalty by

$$\propto P_N(v_i)$$

  - Missed detection cost must go up proportional to this amount too.
  - Can show this requires

$$\alpha_i = \frac{1-p}{p}\frac{c_d}{c_{fa}}P_N(v_i)$$

Some Nash Equilibrium Insights (3)

- Similarly, attacker "needs" to keep defender indifferent on her support
  - Can show this requires

$$\beta_i = \frac{R(v_i) - R(v_{i-1})}{c_d}$$

Nash equilibrium computation

- Search over possible starting point of the support
  - For each starting point, 2 possible forms differing at the ends
- Can be “ties” in which case, equilibria defined by convex hull.

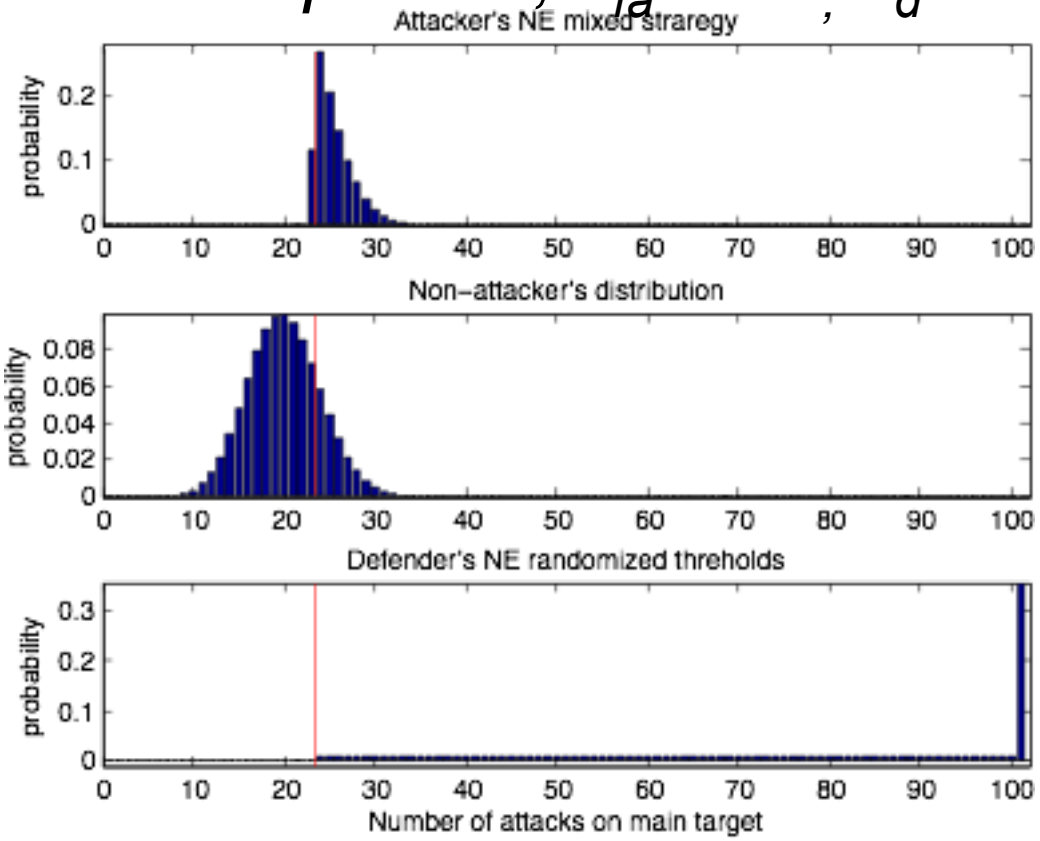
Example

$V = \{ 0, 1, \ldots, 100 \}$

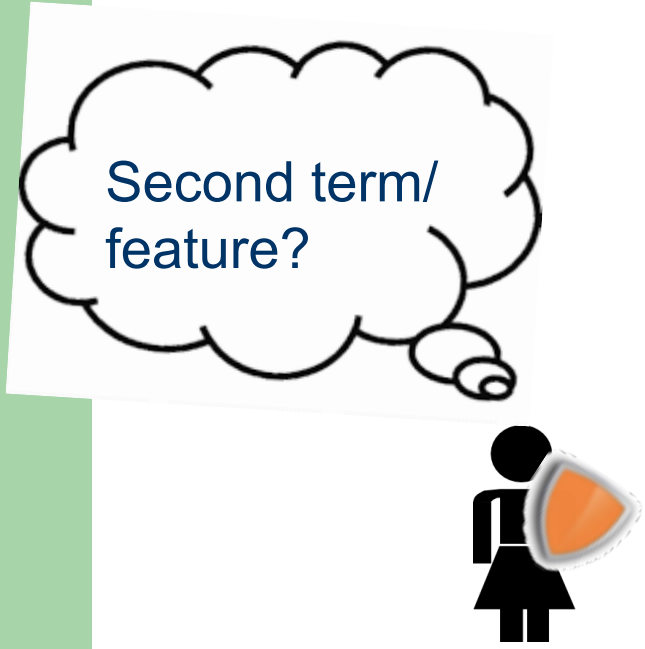
$R(i) = i$

$P_N \sim \text{Binomial}(100, 0.2)$

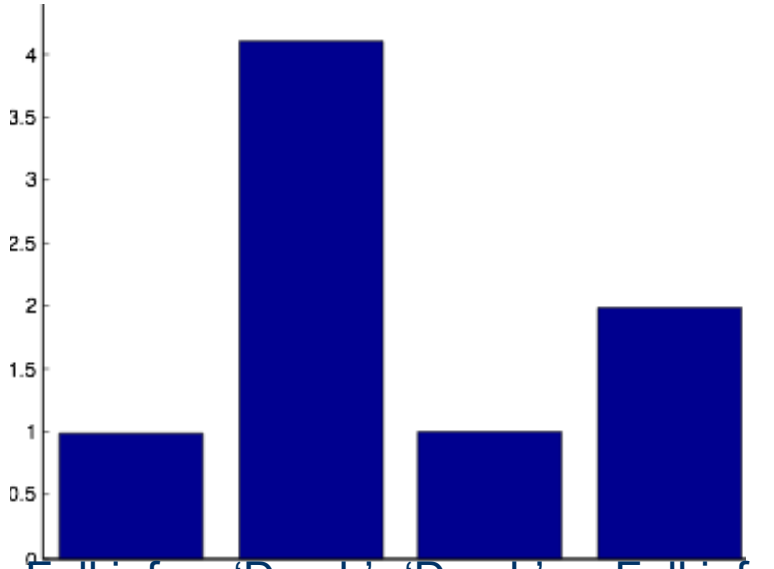
$p = 0.2,\ c_{fa}=140,\ c_d=120$



To invest or not to invest? (It depends.)



Defender's NE payoff

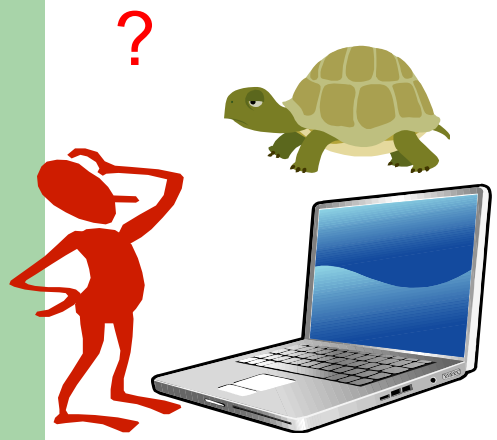


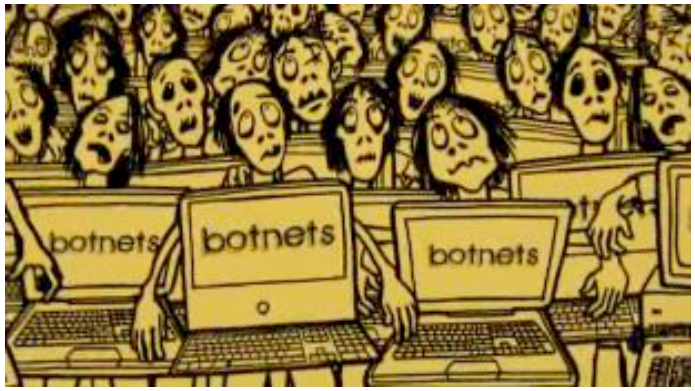
Overview

- **Part I: Adversarial classification [Dritsoula, Loiseau, Musacchio]**
  - Single-feature threshold-based classification [CDC 2012]
  - Generalized payoff structure [GameSec 2012]
  - Optimal defense and attack strategies [to be submitted to TIFS 2015]

- **Part 2: Botnet detection games [Soper, Musacchio]**
  - Homogeneous Agents [Allerton 14]
  - Heterogeneous Agents [NetGCooP 14]

Do I have a botnet?





Threshold ?

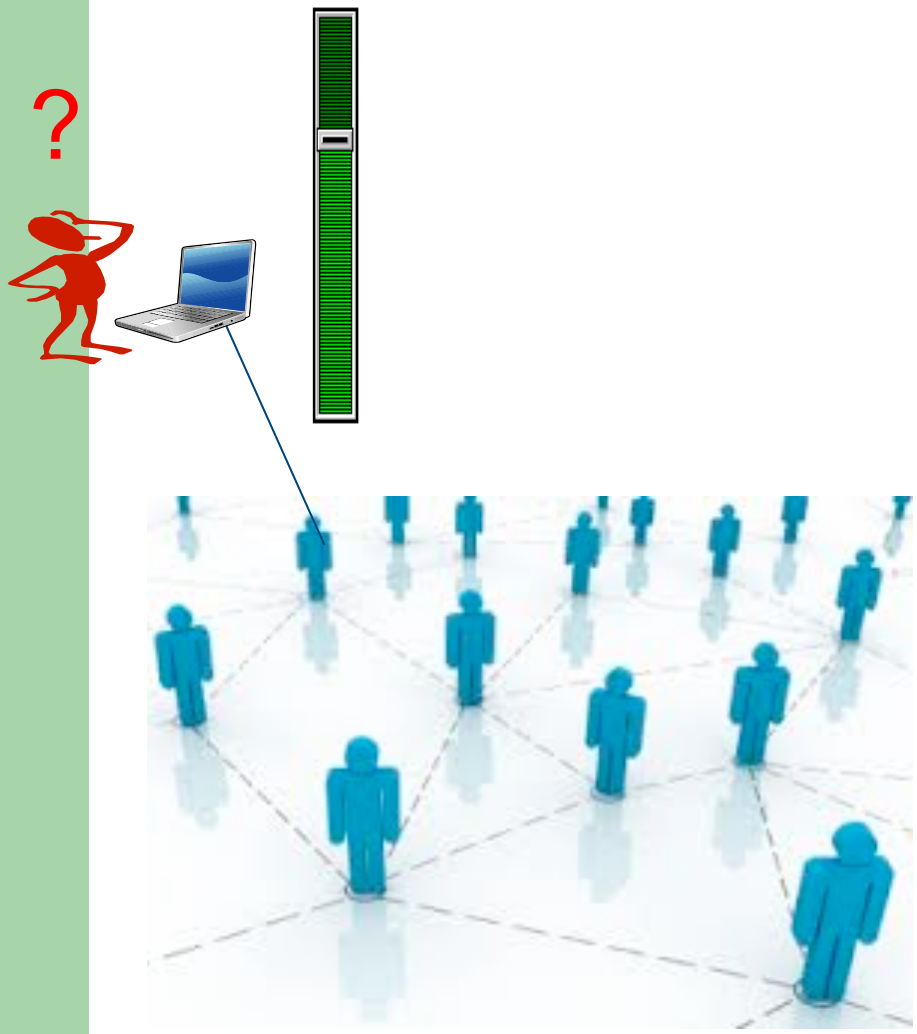
"Slowness"

How aggressively should I use bot?

Aggressive

subtle

Some questions



- How do network externalities influence a user's threshold choice?
- Would a social planner tell people to be more sensitive or less?

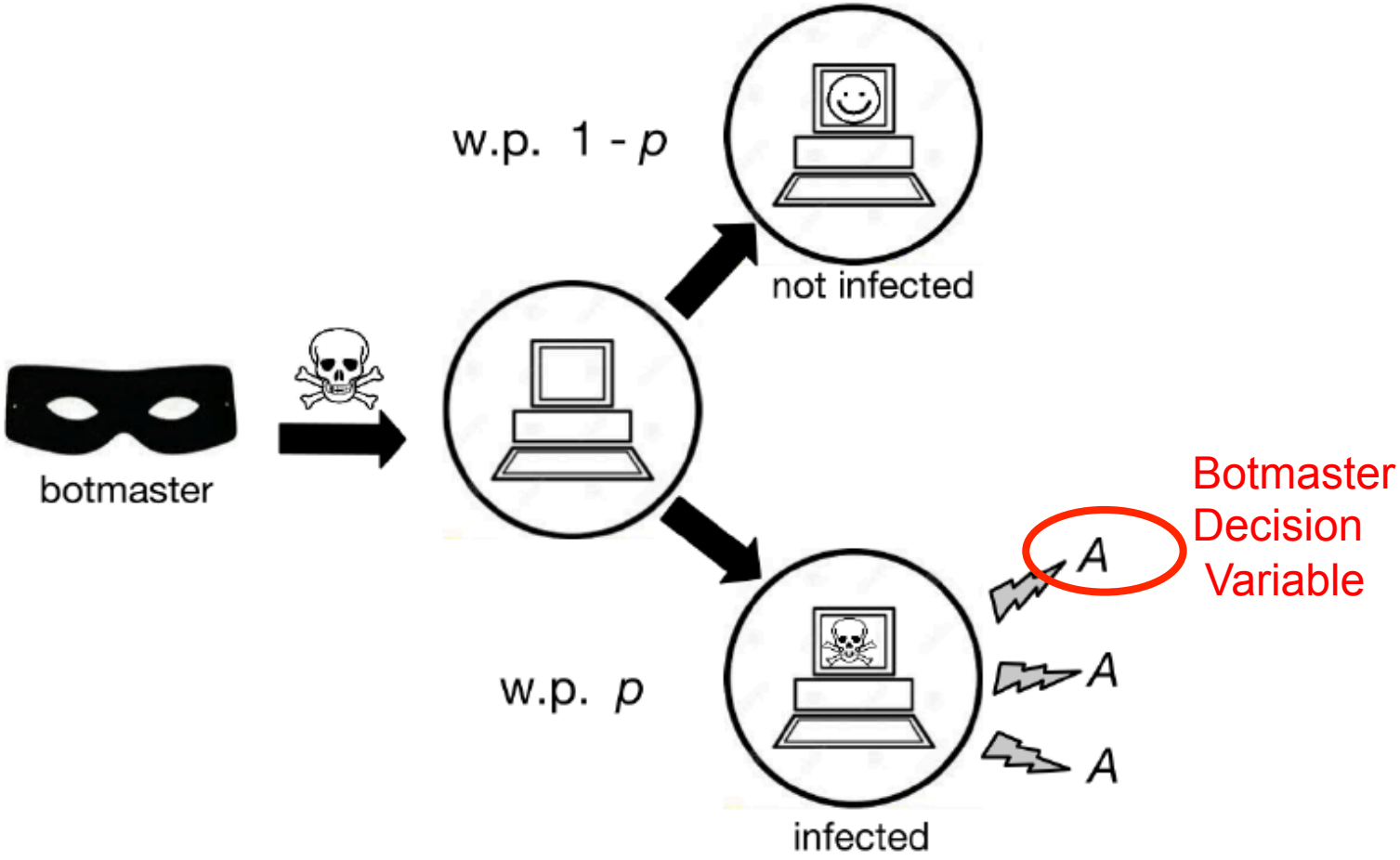
Prior Work

- Mathematical models of botnets
  - Dragon et. al. (06)
  - Song et. al. (08)
  - Vratonjic et. al. (10)
  - Li and Strengel (09)
  - Bensoussan (10)
  - Lelarge and Bolot (08)
    - Capture network effects in large graph
    - Interdependent security investment decisions
- Security vs. Reliability
  - Difficult to distinguish failures from breaches
  - Honeyman and Schwartz (07)

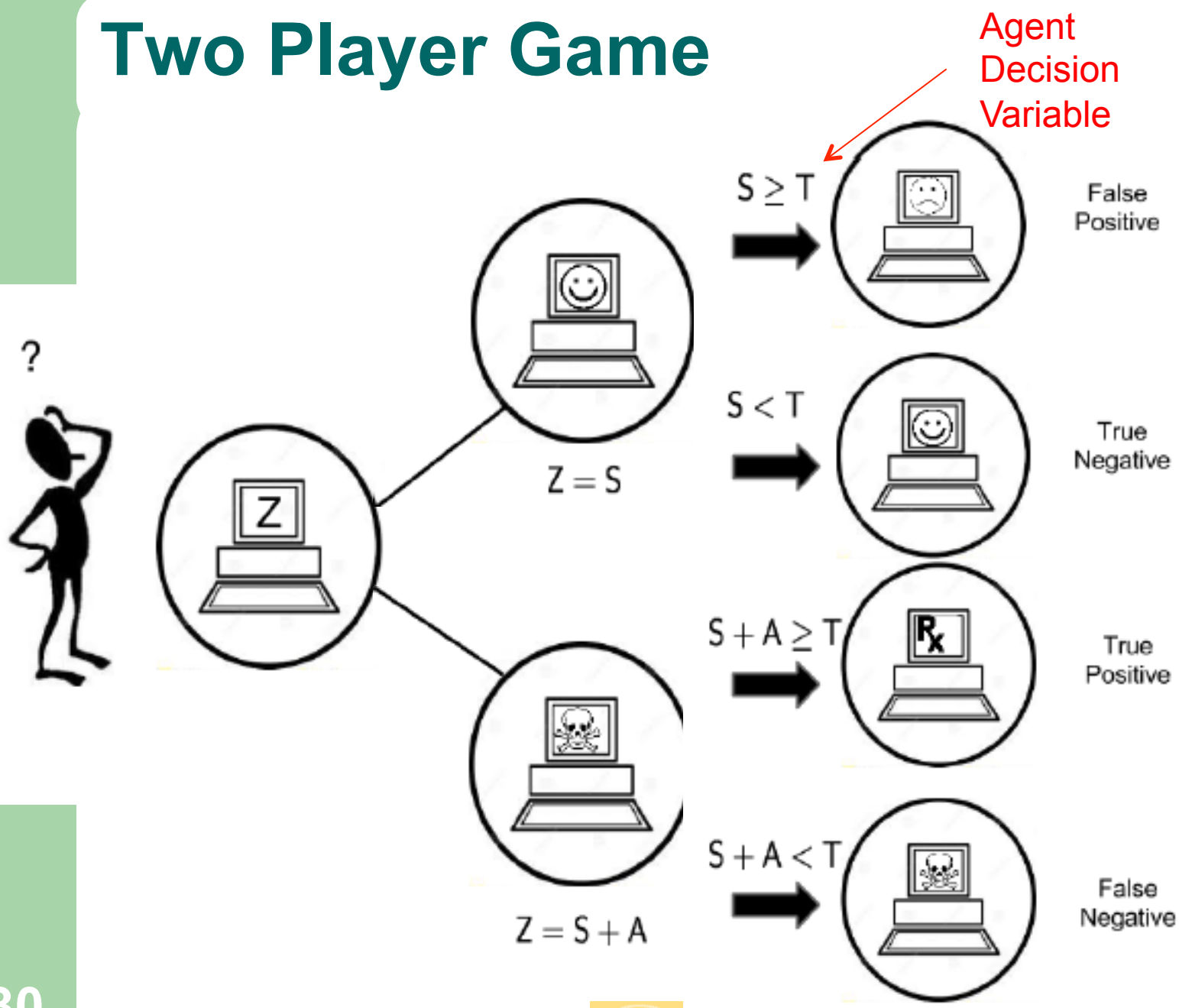
Agenda to Follow

- Two player version of game 
- Large multi-agent game
- Numerical results

Two Player Game



Two Player Game



Two Player Game

- Parameters:

  c: cost of a false alarm

  ℓ: loss due to a missed detection

  $F_S(\cdot)$: c.d.f of S

- Cost to agent

$$C(A,T) = c\mathbb{P}(\text{false alarm}) + \ell\mathbb{P}(\text{missed detection})$$

? $$= c[1 - F_S(T)](1-p) + \ell F_S(T-A)p$$

Two Player Game

- Botmaster utility proportional to aggressiveness

$$U(A,T) = A\mathbb{P}(\text{missed detection})$$
$$= AF_S(T-A)p$$

Two Player Game

- Result:
  - For $S \sim \exp()$ there exists a unique pure Nash equilibrium.

If $\frac{c}{\ell}\frac{1-p}{p} \leq 1$

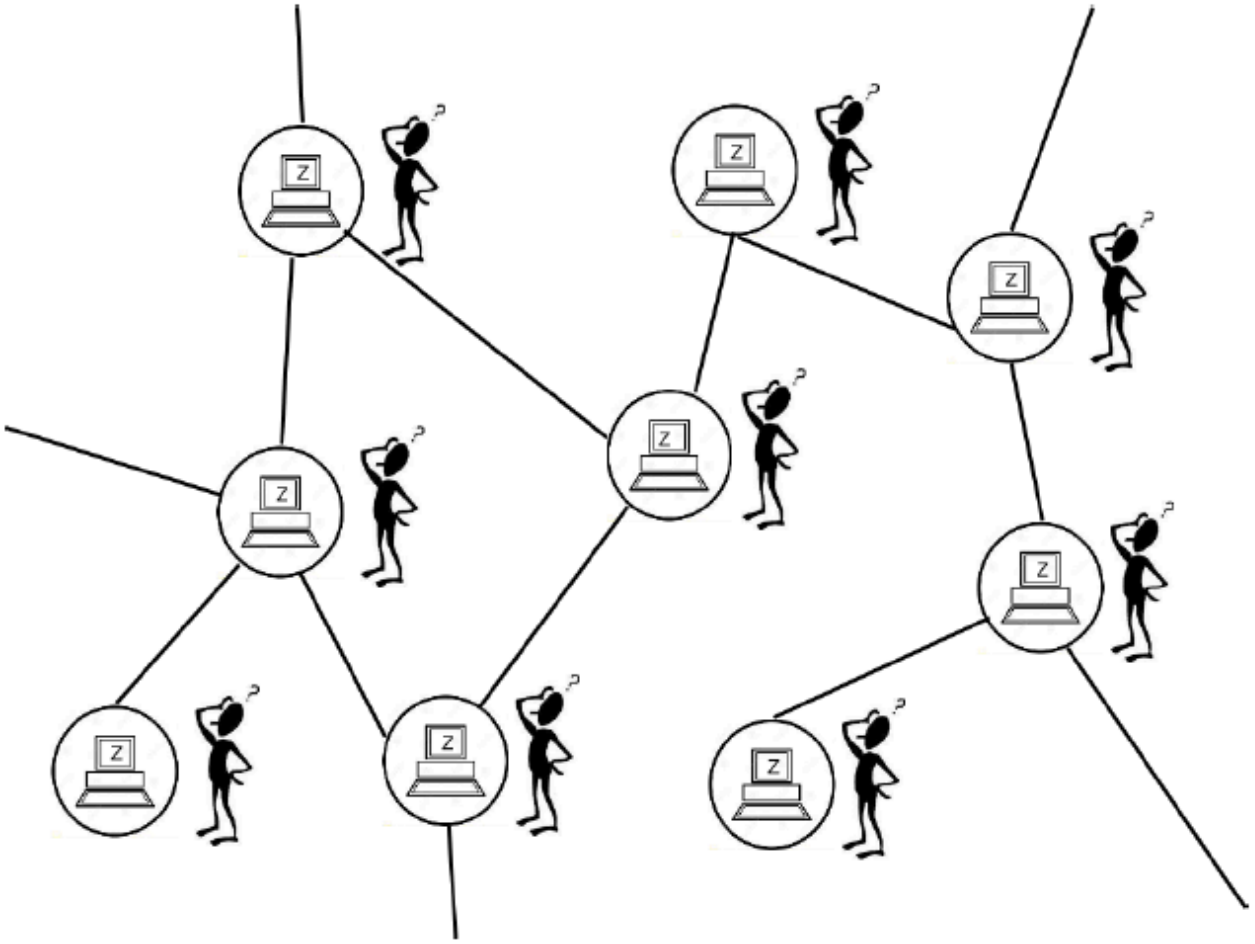
$$A^* = T^* = 0.$$

If $\frac{c}{\ell}\frac{1-p}{p} > 1$

$$A^* = \frac{1}{\beta}\log\left(\frac{c}{\ell}\frac{1-p}{p}\right),$$

$$T^* = \frac{1}{\beta}\log\left(\frac{c}{\ell}\frac{1-p}{p}\left[1+\log\left(\frac{c}{\ell}\frac{1-p}{p}\right)\right]\right)$$

Multi-Agent Game



Multi-Agent Game

- Capture network effect using technique of Lelarge and Bolot (08)
  - Objective method Aldous and Steele (04)
  - Show a sequence of graphs converges in distribution to a limiting graph
  - In particular

$$G(n, \lambda/n) \xrightarrow{d} T(\lambda)$$

Erdos-Renyi

Galton-Watson branching w/ Poisson(λ) offspring

Multi-Agent Game

- Focus on $T(\lambda)$ graphs
- Notation
  - A Botmaster Aggressiveness
  - T_i Threshold agent i
  - S_i Noise on agent i's observation (iid)
  - W_i indicator on infection reaching agent i from children
  - Y_i indicator on missed detection event
    - (infection reaches i and is not detected)
- If an agent detects infection, she it cannot pass it on

Recursive Tree Process

$$W_i = 1 - (1 - \chi_i) \prod_{k \to i} (1 - B_{ki} Y_k)$$

$$Z_i = A W_i + S_i$$

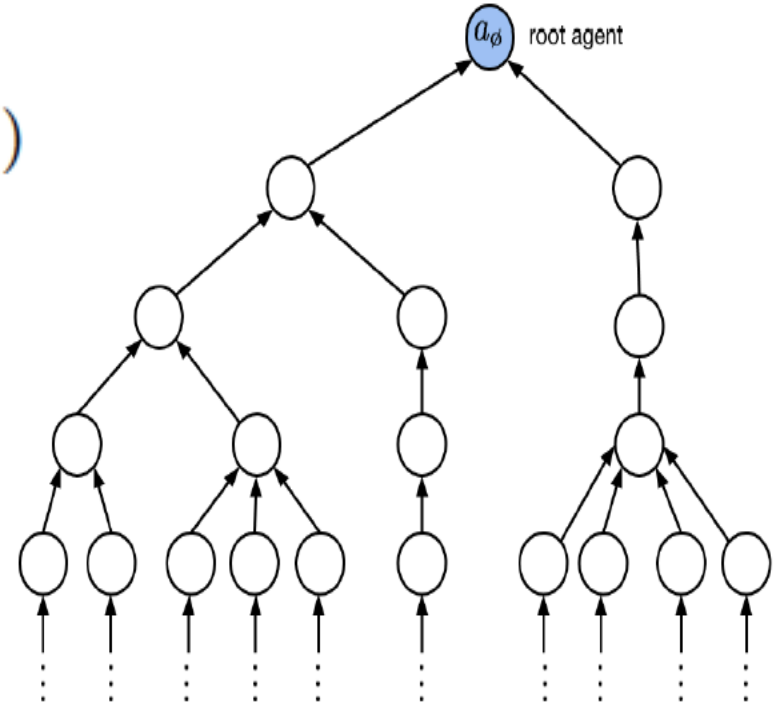
$$Y_i = W_i \mathbb{1}_{\{T_i > Z_i\}}$$

$$\chi_i \overset{iid}{\sim} \text{Bernoulli}(p)$$

Direct infection indicator

$$B_{ki} \overset{iid}{\sim} \text{Bernoulli}(q)$$

Infection can spread indicator



For Comparison, RTP from Lelarge and Bolot (08):

$$X_i = 1 - (1 - \chi_i) \prod_{i \sim j} (1 - B_{ij} X_j)$$

Recursive Distributional Equation (RDE)

Approach

- Consider Symmetric Strategies
- Look for invariant solution of RTP

i.i.d. Bernoulli

$$W \stackrel{d}{=} 1 - (1-\chi)\prod_{k=1}^{N}(1 - B_k Y_k)$$

Poisson(λ)

$$Y \stackrel{d}{=} W \mathbb{1}_{\{T > AW+S\}}$$

Bernoulli

Can show solution is:

$$\mathbb{P}(Y=1) = h \stackrel{\Delta}{=} F_S(T-A)[1-(1-p)e^{-\lambda q h}]$$

$$\mathbb{P}(W=1) = 1-(1-p)e^{-\lambda q h}$$

See Aldous and Bandyopadhyay (05) for RDE background

Multi-Agent Game

- Botmaster utility

$$U(A,T) = AE[Y] = Ah(A,T)$$

- Proposition:
  - If *S* has a gamma(α , β) distribution, botmaster has unique best response
- Key step of proof
  - *A* = *G(A,T)* fixed point unique

$$G(A,T) = \frac{F_S(T-A)}{f_S(T-A)} \left[1 - F_S(T-A)(1-p)\lambda q e^{-\lambda q h} \right]$$

Agent vs. Agent

- Suppose all but one agent play T
  - "root" agent plays $T_\emptyset$

$$C_\emptyset(A, T, T_\emptyset) = c\mathbb{P}(\text{false alarm}) + \ell\mathbb{P}(\text{missed detection})$$
$$= c[1 - F_S(T_\emptyset)](1-p)e^{-\lambda q h} + \ell F_S(T_\emptyset - A)[1 - (1-p)e^{-\lambda q h}]$$

- Key idea: deviation does not effect chance infection reaches root. Use this to show

$$\frac{\partial C_\emptyset}{\partial T_\emptyset} = 0 \implies \frac{f_S(T_\emptyset - A)}{f_S(T_\emptyset)} = \frac{c}{\ell}\frac{(1-p)e^{-\lambda q h}}{[1-(1-p)e^{-\lambda q h}]}$$

Agent vs. Agent

$$\frac{\partial C}{\partial T_{\emptyset}} \lesseqgtr 0 \iff \frac{f_S(T_{\emptyset} - A)}{f_S(T_{\emptyset})} \lesseqgtr \frac{c}{\ell} \frac{(1-p)e^{-\lambda q h}}{1-(1-p)e^{-\lambda q h}}$$

Depends only on A and $T_{\emptyset}$

Depends only on A and T
Nonincreasing in T

- If *S*~gamma(α, β) distribution ($\alpha \geq 1$, $\beta > 0$)
  - Then $\frac{f_S(T_{\emptyset}-A)}{f_S(T_{\emptyset})}$ increasing or constant in $T_{\emptyset}$
  - Also $C_{\emptyset}(A, T, T_{\emptyset})$ quasiconvex and there exists a fixed point to

$$T^* \in \sigma_{\emptyset}(A, T^*)$$

Best response correspondence

Nash equilibrium

- So far:
  - For a given A we know there is a T*
    - (under conditions on *S*)
  - For all agents playing a given T, there is an A*
    - (under conditions on *S*)

- Using continuity and asymptotic arguments, can show:
  - If $S \sim \text{Gamma}(\alpha, \beta)$ $\alpha \geq 1$, $\beta > 0$, there exists a pure Nash equilibrium (A^*, T^*) of the multi agent botnet game

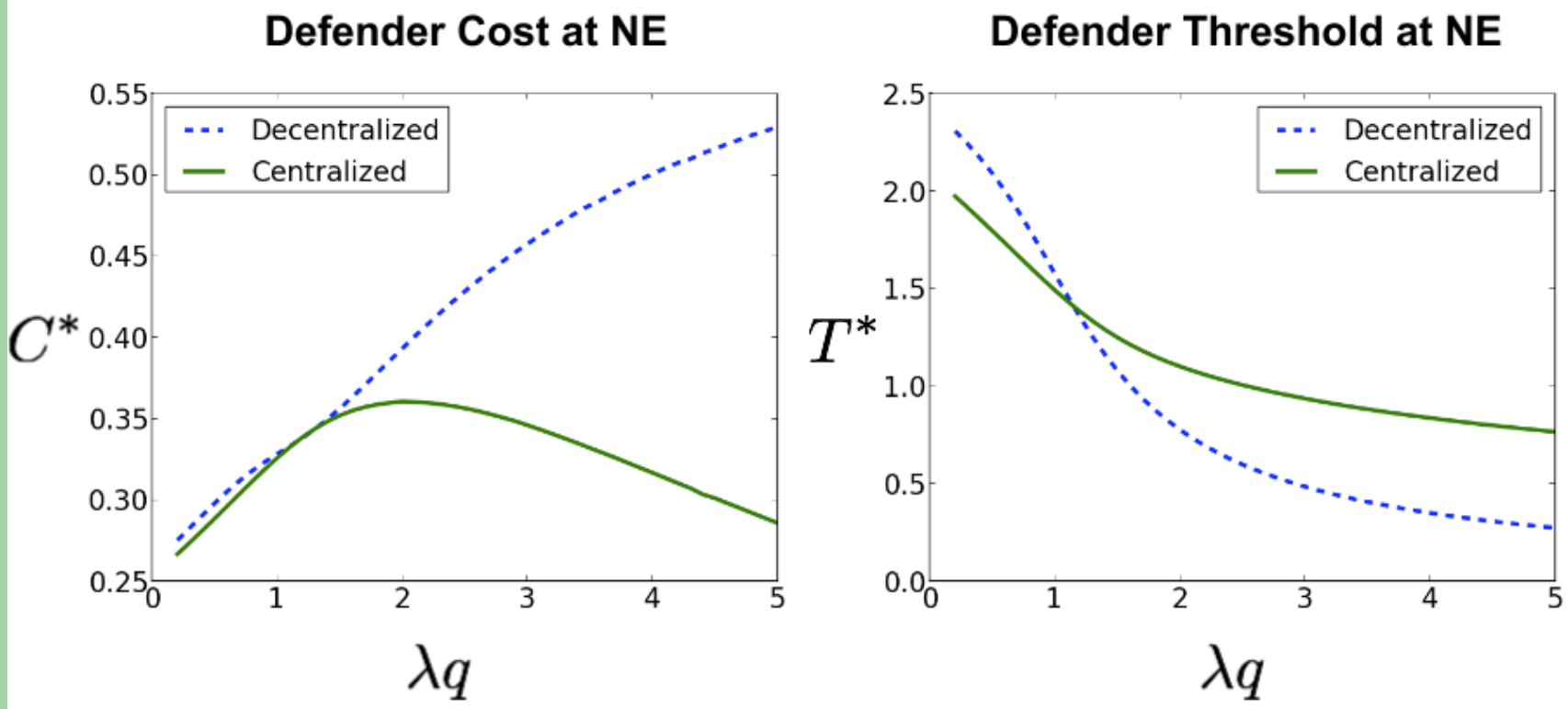
Central Planner

- Suppose a central planner chooses the T everyone uses

$$\Longrightarrow \frac{\partial h}{\partial T} \geq 0$$

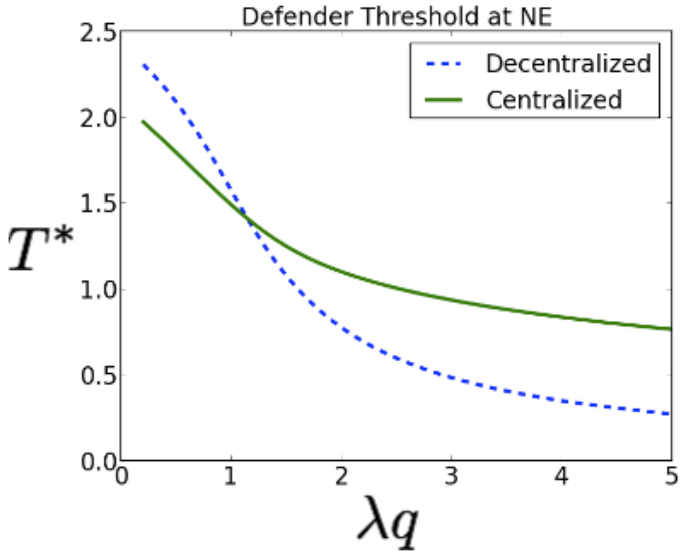
- Yet, still can show
  - If $S \sim \text{Gamma}(\alpha, \beta)$, $\alpha \geq 1$, $\beta > 0$, there exists a pure Nash equilibrium (A^*, T^*) of the centrally planned, multi-node botnet game

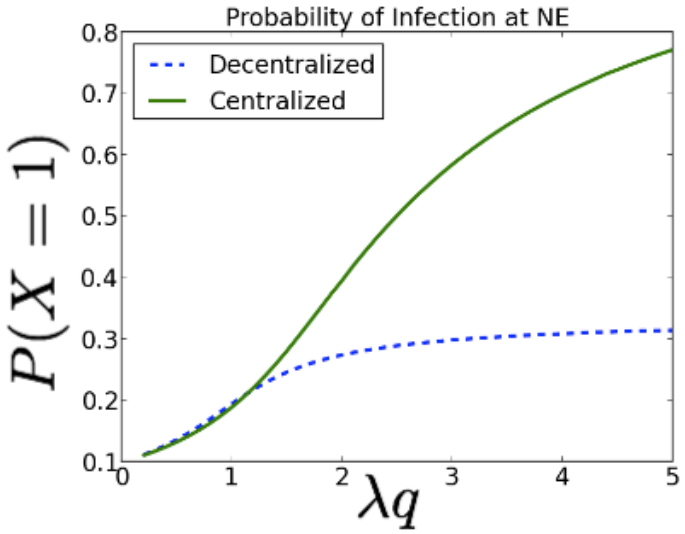
Numerical Results

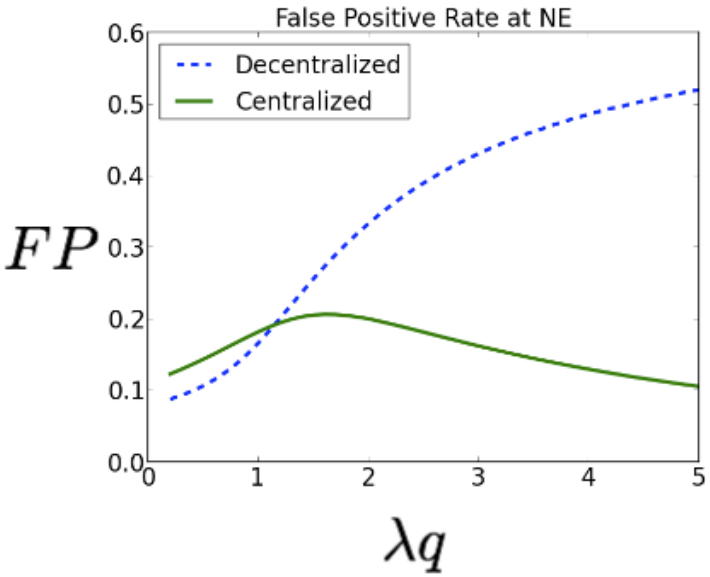


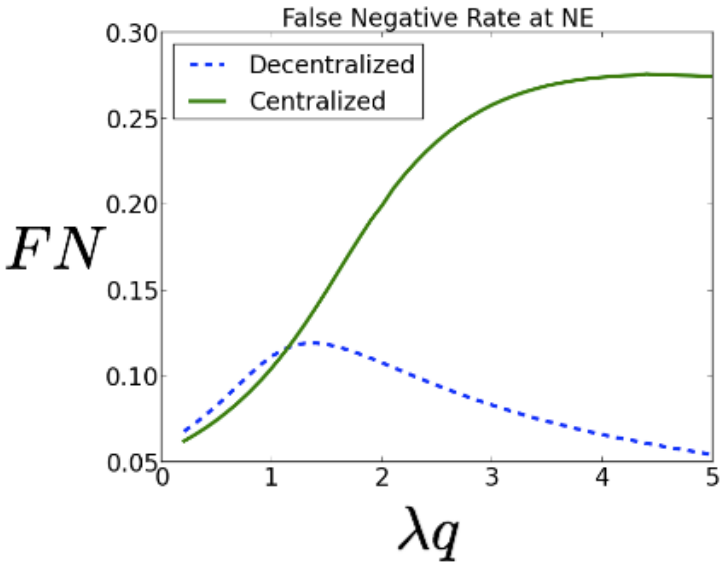
$$S \sim \exp(1), p = 0.1, c = 1, \ell = 2A$$

More results









Current and Future Directions

- Dynamics
  - Sequential hypothesis testing
- Heterogeneous Agents
- Non Symmetric equilibrium