# Interdependent Defense Games: Modeling Interdependent Security under Deliberate Attacks (Extended Abstract)

Michael Ceyko *

Department of Computer Science
Harvard University
Cambridge, MA 02138

Hau Chan & Luis E. Ortiz †

Department of Computer Science
Stony Brook University
Stony Brook, NY 11778

**Abstract**

Inspired by problems in cyber defense, we propose *interdependent defense (IDD) games*, a computational game-theoretic framework to study aspects of interdependence of risk and security under deliberated external attacks. Our model adapts *interdependent security (IDS) games*, a model due to Heal and Kunreuther, to explicitly model the source of the risk: the attackers' behavior. We provide a complete characterization of the set of Nash equilibria (NE) of an important subclass of IDD games. Some interesting properties of the (almost surely unique) NE immediately fall off the characterization, as well as the design of a simple polynomial-time algorithm for computing NE in that subclass. We propose a generator of random instances of IDD games based on the real-world Internet-derived AS graph ($\sim 27K$ nodes and $\sim 100K$ edges as measured in March 2010 by the DIMES project). Preliminary experiments applying simple learning-in-games

*This article includes work performed by the first author as part of his Undergraduate Honors' Project in Computer Science at Stony Brook under the guidance of the third author.

†Contact author's email: `leortiz@cs.sunysb.edu`; webpage: `http://www.cs.sunysb.edu/~leortiz`

heuristics to compute (approximate) NE in such randomly-generated game instances are promising. Finally, we discuss several extensions, current and future work, and present several open problems.

# 1 Introduction

*Interdependent security (IDS) games*, originally introduced by economists Kunreuther and Heal [Kunreuther and Heal, 2003], model general abstract security problems in which an individual within a population considers whether to voluntarily invest in some protection mechanisms or security against a risk they may face, knowing that the cost-effectiveness of the decision depends on the investment decisions of others in the population because of transfer risks (i.e., the "bad event" may be transferable from a compromised individual to another). In IDS games, each player's pure strategies are whether to invest or not. The cost (i.e., negative payoff) functions encode the fact that the actual level of safety afforded by the investment is a function not only of the individual player's choice but also on the choices of other players in the population. Kunreuther and Heal [2003] devised the model for these games and studied some of their properties analytically, while Kearns and Ortiz [2003] later proposed algorithms for computing equilibria.

As a canonical example of the real-world relevance of IDS settings and the applicability of IDS games, Kunreuther and Heal used this model to describe problems such as airline baggage security [Heal and Kunreuther, 2005]. Individual airlines may choose to invest in additional complementary equipment to screen passengers' bags and check for hazards such as bombs that could cause damage to their passengers, planes, buildings, or even reputations. Regardless of the outcome, airlines wish to avoid the bad event. However, mainly due to the large amount of traffic volume, it is impractical for an airline to go beyond applying security checks to bags incoming from passengers and include checks to baggage or cargo transferred from other airlines. So, even if an airline invests in security, they can still experience a bad event if the bag was transferred from an airline that does not screen incoming bags, rendering their investment useless. Even if full screening were performed, the Chrismas day episode in Detroit last year [O'Connor and Schmitt, 2009] is a painful reminder that transfer risk still exists. Thus, we can see how the cost-effectiveness of an investment can be highly dependent on others' investment decisions.

In cyberspace, the situation is similar but slightly different in nature. Consider a network where all computers fully trust all other computers and freely exchange information. Each user has complete control over his own computer and can decide if he wants to protect his computer from hackers by installing a firewall, for example. However, he cannot control if others on the network protect themselves as well. So, in order for one to feel secure in storing his information on the network, he not only has to think about his own security, but also the security of *other* computers on the internal network, because any other computer may access his as well. If any computer were hacked, his information would potentially be exposed to the outside world.

Two potential outcomes immediately arise out of the cyber security scenario. If one doesn't think enough people have invested in security, then one will not invest either, because any investment will contribute negligibly to the overall protection of one's data. If none of the other people invest, one would not want to invest. Also, and this is the aspect that perhaps differentiates the most the cyberspace from the airline security scenario previously discussed, if nearly everyone has invested in security, one may no longer feel the need to protect oneself because the network is already mostly secure and the amount of work required to protect oneself outweighs the minimal change in overall security. Thus, as many invest, fewer may want to invest.

In general, scenarios like the ones just described have two important components: a *bad event* (e.g., virus, worm, hacking) which all players attempt to avoid as it will cause losses (e.g., monetary, privacy, data) for the player, and an *investment* decision which reduces the risk of the bad event but has an associated cost (e.g., monetary, time, work) to the player.

But there is a *third* important component implicit in the previous discussion: the source of the risk! In both the airline and cyber security scenario the source of the risk results from the potential *deliberate attack* or action taken by agents often "external" to the system (e.g., the terrorist and/or the hacker).

Inspired in part by problems in the cyber domain, in this work, we adapt IDS games to cases in which the abstract "bad event" results from the *deliberate* action of one or several "external agents," whom we refer to (somewhat interchangeably) as the "attackers," "aggressors" or "bad actors." The "internal agents" (e.g., airlines and computer network users), whom we also often refer to as "sites," have the voluntary choice to individually invest in security to defend themselves against a direct or indirect offensive attack, modulo, of course, the cost-effectiveness to do so. As a result of the adapta-

tion, we formally define a new model we call *interdependent defense (IDD) game*. We study this new class of games and provide preliminary results on characterizations of Nash equilibria, as well as representational and computational properties. We also propose a process to generate random instances of large interdependent defense games based on the actual network structure of real ISPs in the current Internet as a benchmark to evaluate the quality and study the behavior of algorithms for computing equilibria in these games. We report preliminary results evaluating a simple learning-in-games heuristic [Fudenberg and Levine, 1999] as a way to obtain an approximate Nash equilibrium of large interdependent game generated according to the proposed process. We conclude with a brief discussion of extensions, future work, open problems and a summary of our contributions.

# 2 Interdependent Security Games

We start by looking at *IDS games* and define the parameters and rules governing this model. Each player $i$ has a choice of whether or not to invest, which we denote by $a_i$ such that $a_i \in \{0, 1\}$ where "1" corresponds to investing and "0" to not investing. For player $i$, an investment will cost $C_i$ and the bad event will induce a loss $L_i$. Naturally, the case where $C_i > L_i$ is not interesting because the player *always* reduces his costs more by not investing. Hence, IDS models are mostly interesting in cases when $L_i \gg C_i$ so that the player can potentially greatly reduce his overall cost by investing. More important is the ratio of the two parameters, the player's "cost-to-loss" ratio, which we define as $\rho_i \equiv C_i/L_i$.

Bad events can occur through both *direct* and *indirect* means. *Direct risk*, or *internal risk,* is the chance of a player experiencing a bad event due to direct contamination, e.g., if you didn't have anti-virus software and downloaded a virus from the web to your own computer. The IDS model assumes that investing will completely protect the player from direct contamination; hence, internal risk is only possible when $a_i = 0$. We denote by $p_i$ the probability that player $i$ will experience a bad event because of internal risk. *Indirect risk* results from the possibility that a player experiences a bad event because of a *transfer* from other player; from another computer on the network, for example. The IDS model also assumes that the interactions between players are unaffected by investment, so regardless of one's investment, one's *transferred risk* is the same. We denote by $q_{ji}$ the probability

that player $j$ is directly contaminated/infected/targeted, does not experience the bad event but transfers it to player $i$ who ends up experiencing the bad event. Note that there is an implicit global constraint on these parameters, by the axioms of probability: $p_i + \sum_{j=1}^{n} q_{ij} \leq 1$ for all $i$.

We now formally define a (directed) graphical-games [Kearns et al., 2001, Kearns, 2007] version of IDS games, as first introduced by Kearns and Ortiz [2003]. Denote by $[n] \equiv \{1, \ldots, n\}$ the set of $n$ players. Note that the parameters $q_{ij}$'s induce a directed graph $G = ([n], E)$ such that, for all $i, j \in [n]$, directed edge $(i, j) \in E$ if and only if $q_{ij} > 0$; more formally, $E \equiv \{(i, j) \mid q_{ij} > 0\}$. Let $\mathrm{Pa}(i) \equiv \{j \mid q_{ji} > 0\}$ be the set of players that are *parents* of player $i$ in $G$ (i.e., the set of players that player $i$ is exposed to via transfers), and by $\mathrm{PF}(i) \equiv \mathrm{Pa}(i) \cup \{i\}$ the *parent family* of player $i$, which includes $i$. Denote by $k_i \equiv |\mathrm{PF}(i)|$ the size of the parent family of player $i$. Similarly, let $\mathrm{Ch}(i) \equiv \{j \mid q_{ij} > 0\}$ be the set of players that are *children* of player $i$ (i.e., the set of players to whom player $i$ can present a risk via transfer) and $\mathrm{CF}(i) \equiv \mathrm{Ch}(i) \cup \{i\}$ the *(children) family* of player $i$, which includes $i$. .

The *probability that player $i$ is safe from player $j$*, as a function of player $j$'s decision, is

$$e_{ij}(a_j) \equiv a_j + (1 - a_j)(1 - q_{ji}) = (1 - q_{ji})^{1-a_j} \tag{1}$$

because if $j$ invests, then it is impossible for $j$ to transfer the bad event, while if $j$ does not invest, then $j$ either experiences the bad event or transfers it to another player, [1] but never both.

Denote by $\mathbf{a} \equiv (a_1, \ldots, a_n) \in \{0, 1\}^n$ the *joint action* of all $n$ players. Also denote by $\mathbf{a}_{-i} \equiv (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n)$ the joint-action of all players except $i$. For any subset $I \subset [n]$ of players, [2] denote by $\mathbf{a}_I \equiv (a_i : i \in I)$ the sub-component of the joint action corresponding to those players in $I$ only. We define $i$'s *overall safety* from *all* other players as [3]

$$s_i \equiv s_i(\mathbf{a}_{\mathrm{Pa}(i)}) \equiv \prod_{j \in \mathrm{Pa}(i)} e_{ij}(a_j) = \prod_{j \in \mathrm{Pa}(i)} (1 - q_{ji})^{1-a_j} \tag{2}$$

[1] The player receiving the transfer still has the chance of not experiencing the bad event. However, without some form of screening of transfers, this chance is usually very low.

[2] Here we use as convention that $\subset$ need not mean "a proper subset" (i.e., $A \subset B$, includes $A = B$).

[3] Throughout the document, when clear form context, we often drop the arguments of functions to reduce notational clutter.

and equivalently his *overall risk* from *some* other players is

$$r_i \equiv r_i(\mathbf{a}_{\mathrm{Pa}(i)}) \equiv 1 - s_i(\mathbf{a}_{\mathrm{Pa}(i)}) = 1 - s_i. \tag{3}$$

Note that each players' external safety (and risk) is a direct function of its parents only, *not* all other players.

From these definitions, we obtain player $i$'s overall cost, the cost of joint action $\mathbf{a} \in \{0,1\}^n$, corresponding to the (binary) investment decision of all players, is

$$\begin{aligned} M_i &\equiv M_i(\mathbf{a}_{\mathrm{PF}(i)}) \equiv M_i(a_i, \mathbf{a}_{\mathrm{Pa}(i)}) \\ &\equiv \begin{cases} C_i + r_i L_i, & \text{if } a_i = 1 \text{ (player } i \text{ invests)}, \\ (p_i + (1-p_i) r_i) L_i, & \text{if } a_i = 0 \text{ (player } i \text{ does not invest)}. \end{cases} \\ &= a_i(C_i + r_i L_i) + (1 - a_i)(p_i + (1-p_i) r_i) L_i \ . \end{aligned} \tag{4}$$

Whether players invest is dependent solely on what they can gain or lose by investing. If the overall cost of investing is less than the overall cost of not investing, the player will invest. Applying this logic to cost function $M_i$, player $i$ will invest if

$$C_i + r_i L_i < [p_i + (1-p_i) r_i] L_i$$

so that the investment cost and the losses due to a transferred event do not outweigh the losses from an internal or transferred bad event. Similarly, if the inequality in the last expression is reversed or is replaced by equality, player $i$ will not invest or would be indifferent, respectively. Rearranging the expression for the best-response conditions given in the last equation and letting $\Delta_i \equiv \rho_i / p_i = \frac{C_i}{p_i L_i}$, the *cost-to-expected-loss ratio* of player $i$, we get the following *best-response correspondence* $\mathcal{BR}_i : \{0,1\}^{k_i - 1} \to 2^{\{0,1\}}$ for player $i$:

$$\mathcal{BR}_i(\mathbf{a}_{\mathrm{Pa}(i)}) \equiv \begin{cases} \{1\}, & \text{if } s_i > \Delta_i, \\ \{0\}, & \text{if } s_i < \Delta_i, \\ \{0,1\}, & \text{if } s_i = \Delta_i. \end{cases} \tag{5}$$

In other words, whether it is cost-effective for player $i$ to invest or not depends on a simple threshold condition on his safety: Does he feel safe enough from others?

**Definition 1.** *A joint-action* $\mathbf{a}^* \in \{0,1\}^n$ *is a* pure-strategy Nash equilibrium (PSNE) of an IDS game *if* $a_i^* \in \mathcal{BR}_i(\mathbf{a}^*_{Pa(i)})$ *for all players* $i$ *(i.e.,* $\mathbf{a}^*$ *is a mutual best-response).*

# 3 Interdependent Defense Games

In this section, we present our adaptation of IDS to the setting of deliberate attacks. From now on we will call this setting *interdependent defense* and use the acronym *IDD* to emphasize its connection to IDS. eventually leading to our new proposed model: *interdependent defense (IDD) games*. The formulas up to this point have mostly already been defined in previous work. We begin by introducing some new concepts that are particularly relevant to settings in which the source of the risk of experiencing the bad event are deliberate attacks, such as in the cyber domain.

## 3.1 Protection Investment May Reduce Transfer Risk

The first modification we introduce to the traditional IDS games is to allow the possibility that investing in protection not only makes us safe from direct attack but may also partially reduce (or even eliminate) the transfer risk. The cyber domain provides a motivation for the modification. [4] The cost of checking all communications or transfers within a computer network can be significantly high, but at least some data can be checked in a reasonable amount of time. We can imagine that if a player invests, *some* checking can be done for *transferred events*. We incorporate this factor by introducing a new real-valued parameter $\alpha_i \in [0, 1]$ representing the probability that a transfer of a potentially bad event will go *unblocked* by $i$'s security, assuming $i$ has invested. Thus, we redefine player $i$*'s* overall cost as [5]

$$M_i \equiv M_i(\mathbf{a}_{\mathrm{PF}(i)}) \equiv a_i[C_i + \alpha_i r_i L_i] + (1 - a_i)[p_i L_i + (1 - p_i) r_i L_i] \quad (6)$$

A similar extension was also proposed by Heal and Kunreuther [2007].

## 3.2 Introducing the Aggressor

We also change how bad events are modeled. We introduce an additional player, the *aggressor* or *attacker*, who *deliberately* initiates bad events. (So that now bad events are no longer "chance occurances" without any deliberation.) Instead of having an investment decision like other players, the

[4] Problems related to vaccination against a contagious disease are other source of motivation for the proposed modification.

[5] A possible generalization, which we do not pursue here, may also consider $\alpha_i$ a function of $\mathrm{Pa}(i)$.

aggressor has a *target decision* for each player - a choice whether or not to attack that player. We introduce a new action variable for each player, $b_i \in \{0, 1\}$, where $b_i = 1$ represents that a bad event is attempted to be initiated on player $i$. Hence, the aggressor's pure strategy is denoted by the vector $\mathbf{b} \in \{0, 1\}^n$.

Changing from "random" non-strategic attacks whose probability of occurrence is determined independent of the actions of the internal players, to intentional attacks, ones that are deliberately carried out by an external actor, gives reason for us to alter $p_i$ and $q_{ij}$ because their original definitions actually imply extra meaning with respect to the new aggressor.

### 3.2.1 Risk Parameters Depend on Aggressor's Actions

The game parameter $p_i$ is the probability that player $i$ will experience a bad event due to internal risk, which implicitly "encodes" $b_i$ because $b_i = 0$ implies $p_i = 0$. Thus, we redefine

$$p_i \equiv p_i(b_i) \equiv b_i \widehat{p}_i$$

so that player $i$ has *intrinsic risk* $\widehat{p}_i$, and only has *internal risk* if targeted (i.e, $b_i = 1$). The new parameter $\widehat{p}_i$ represents the (*conditional*) probability that an attack is successful at player $i$ *given* that player $i$ was directly targeted and did not invest in protection.

A similar situation arises for $q_{ij}$, the probability that player $i$ will transfer a *bad event* to player $j$ (implying that $q_{ij}$ "encodes" $b_i = 1$, because a prerequisite is that $i$ is targeted before it can transfer the bad event to $j$). We redefine

$$q_{ij} \equiv q_{ij}(b_i) \equiv b_i \widehat{q}_{ij}$$

so that $\widehat{q}_{ij}$ is the intrinsic transfer probability from player $i$ to player $j$, independent of $b_i$. The new parameter $\widehat{q}_{ij}$ represents the (*conditional*) probability that an attack is successful at player $j$ *given* that it originated at player $i$, did not occur at $i$ but was transferred undetected to $j$.

Note that just as it was the case with traditional IDS games, there is an implicit constraint on the risk-related parameters: $\widehat{p}_i + \sum_{j \in \mathrm{Ch}(i)} \widehat{q}_{ij} \leq 1$, for all $i$.

### 3.2.2 The Internal Players' Costs

Because the $p_i$'s and $q_{ij}$'s depend on the aggressor's action $\mathbf{b}$, so does the safety and risk functions. In particular, we now have

$$e_{ij}(a_j, b_j) \equiv a_j + (1 - a_j)(1 - b_j \widehat{q}_{ji}) = (1 - \widehat{q}_{ji})^{b_j(1-a_j)},$$

$$s_i \equiv s_i(\mathbf{a}_{\mathrm{Pa}(i)}, \mathbf{b}_{\mathrm{Pa}(i)}) \equiv \prod_{j \in \mathrm{Pa}(i)} e_{ij}(a_j, b_j) = \prod_{j \in \mathrm{Pa}(i)} (1 - \widehat{q}_{ji})^{b_j(1-a_j)}$$

and

$$r_i \equiv r_i(\mathbf{a}_{\mathrm{Pa}(i)}, \mathbf{b}_{\mathrm{Pa}(i)}) \equiv 1 - s_i.$$

Hence, for each player $i$, the *cost* function becomes

$$\begin{aligned} M_i &\equiv M_i(\mathbf{a}_{\mathrm{PF}(i)}, \mathbf{b}_{\mathrm{PF}(i)}) \\ &\equiv \begin{cases} C_i + \alpha_i r_i L_i, & \text{if } a_i = 1 \text{ (player } i \text{ invests)}, \\ b_i \widehat{p}_i L_i + (1 - b_i \widehat{p}_i) r_i L_i, & \text{if } a_i = 0 \text{ (player } i \text{ does not invest)}, \end{cases} \\ &= a_i [C_i + \alpha_i r_i L_i] + (1 - a_i)[b_i \widehat{p}_i L_i + (1 - b_i \widehat{p}_i) r_i L_i] \ . \end{aligned} \tag{7}$$

**The Internal Players' Best-Responses.** Letting $\widehat{\Delta}_i \equiv \rho_i / \widehat{p}_i = \frac{C_i}{\widehat{p}_i L_i}$ the adapted *cost-to-expected-loss ratio* of internal player $i$, and

$$\begin{aligned} \widehat{s}_i &\equiv \widehat{s}_i(\mathbf{a}_{\mathrm{Pa}(i)}, \mathbf{b}_{\mathrm{PF}(i)}) \\ &\equiv b_i s_i + \frac{1 - \alpha_i}{\widehat{p}_i} r_i = b_i s_i + \frac{1 - \alpha_i}{\widehat{p}_i}(1 - s_i) \\ &= \left( b_i - \frac{1 - \alpha_i}{\widehat{p}_i} \right) s_i + \frac{1 - \alpha_i}{\widehat{p}_i} \ , \end{aligned}$$

we get that the following best-response correspondence $\mathcal{BR}_i : \{0,1\}^{k_i - 1} \times \{0,1\}^{k_i} \to 2^{\{0,1\}}$:

$$\mathcal{BR}_i(\mathbf{a}_{\mathrm{Pa}(i)}, \mathbf{b}_{\mathrm{PF}(i)}) \equiv \begin{cases} \{1\}, & \text{if } \widehat{s}_i > \widehat{\Delta}_i, \\ \{0\}, & \text{if } \widehat{s}_i < \widehat{\Delta}_i, \\ \{0,1\}, & \text{if } \widehat{s}_i = \widehat{\Delta}_i. \end{cases} \tag{8}$$

### 3.2.3 The Aggressor's Payoff and Best-Response

The aggressor is assumed to have opposite goals to those of all other internal players: to cause as much damage as possible. One possible *utility/payoff* function $U$ quantifying the objective of the aggressor is [6]

$$U \equiv U(\mathbf{a}, \mathbf{b}) \equiv \sum_{i=1}^{n} u_i(\mathbf{a}_{\mathrm{PF}(i)}, \mathbf{b}_{\mathrm{PF}(i)}) - b_i C_i^0,$$

where

$$u_i(\mathbf{a}_{\mathrm{PF}(i)}, \mathbf{b}_{\mathrm{PF}(i)}) \equiv M_i - a_i C_i = (a_i \alpha_i r_i + (1 - a_i)(b_i \widehat{p}_i + (1 - b_i \widehat{p}_i) r_i)) L_i \ , \quad (9)$$

which adds the expected players costs (for targeted and transferred bad events) over all players, minus $C_i^0$, the aggressor's own *"cost" to target player* $i$. [7]

We close out this section by presenting the attacker's *best-response correspondence* $\mathcal{BR}_0 : \{0,1\}^n \to 2^{\{0,1\}^n}$:

$$\mathcal{BR}_0(\mathbf{a}) \equiv \arg\max_{\mathbf{b} \in \{0,1\}^n} U(\mathbf{a}, \mathbf{b}) \ . \quad (10)$$

**Definition 2.** *A joint-action* $(\mathbf{a}^*, \mathbf{b}^*) \in \{0,1\}^{2n}$ *is a* PSNE of an interdependent defense game *if, for each player* $i$, $a_i^* \in \mathcal{BR}_i(\mathbf{a}^*_{P_{a(i)}}, \mathbf{b}^*_{\mathrm{PF}(i)})$, *and for the aggressor,* $\mathbf{b}^* \in \mathcal{BR}_0(\mathbf{a}^*)$.

## 3.3 Limiting the Aggressor's Power

Note that the aggressor has in principle an *exponential* number of pure strategies! That is, the aggressor has the ability to attack *any* subset of the sites. We begin our study assuming that the aggressor has limited capabilities and restrict the number of possible attacks (e.g., at most $k$ sites can be simultaneously attacked). Future work will consider restricting the space of pure strategies by, for example, limiting distributional assumptions over possible pure strategies (e.g., the aggressor's set of pure strategies belongs to the set of joint-attack vectors corresponding to at most $K$ attacks) .

[6] Note that we can also consider the case in which the aggressor can target only a subset $\mathcal{T} \subset \{1, \ldots, n\}$ of all the sites. Of course, in that case, only nodes in $\mathcal{T}_{\mathrm{ext}} \equiv \cup_{i \in \mathcal{T}} CF(i)$ can be affected by an attack, either direct or indirect. Thus, under reasonable values of the parameters, $\mathbf{a}^*_{[n]-\mathcal{T}_{\mathrm{ext}}}$ equals either $\mathbf{1}$ or $\mathbf{0}$.

[7] Note that in this model, this cost can include the cost of getting caught or retaliated, among other things.

### 3.3.1 Limiting the Number of Simultaneous Attacks

Let us start simple and assume that there is at most a single simultaneous attack.

**Assumption 1.** *The set of pure strategies of the aggressor is*

$$\mathcal{B} = \left\{ \mathbf{b} \in \{0,1\}^n \;\middle|\; \sum_{i=1}^{n} b_i \leq 1 \right\}. \tag{11}$$

Any pure strategy in $\mathcal{B}$ is either a vector of all 0's, or exactly one 1.

We first note that in the case of at most one attack (Assumption 1), some of the expressions involving external risk/safety simplify considerably. For instance, in this case, we have

$$\begin{aligned} s_i \equiv s_i(\mathbf{a}_{\mathrm{Pa}(i)}, \mathbf{b}_{\mathrm{Pa}(i)}) &= \begin{cases} \sum_{j \in \mathrm{Pa}(i)} b_j e_{ij}(a_j), & \text{if } b_j = 1 \text{ for some } j \in \mathrm{Pa}(i), \\ 1, & \text{if } b_j = 0 \text{ for all } j \in \mathrm{Pa}(i), \end{cases} \\ &= 1 - \sum_{j \in \mathrm{Pa}(i)} b_j (1 - a_j) \widehat{q}_{ji}, \end{aligned}$$

so that

$$r_i(\mathbf{a}_{\mathrm{Pa}(i)}, \mathbf{b}_{\mathrm{Pa}(i)}) = \sum_{j \in \mathrm{Pa}(i)} b_j (1 - a_j) \widehat{q}_{ji},$$

and

$$b_i s_i = b_i.$$

Also, if the interdependent defense game has a PSNE $(\mathbf{a}^*, \mathbf{b}^*)$ in the single-attack case, then the aggressor's payoff in it is

$$U(\mathbf{a}^*, \mathbf{b}^*) = \left[ \max_{i \in [n]} (1 - a_i^*) \left( \widehat{p}_i L_i + \sum_{j \in \mathrm{Ch}(i)} \widehat{q}_{ij} (a_j^* \alpha_j + (1 - a_j^*)) L_j \right) - C_i^0 \right]^+$$

where for any real number $z \in \mathbb{R}$, the operator $[z]^+ \equiv \max(z, 0)$; in addition, if $b_k^* = 1$ for some $k \in [n]$, then

$$\begin{aligned} (1 - a_k^*) \left( \widehat{p}_k L_k + \sum_{j \in \mathrm{Ch}(k)} \widehat{q}_{kj} (a_j^* \alpha_j + (1 - a_j^*)) L_j \right) - C_k^0 \geq \\ \left[ \max_{i \in [n] - \{k\}} (1 - a_i^*) \left( \widehat{p}_i L_i + \sum_{j \in \mathrm{Ch}(i)} \widehat{q}_{ij} (a_j^* \alpha_j + (1 - a_j^*)) L_j \right) - C_i^0 \right]^+ . \end{aligned} \tag{12}$$

We will now show that in this case, under reasonable parameter values, there is no PSNE in interdependent defense games! We begin by introducing some assumptions on the game parameters.

The first assumption states that every site's investment cost is positive and (strictly) smaller than the *conditional* expected *direct* loss if the site were to be directly attacked ($b_i = 1$).

**Assumption 2.** *For all sites $i \in [n]$, $0 < C_i < \widehat{p}_i L_i$.*

The next assumption states that, for all sites $i$, the aggressor's cost to attack $i$ is positive and (strictly) smaller than the expected loss (i.e., gains from the perspective of the attacker) achieved if an attack initiated at site $i$ is successful, either directly at $i$ or at one of its children (after transfer). Or, roughly speaking, for every site, there is always the possibility that the site could be attacked if the conditions are "right" (i.e., neither the site nor its children are protected).

**Assumption 3.** *For all sites $i \in [n]$, $0 < C_i^0 < \widehat{p}_i L_i + \sum_{j \in Ch(i)} \widehat{q}_{ij} \alpha_j L_j$.*

The following proposition eliminates PSNE as a solution concept for "natural" interdependent defense games in which at most one attack is possible. The main significance of this result is that it allows us to concentrate our efforts on *mixed-strategy Nash equilibria (MSNE)*, a topic we consider in the following section.

**Proposition 1.** *No interdependent defense game in which Assumptions 1, 2 and 3 hold has a PSNE.*

**A Short Remark on Extensions.** We can further relax our restrictions on the set of pure strategies of the attacker in several directions. One is by extending our previous discussion and caping the number of attacks to $K > 1$. A further generalization along the same lines is to impose other more complex restrictions such as subdividing or creating a partition of the players/sites into groups, and limiting the maximum number of attacks both globally and within each group. The analysis under those alternative restrictions is likely to be significantly more involved than the case of one attack presented earlier, of course. We leave such an analysis for future work.

# 4 Allowing Mixed Strategies

We now extend the capabilities of the players by allowing them to play mixed strategies.

## 4.1 Notation and Preliminaries

For all player $i$, denote by $x_i \equiv \mathbf{P}(A_i = 1)$ the *mixed strategy of player $i$*: the probability that player $i$ invests. [8] In other words, the random variable $A_i \sim \text{Bernoulli}(x_i)$ models the (possibly randomized) decision of player $i$, with $\{A_i = 1\}$ corresponding to the event that player $i$ invests. Similarly, $\mathbf{B} \equiv (B_1, \ldots, B_n)$ is a vector of Bernoulli, not necessarily independent random variables, $y \equiv y_{\mathbf{B}} : \{0,1\}^n \to [0,1]$ denotes the joint PMF of $\mathbf{B}$ corresponding to the *aggressor's mixed strategy* so that for all $\mathbf{b} \in \mathcal{B}$, $y(\mathbf{b}) \equiv y_{\mathbf{B}}(\mathbf{b}) \equiv \mathbf{P}(\mathbf{B} = \mathbf{b})$ is the probability that the aggressor chooses joint-attack vector $\mathbf{b}$. [9]

Our model works within a non-cooperative setting (i.e., each player plays based on its own mixed strategy independently). Thus, the *joint mixed-strategy* of play is, for all joint pure-strategies $(\mathbf{a}, \mathbf{b}) \in \mathcal{A} \times \mathcal{B}$,

$$\mathbf{P}(\mathbf{A} = \mathbf{a}, \mathbf{B} = \mathbf{b}) = \mathbf{P}(\mathbf{B} = \mathbf{b}) \prod_{i=1}^{n} \mathbf{P}(A_i = a_i) = y(\mathbf{b}) \prod_{i=1}^{n} x_i^{a_i} (1 - x_i)^{1 - a_i}$$

where $\mathcal{A} \equiv \{0,1\}^n$ denotes the joint pure-strategy of the sites.

For simplicity, it will be convenient to denote the marginal PMF over a subset $I \subset [n]$ of the internal players by $y_I \equiv y_{\mathbf{B}_I}$ such that for all $\mathbf{b}_I \in \mathcal{B}_I$, $y_I(\mathbf{b}_I) \equiv y_{\mathbf{B}_I}(\mathbf{b}_I) \equiv \mathbf{P}(\mathbf{B}_I = \mathbf{b}_I) = \sum_{\mathbf{b}_{-I} \in \mathcal{B}_{-I}} y_{\mathbf{B}_I, \mathbf{B}_{-I}}(\mathbf{b}_I, \mathbf{b}_{-I})$ is the (marginal) probability that the aggressor chooses a joint-attack vector in which the sub-component decisions corresponding to players in $I$ are as in $\mathbf{b}_I$. It will also be convenient to denote by $y_i \equiv y_{\{i\}}(1) \equiv \mathbf{P}(B_i = 1)$ the marginal probability that the aggressor chooses an attack vector in which player $i$ is directly targeted.

[8] We can also view $x_i$ as a kind of investment level of protection of player $i$ and define the game such that the pure strategies of each player $i$ are given by the interval $[0,1] \subset \mathbb{R}$.

[9] We are intentionally considering a general set of possible pure strategies of the aggressor, denoted by $\mathcal{B}$, to allow for possible restrictions on this set, some of which were presented and discussed in the previous section. Of course, $\mathcal{B} = \{0,1\}^n$ when the aggressor's pure strategies are unconstrained, and the aggressor is thus allowed to consider an attack pure-strategy on any possible subset of the sites.

Slightly abusing notation, we redefine some of the functions previously introduce in the context of pure-strategies to the mixed strategy setting. For instance, the function $e_{ij}$ for how safe $i$ is from $j$ is now of type $[0,1]^2 \to \mathbb{R}$ so that

$$e_{ij}(x_j, y_j) \equiv \mathbf{E}[e_{ij}(A_j, B_j)] = x_j + (1 - x_j)(1 - y_j \widehat{q}_{ji}).$$

Because $e_{ij}(x_j, 0) = 1$, in the discussion below we abuse notation by ignoring the dependency on $y_j$ and use

$$e_{ij}(x_j) \equiv e_{ij}(x_j, 1) = x_j + (1 - x_j)(1 - \widehat{q}_{ji}) = 1 - (1 - x_j)\widehat{q}_{ji}$$

when it is clear from context. (Similarly for other previously-defined functions such as $s_i$, $r_i$, etc.)

To simplify notation, we define the random variables $S_i \equiv s_i(\mathbf{A}_{\mathrm{Pa}(i)}, \mathbf{B}_{\mathrm{Pa}(i)})$ and $R_i \equiv 1 - S_i$ .

## 4.2 The Expected Costs of the Internal Players

In general, the *expected* cost of protection to site $i$, with respect to a joint mixed-strategy $(\mathbf{x}, y)$, is now

$$\begin{aligned} M_i(\mathbf{x}, y) &\equiv M_i(\mathbf{x}_{\mathrm{PF}(i)}, y_{\mathrm{PF}(i)}) \equiv \mathbf{E}[M_i(\mathbf{A}_{\mathrm{PF}(i)}, \mathbf{B}_{\mathrm{PF}(i)})] \\ &= x_i \left(C_i + \alpha_i \mathbf{E}[R_i] L_i\right) + (1 - x_i) \left(\widehat{p}_i \mathbf{E}[B_i S_i] + \mathbf{E}[R_i]\right) L_i \ , \end{aligned} \tag{13}$$

where $\mathbf{E}[R_i] = 1 - \mathbf{E}[S_i]$,

$$\mathbf{E}[S_i] = \sum_{I \subset \mathrm{Pa}(i)} y_{I, \mathrm{Pa}(i) - I}(\mathbf{1}, \mathbf{0}) \prod_{j \in I} (1 - (1 - x_j)\widehat{q}_{ji}) \ ,$$

and

$$\mathbf{E}[B_i S_i] = \sum_{I \subset \mathrm{Pa}(i)} y_{\{i\}, I, \mathrm{Pa}(i) - I}(1, \mathbf{1}, \mathbf{0}) \prod_{j \in I} (1 - (1 - x_j)\widehat{q}_{ji}) \ .$$

## 4.3 The Attacker's Expected Payoff

For each internal player $i$, define the random variable $U_i \equiv u_i(\mathbf{A}_{\mathrm{PF}(i)}, \mathbf{B}_{\mathrm{PF}(i)})$. Define the random variable $U \equiv \sum_i U_i - b_i C_i^0$.

The expected payoff of the aggressor is simply

$$U(\mathbf{x}, y) \equiv \mathbf{E}[U] = \mathbf{E}\left[\sum_i U_i - b_i C_i^0\right] = \sum_i \mathbf{E}[U_i] - y_i C_i^0$$

where

$$\begin{aligned}\mathbf{E}[U_i] &= ((1 - x_i)\widehat{p}_i \mathbf{E}[B_i S_i] + (x_i \alpha_i + (1 - x_i))\mathbf{E}[R_i])L_i \\ &\equiv u_i(\mathbf{x}_{\mathrm{PF}(i)}, y_{\mathrm{PF}(i)}) \ .\end{aligned}$$

## 4.4 Best-Response Correspondences of Internal Players

Let

$$\hat{s}_i \equiv \hat{s}_i(\mathbf{x}_{\mathrm{Pa}(i)}, y_{\mathrm{PF}(i)}) \equiv \mathbf{E}[B_i S_i] + \frac{1 - \alpha_i}{\widehat{p}_i}\mathbf{E}[R_i].$$

Denote by $\mathcal{P}^m$ the set of all possible joint probability mass functions over $m$ Bernoulli random variables. Recall that $k_i \equiv |\,\mathrm{PF}(i)|$. The best-response correspondence $\mathcal{BR}_i : [0,1]^{k_i - 1} \times \mathcal{P}^{k_i} \to [0,1]$ of site $i$ is

$$\mathcal{BR}_i(\mathbf{x}_{\mathrm{Pa}(i)}, y_{\mathrm{PF}(i)}) \equiv \begin{cases} \{1\}, & \text{if } \hat{s}_i > \widehat{\Delta}_i, \\ \{0\}, & \text{if } \hat{s}_i < \widehat{\Delta}_i, \\ [0,1], & \text{if } \hat{s}_i = \widehat{\Delta}_i. \end{cases} \tag{14}$$

It is important to remember that $\mathcal{BR}_i$ of player $i$ depends only on its parents' mixed strategies $\mathbf{x}_{\mathrm{Pa}(i)}$ and aggressor's mixed-strategy marginal over $i$'s parent family. This point will become particularly relevant when considering restrictions on the complexity of the aggressor's mixed strategies.

## 4.5 The Aggressor's Best-Response Correspondence

The best-response correspondence for the aggressor is simply

$$\mathcal{BR}_0(\mathbf{x}) \equiv \arg\max_{y \in \mathcal{P}^n} U(\mathbf{x}, y). \tag{15}$$

We can also express the conditions for the aggressor's best response as, $y^* \in \mathcal{BR}_0(\mathbf{x})$ if and only if

$$\sum_{i=1}^n u_i(\mathbf{x}_{\mathrm{PF}(i)}, y^*_{\mathrm{PF}(i)}) - y_i^* C_i^0 \geq \sum_{i=1}^n u_i(\mathbf{x}_{\mathrm{PF}(i)}, \mathbf{b}_{\mathrm{PF}(i)}) - b_i C_i^0$$

for all $\mathbf{b} \in \{0,1\}^n$.

**Definition 3.** *A joint mixed-strategy* $(\mathbf{x}^*, y^*)$ *is a* mixed-strategy Nash equilibrium (MSNE) *of the interdependent defense game if (1) for all sites* $i \in [n]$, $x_i^* \in \mathcal{BR}_i(\mathbf{x}^*_{\mathrm{PF}(i)}, y^*_{\mathrm{PF}(i)})$ *and (2)* $y^* \in \mathcal{BR}_i(\mathbf{x}^*)$.

## 4.6 A Characterization of the MSNE

Recall that the space of pure strategies of the aggressor is, in its most general form, exponential in the number of internal players. This is an obstacle to tractable computational representations in large-population games.

In the following result, we establish an equivalence class over the aggressor's mixed strategy that allows us to only consider "simpler" mixed strategies in terms of their probabilistic structure.

**Proposition 2.** *For any mixed strategy of the interdependent defense game* $(\mathbf{x}^*, y^*)$*, with the aggressor's utility defined as above, there exists another mixed strategy* $(\mathbf{x}^*, \widetilde{y})$*, such that*

1. *the joint PMF* $\widetilde{y}$ *decomposes as* [10]

$$\widetilde{y}(\mathbf{b}) \propto \prod_{i=1}^{n} \Phi_{\mathrm{PF}(i)}(\mathbf{b}_{\mathrm{PF}(i)})$$

   *for some non-negative functions* $\Phi_{\mathrm{PF}(i)} : \{0,1\}^{k_i} \to [0,\infty)$*, and all* $\mathbf{b} \in \{0,1\}^n$,

2. *for all* $i \in [n]$*, the parent-family marginal PMFs* $\widetilde{y}_{\mathrm{PF}(i)} = y^*_{\mathrm{PF}(i)}$ *agree, and*

3. *the sites and the aggressor achieve the same expected cost and utility, respectively, in* $(\mathbf{x}^*, \widetilde{y})$ *as in* $(\mathbf{x}^*, y^*)$*: for all* $i \in [n]$,

$$M_i(\mathbf{x}^*_{\mathrm{PF}(i)}, \widetilde{y}_{\mathrm{PF}(i)}) = M_i(\mathbf{x}^*_{\mathrm{PF}(i)}, y^*_{\mathrm{PF}(i)}),$$

   *and*

$$U(\mathbf{x}^*, \widetilde{y}) = U(\mathbf{x}^*, y^*).$$

[10] In other words, $\widetilde{y}$ is a Gibbs distribution with respect to the undirected "moralized" graph that results from adding an (undirected) edge among every pair of parents of every node to the original directed graph of the game and ignoring the directions of the edges in the original game graph.

**Corollary 1.** *For any interdependent defense game, let* $k_{\max} \equiv \max_{i\in[n]} k_i$ *be the size of the largest parent-family in the game graph. The representation size of any mixed strategy of the aggressor in the game is* $O\left(2^{k_{\max}}\right)$*, modulo expected-payoff equivalence.*

### 4.7 The Case of at Most One Attack

In this case, the expressions given previously simplify considerably. [11] Let $y_0 \equiv P(\mathbf{b} = \mathbf{0})$. In particular, we have $\sum_{i=0}^{n} y_i = 1$,

$$\mathbf{E}[S_i] = 1 - \sum_{j\in \mathrm{Pa}(i)} y_j(1 - x_j)\widehat{q}_{ji} \ ,$$

so that $\mathbf{E}[R_i] = \sum_{j\in \mathrm{Pa}(i)} y_j(1 - x_j)\widehat{q}_{ji}$, and

$$\mathbf{E}[B_i S_i] = \mathbf{P}(B_i = 1, \mathbf{B}_{\mathrm{Pa}(i)} = \mathbf{0}) = y_i \ .$$

This leads to a simplification of the expected cost of the internal players:

$$\begin{aligned}\mathbf{E}[M_i] = x_i C_i + (1 - x_i) y_i \widehat{p}_i L_i + \\ (x_i\alpha_i + (1 - x_i))L_i \sum_{j\in \mathrm{Pa}(i)} y_j(1 - x_j)\widehat{q}_{ji} \ .\end{aligned}$$

# 5 Interdependent Defense Games under Single Attack and Full Transfer Vulnerability

In this section, we provide a polynomial-time algorithm to compute a MSNE in interdependent defense games in which investment in security provides no protection from transfers (i.e., $\alpha_i = 1$ for all internal players $i$) and the aggressor's capabilities are limited to at most one attack (i.e., Assumption 1 holds).

[11] Also, in this case, we can view $y_i$ as a kind of level of effort by the aggressor to each target $i \in [n]$, similar to the alternative view of $x_i$ presented previously, and define the game such that the pure strategies of the aggressor are given by the $n$-simplex $\{(z_0, z_1, \ldots, z_n) \in \mathbb{R}^{n+1} \mid \sum_{i=0}^{n} z_n = 1 \text{ and } z_i \geq 0 \text{ for all } i\}$.

## 5.1 Preliminaries

We begin by formally stating our assumption on the $\alpha$'s.

**Assumption 4.** *For all internal players $i \in N$, the probability $\alpha_i = 1$ that player $i$'s investment in security does not protect the player from transfers.*

This is the same assumption implicit in the original IDS games.

**Definition 4.** *We say an interdependent defense game is* transfer-vulnerable *if Assumption 4 holds.*

Also, recall that Assumption 1, in the context of mixed strategies, implies $\sum_i^n y_i + y_0 = 1$.

**Definition 5.** *We say an interdependent defense game is a* single-attack *game if Assumption 1 holds (i.e., at most one attack is possible).*

Combining Assumptions 1 and 4 allows us to greatly simplify the best-response condition of the internal players because now $\widehat{s}_i = y_i$.

Let $L_i^0(x_i) \equiv (1 - x_i)(\widehat{p}_i L_i + \sum_{j \in \text{Ch}(i)} \widehat{q}_{ij} L_j)$. Later it will be convenient to denote by $\overline{L}_i^0 \equiv L_i^0(0) = \widehat{p}_i L_i + \sum_{j \in \text{Ch}(i)} \widehat{q}_{ij} L_j$, so that we can express $L_i^0(x_i) = (1 - x_i)\overline{L}_i^0$, to highlight that $L_i^0$ is linear function of $x_i$. Similarly, it will also be convenient to let $M_i^0(x_i) \equiv L_i^0(x_i) - C_i^0$, and denote $\overline{M}_i^0 \equiv M_i^0(0) = \overline{L}_i^0 - C_i^0$. Let $\eta_i^0 \equiv C_i^0 / \overline{L}_i^0$.

The best-response condition of the attacker also simplifies under the same assumptions because now $E[U] = \sum_{i=1}^n y_i M_i^0(x_i)$.

We will make use of Assumption 3 stated earlier. That assumption is also reasonable in our new context because, under Assumption 4, if there were a player $i$ with $\eta_i^0 > 1$, the aggressor would never attack $i$, and as a result player $i$ would never invest. In that case, we can safely remove $j$ from the game, without any loss of generality.

## 5.2 Characterizing the MSNE

We now characterize the space of MSNE in interdependent defense games in a way that leads immediately to a polynomial-time algorithm for computing a single MSNE.

We will start by partitioning the space of games into three, based on whether $\sum_{i=1}^n \widehat{\Delta}_i$ is (1) less than, (2) equal to, or (3) greater than 1.

**Proposition 3.** *The joint mixed-strategy* $(\mathbf{x}, \mathbf{y})$ *is an MSNE of a single-attack transfer-vulnerable interdependent defense game in which* $\sum_{i=1}^{n} \widehat{\Delta}_i < 1$ *if and only if it satisfies the following properties.*

1. *There may not be an attack:* $1 > y_0 = 1 - \sum_{i=1}^{n} \widehat{\Delta}_i > 0$.

2. *Every internal player has non-zero chance of being attacked, and this probability equals the respective internal player's cost-to-expected-loss ratio: for all internal players* $i \in [n]$, $y_i = \widehat{\Delta}_i > 0$.

3. *Every internal player invests some but none does fully, and in particular, the probability a player does* not *invest equals the respective cost-to-loss ratio to the attacker: for all internal players* $i \in [n]$, $0 < x_i = 1 - \eta_i^0 < 1$.

**Proposition 4.** *The joint mixed-strategy* $(\mathbf{x}, \mathbf{y})$ *is an MSNE of a single-attack transfer-vulnerable interdependent defense game in which* $\sum_{i=1}^{n} \widehat{\Delta}_i = 1$ *if and only if it satisfies the following properties.*

1. *There is always an attack:* $y_0 = 0$.

2. *Every internal player has non-zero chance of being attacked, and this probability equals the respective internal player's cost-to-expected-loss ratio: for all internal players* $i \in [n]$, $y_i = \widehat{\Delta}_i > 0$.

3. *No internal player invests fully, and the possible investment probabilities are connected by a bounded 1-d interval (manifold) in* $\mathbb{R}^n$*:*

$$x_i = 1 - \frac{v + C_i^0}{\overline{L}_i^0} \text{ for all } i \in [n]$$

*with* $0 \leq v \leq \min_{i \in [n]} \overline{M}_i^0$.

**Proposition 5.** *The joint mixed-strategy* $(\mathbf{x}, \mathbf{y})$ *is an MSNE of a single-attack transfer-vulnerable interdependent defense game in which* $\sum_{i=1}^{n} \widehat{\Delta}_i > 1$ *if and only if it satisfies the following properties.*

1. *There is always an attack:* $y_0 = 0$.

2. *There exists a non-singleton, non-empty subset* $I \subset [n]$*, such that* $\min_{i \in I} \overline{M}_i^0 \geq \max_{k \notin I} \overline{M}_k^0$*, if* $I \neq [n]$*, and the following holds.*

(a) *No internal player outside $I$ invests or is directly attacked:* $x_k = 0$ *and* $y_k = 0$ *for all* $k \notin I$.

(b) *Let* $J \equiv \arg\min_{i \in I} \overline{M}_i^0$. *No internal player in* $J$ *invests and the probability of that player being directly attacked is at most the player's cost-to-expected-loss ratio: for all* $i \in J$, $x_i = 0$ *and* $0 \leq y_i \leq \widehat{\Delta}_i$; *in addition,* $\sum_{i \in J} y_i = 1 - \sum_{t \in I - J} \widehat{\Delta}_i$.

(c) *Every internal player in* $I - J$ *partially invests and is attacked with probability equal to the player's cost-to-expected-loss ratio: for all* $i \in I - J$, $y_i = \widehat{\Delta}_i$ *and*

$$0 < x_i = 1 - \frac{\min_{t \in I} \overline{M}_t^0 + C_i^0}{\overline{L}_i^0} < 1.$$

Hence, from the proof of the last proposition we can infer that if the $\overline{M}_l^0$'s form a complete order, then the last condition allows us to search for an MSNE by exploring only $n - 2$ sets, as opposed to $2^{n-2}$ if done naively.

It turns out a complete order is not necessary. The following claim allows us to safely move all the internal players with the same value of $\overline{M}_i^0$ in a group as a whole inside or outside $I$.

**Claim 1.** *Let* $I \subset [n]$, *such that* $I' \subset I$, $|I'| < |I| < n - 1$. *Suppose we find an MSNE* $(\mathbf{x}, \mathbf{y})$ *such that* $I' = \{i \mid y_i > 0\}$, *with the property that* $\min_{l \in I'} \overline{M}_l^0 = \max_{k \notin I'} \overline{M}_k^0$. *In addition, suppose* $I$ *satisfies* $\min_{l \in I'} \overline{M}_l^0 = \min_{l \in I} \overline{M}_l^0 \geq \max_{k \notin I} \overline{M}_k^0$. *Then, we can also find* $(\mathbf{x}, \mathbf{y})$ *using partition* $I$.

## 5.3 Algorithms

Propositions 3, 4 and 5 provide an extremely simple characterization of the MSNE of single-attack transfer-vulnerable interdependent defense games. That characterization leads immediately to a polynomial-time algorithm for computing (essentially) *all* MSNE in these games. First note that the equilibrium in the case of IDD games with $\sum_{i=1}^n \widehat{\Delta}_i \leq 1$ has essentially an analytic closed-form. Hence, we concentrate on the remaining case of $\sum_{i=1}^n \widehat{\Delta}_i > 1$.

Armed with Proposition 4 and Claim 1, we now describe the part of the algorithm corresponding to single-attack transfer-vulnerable interdependent defense games with $\sum_{i=1}^n \widehat{\Delta}_i > 1$. We start by sorting the indices of the

internal players in descending order based on $\overline{M}_i^0$'s. Let $\text{Val}(l)$ and $\text{Idx}(l)$ be the $l$th value and index in the resulting sorted list, respectively. Find $t$ such that $1 - \widehat{\Delta}_{\text{Idx}(t)} \leq \sum_{l=1}^{t-1} \widehat{\Delta}_{\text{Idx}(l)} < 1$. Let $k = \arg\max\{l \geq t \mid \text{Val}(l) = \text{Val}(t)\}$ (i.e., continue down the sorted list of values until a change occurs). For $i = 1, \ldots, t-1$, let $l = \text{Idx}(i)$ and set $x_l = 1 - \frac{\text{Val}(t)+C_l^0}{\overline{L}_l^0}$ and $y_l = \widehat{\Delta}_l$. For $i = k+1, \ldots, n$, let $l = \text{Idx}(i)$ and set $x_l = 0$ and $y_l = 0$. For $i = t, \ldots, k$, let $l = \text{Idx}(i)$ and set $x_l = 0$. Finally, represent the simplex defined by the following constraints: for $i = t, \ldots, k$, let $l = \text{Idx}(i)$ and $0 \leq y_l \leq \widehat{\Delta}_l$; $\sum_{i=t}^{k} y_{\text{Idx}(i)} = 1 - \sum_{i=1}^{t-1} \widehat{\Delta}_{\text{Idx}(i)}$. The running time of the algorithm is $O(n \log n)$ (because of the sorting needed).

Having established and discussed the appropriate definitions, we now state our main computational result.

**Theorem 1.** *There exists a polynomial-time algorithm to compute all MSNE of a single-attack transfer-vulnerable interdependent defense game.*

To re-emphasize, note that in the cases in which the equilibria is not unique, it can be generated via simple sampling of either a simple linear system or a simplex. In either case, one can compute a single MSNE from that infinite set in polynomial time.

Let us revisit the types of games that may have an infinite MSNE set. Note that the case in which $\sum_{i=1}^{n} \widehat{\Delta}_i = 1$ has measure zero within the space of parameters. It is also quite brittle in the sense that adding or removing a player breaks the equality. For the case in which $\sum_{i=1}^{n} \widehat{\Delta}_i > 1$, which seems like the most reasonable of all three cases, if the value of the $\overline{M}_i^0$'s are distinct, [12] then there is a unique MSNE!

## 5.4 Some Remarks

Note that if, at equilibrium, an internal player invests in security with positive probability, then that probability of investing not only depends on the cost to the attacker (which is expected), but also on the expected losses the player's non-investing could cause to the player's *children*. So, the player's probability of investing is a function of the player's *family* (i.e., the player and the player's children), which includes the player itself. More formally, at equilibrium $\mathbf{x}$,

[12] Actually, as can be seen from Proposition 5 and the description of the algorithm, a weaker requirement is enough to achieve a unique MSNE. All we need is for the set at which the sum goes over one to guarantee unique MSNE.

if player $i$'s probability of investing $x_i > 0$, the probability of *not* investing is proportional to the cost $C_i^0$ incurred by the aggressor to attack $i$ and *inversely* proportional to the expected loss $\widehat{p}_i\overline{L} + \sum_{j\in \mathrm{Ch}(i)} q_{ij}L_j$ that could cause to the player $i$'s family an attack on that player $i$ should $i$ not invest in security. It is kind of reassuring the at equilibrium, which is the (almost-surely) unique stable outcome of the system, the probability of investing increases with the potential loss a player's non-investment decision could cause to the system: The higher the expected loss from not investing to a player's children, the higher the probability of that player investing.

Hence, behavior in a stable system implicitly "forces" all players to indirectly account for or take care of their own children. This may sound a bit paradoxical at first given that we are working within "noncooperative" setting and each player's cost function is only dependent on the investment decision of the player's *parents*. What is happening here is that the existence of the attacker in the system is inducing an (almost-surely) unique stable outcome in which an implicit form of "cooperation" occurs. An internal player's best response is independent of their parents, the source of transfer risk, if investment in security does nothing to protect that player from transfers (i.e., $\alpha_i = 1$). This makes sense because the player cannot control the transfer risk. Said differently, there is nothing the player can do to prevent the transfer, even though the original potential for transfers does depend on the parents' investment strategies.

In short, rational/optimal noncooperative behavior for each player is not only to protect for the player's own losses but also "cooperate" to protect the player's children.

How does the network structure and the equilibrium relate? As seen above, the values of the equilibrium strategy of each player are local in the sense that they depend on information from the attacker, the player and the player's children only. It was evident from the previous discussion that a player's probability of investing at the equilibrium increases with the expected loss sustained from a "bad event" occurring as a result of a transfer from a player to the player's children. So overall, the higher the player's children losses, the higher the probability of investing.

Let us explore this last point further. Consider the case of uniform-transfer probabilities. Analogous to uniform-transfer IDS games [Kunreuther and Heal, 2003, Kearns and Ortiz, 2003], in that case transfer probabilities are only a function of the source, not the destination: $\widehat{q}_{ij} \equiv \widehat{\delta}_i$. The expression for

the equilibrium probabilities of those players who have a positive probability of investing would simplify to

$$x_i = 1 - \frac{v + C_i^0}{\widehat{p}_i L_i + \delta_i \sum_{j \in \text{Ch}(i)} L_j} \ ,$$

for some constant $v$. The last expression suggests that what differentiates the probability of investing between players is just the sum of the children's losses, $\sum_{j \in \text{Ch}(i)} L_j$. That would suggests the larger the number of children the larger the probability of investing. A scenario that seems to further lead us to that conclusion is when we make the further assumption that the value of all the loss parameter $L_i \equiv L$ is the same for all players. Then, we would get

$$x_i = 1 - \frac{v + C_i^0}{L\left(\widehat{p}_i + \delta_i |\text{Ch}(i)|\right)} \ .$$

We can go even further and assume a homogeneous system in the sense that, in addition to the above, each type of parameter has the same value across all players: $\widehat{p}_i \equiv \widehat{p}$, $\delta_i \equiv \delta$, and $C_i^0 \equiv C^0$. [13] We finally obtain

$$x_i = 1 - \frac{v + C^0}{L\left(\widehat{p} + \delta |\text{Ch}(i)|\right)} \ .$$

So the probability of *not* investing is inversely proportional to the *number* of children the player has.

However, what is missing from the discussion above is that the risk-related parameters are constrained: $\widehat{p}_i + \sum_{j \in \text{Ch}(i)} \widehat{q}_{ij} \leq 1$. Considering the previous two cases of uniform-transfer and homogenous-parameters/players, the resulting conditions would be $\delta_i |\text{Ch}(i)| \leq 1 - \widehat{p}_i$ and $\delta |\text{Ch}(i)| \leq 1 - \widehat{p}$, respectively. So, clearly, we cannot increase the number of children of a player arbitrarily without decreasing the uniform or constant transfer probability. So our discussion above holds as long as the conditions just stated on the risk-related parameters hold. If we fix the maximum number of children, that would impose an upper bound on the transfer probabilities, and *vice versa*. It is important to understand the implicit assumptions we would be making by fixing either the maximum number of children or the transfer probabilities. For example, if we fix the transfer probabilities we are implicitly assuming

[13] Note that this does not mean that the expected loss caused by a player that does not invest but is attacked, $L\left(\widehat{p} + \delta |\text{Ch}(i)|\right)$, is the same for all players.

that adding a new child that a player can transfer to does not change the total probability of transfer to the previously existing children, it just adds to the total probability, so that no normalization takes place.

Let us now explore the properties of the set of internal players that the aggressor could potentially attack at equilibrium (i.e., the support of the attacker's mixed strategy; formally, $I \equiv \{i \mid y_i > 0\}$). This set has the property that internal players for which the attacker's cost-to-expected-loss is higher are "selected" first in the algorithm. In other words, the expected payoff to the attacker is in this sense, the "minimum" possible. In addition, if the size of that set is $t$, and there is a lower bound on the internal players' cost-to-expected-loss ratio $\widehat{\Delta}_i > \widehat{\Delta}$, and $\sum_{i=1}^{n} \widehat{\Delta}_i > 1$, then $t/n < 1/(\widehat{\Delta}n)$ is an upper-bound on the proportion of players that could potentially be attacked (where $t$ is of course the size of that set). Also, if we have a game with homogeneous parameters, then the probability of an attack will be uniform over that set $I$ (almost surely), so that every player in that set is attacked with exactly the same probability. Finally, all but one of the players in that set $I$ invest in security with some non-zero probability (almost surely).

# 6 Experiments

We obtain the structure and topology of the Internet from DIMES (netdimes.org) [Shavitt and Shir, 2005]. Namely, we use the Autonomous Systems (AS) graphs of the Internet. The dataset consists of a set of AS level nodes and AS level edges which were found in March 2010 and were seen at least twice. The connection between any two nodes is defined as the edge between them. The data set consist of 27106 nodes and 100402 edges. The data set consist of 27106 nodes (683 of them are isolated nodes) and 100402 edges. The graph length (diameter) is 6253 and the density (number of edges divided by number of possible edges) is $1.9920 \times 10^{-5}$. Figure 1 shows the indegree and outdegree distribution of the graph.

For this experiments, we consider single-attack IDD games with arbitrary transfer vulnerability $\boldsymbol{\alpha}$. We do not have any algorithm tailored for the case of arbitrary transfer vulnerability. We use best-response gradient-dynamics heuristic, a well-known technique from the work on learning in games [Fudenberg and Levine, 1999, Singh et al., 2000, Kearns and Ortiz, 2003, Heal and Kunreuther, 2005, Kearns, 2005], to search for an $\epsilon$-approximate MSNE, [14]

[14] An $\epsilon$-approximate MSNE is a joint mixed strategy in which the gain for individual

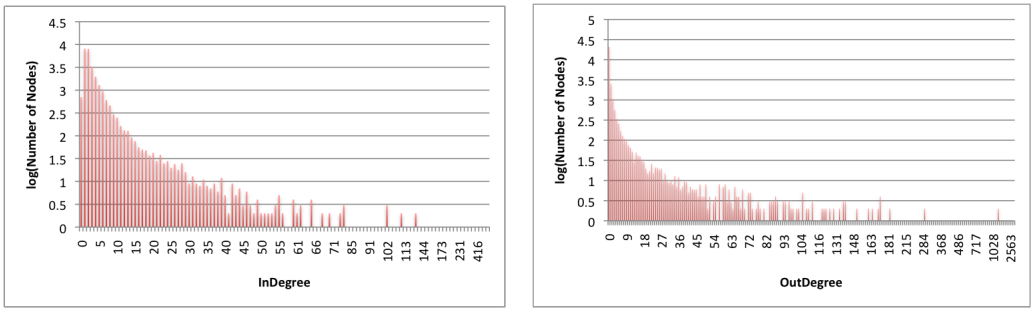


Figure 1: The bar graph on the left and on the right shows the number of node with a particular outdegree value and indegree value, respectively. (The graphs only showing the in/out degrees with non-zero number of node.)

under an properly normalized payoffs/costs functions. During the initialization phase of the program, we construct a (network) graph using the data from DIMES. We also compute the transfer vulnerability and initialize all the parameter values for each node in the network, as explained in the next paragraph. Note that in this step, we randomly initialize the $y_i$ and $x_i$ values of each node. At each iteration, we compute and update each player's utility (normalized), attacker's utility, $y_i$ and $x_i$ for each node, and other data according to the IDD model. In addition, we output the top 360 nodes with the highest $y_i$ along with $x_i$, $\Delta_i$, and other statistics. In the experiments, we run the process until it reaches a $\epsilon$-MSNE or a maximum of 1000 iterations.

Given the heuristic nature of best-response gradient dynamics, we experimentally evaluated the running-time convergence behavior. We ran ten simulations for each of the $\epsilon$ value and recorded the number of iterations until convergence, or 1000 if the run did not converge, which in this case happened only twice, and it occurred when $\epsilon = 0.001$.

In Figure 2, we observe the number of iterations took for varies epsilon-MSNE to converge. We can see that the trend lines of the graph is very close to $\frac{1}{x}$ with $R^2$ relatively close to 0.9 (including the two non-convergence data point for $\epsilon = 0.001$). This is a good indication that the running-time convergence behavior of best-response gradient dynamics is captured by function close to $\frac{1}{\epsilon}$.

We ran two sets of experiments based on randomly generated or fixed

unilateral deviation is no larger than $\epsilon$. Hence, a 0-approximate MSNE is an exact MSNE.

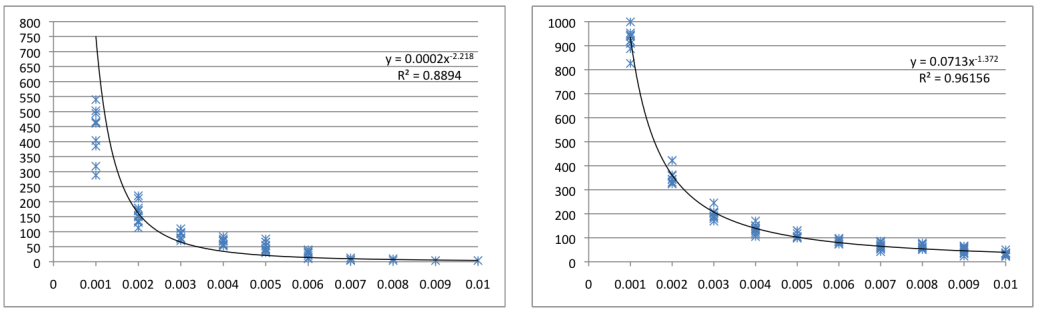


Figure 2: The x-axis represents the $\epsilon$ value and the y-axis represents the number of iteration until convergence (or 1000 iterations) to some $\epsilon$-MSNE (left: fixed parameters, right: randomly generated parameters). The trend line is fitted to the graph and is shown at the upper right corner of the graph along with its $R^2$ value. It roughly behaves like a low degree polynomial of $1/x$, which is somewhat consistent with some theoretical results on learning dynamics.

parameters. In the experiments with fixed parameters, we set the parameters as follow: $\alpha = 0.5/20$, $L_i = 10^8 + (10^9) * 0.5$, $C_i = 10^5 + (10^6) * 0.5$, $p_i = 0.8 + 0.5/10$. For the experiments with randomly generate parameters, we simply replace all the "0.5"'s in the expressions above by a random number distributed uniformly over $[0, 1]$. In both sets of the experiment, the value of the transfer risk probability parameters $q_{ij}$'s is set as $q_{ij} \propto Z_{ij} \times (|\text{Ch}(j)| + |\text{Pa}(j)|)$, where the random (variable) factor $Z_{ij} \sim \text{Uniform}([0.2, 0.4])$ i.i.d., for all $j \in \text{Ch}(i)$. As stated previously, the $y_i$'s and $x_i$'s are independently, identically distributed random variables with a uniform distribution over $[0, 1]$. Finally, we set $\epsilon$ to 0.005 and the maximum number of iterations to 1000. In the sequel, we show typical plots of the solution obtained in each run.

As shown in Figure 3, the experiments suggest that there is a correlation between $x_i$ and $y_i$; the sites that have lower $y_i$ values tend to have a higher $\frac{x_i}{y_i}$ value.

In Figure 4, we plot the topological structure of the top sites (in this case 360) with the highest $y_i$ and their immediately neighbors. We do similarly for $x_i$.

Notice that there are a few numbers of isolated individual nodes and

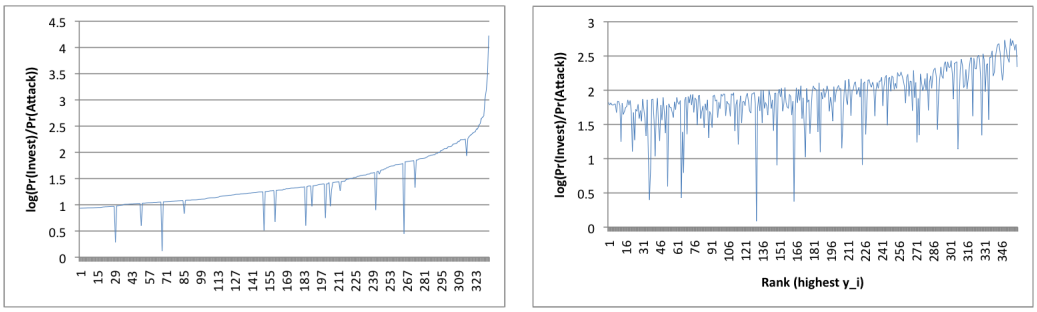


Figure 3: The graphs on the left (fixed parameters) and on the right (randomly generated parameters) show the $\frac{x_i}{y_i}$ value of a node. The nodes are ordered decreasingly based on the $y_i$ value.

few small networks, but in general, the graph tends to have a cluster-like structure as seen in the fixed parameters case. One of the reason the isolated nodes are target is that they do not invest enough and therefore result in a higher utility gain for the attacker.

Figure 4 also shows the number of connected components of the network for the subgraph of the nodes most likely to being attacks (and their neighbors) as well as those of the network for the subgraph of the nodes with the highest probability of investment, along with some additional properties of the graphs.

The plots in Figure 5 show the in-degree and out-degree of the 360 sites with highest $y_i$ and $x_i$. One of the reason that the top 360 nodes (with the highest $y_i$) have higher in degree than out degree in general is because those nodes have high out degree decided to protect themselves, if they don't have "high enough" $x_i$, the attacker would attack them because the attacker would gain more utility compare to those with low out degree.

On the other hand, it might make sense that those do not have high out degree being targeted because they felt safe in turns of the attacker's utility and lower their $x_i$ therefore we are seeing this such relationship. One thing we notice in Figure 5 is that there is not a correlation between the $y_i$ or the $x_i$ and its corresponding in-degree and out-degree (which,. Similarly we cannot see any pattern in the case of fixed parameters. Also, note that the $\alpha_i$ we are using for each player is relatively low (i.e., uniformly distributed over $[0, 1/40]$); yet, interestingly, such a behavior is also predicted by the theory

for the case $\alpha_i = 1$.

# 7 Future Work

**Attackers Can Affect Transfer Probabilities.** We could extend the strategy space of the attacker by allowing the attacker to affect transfer. One particular instantiation of this idea is to have the network graph *edges* represent the attacker's targets, as opposed to just the node. The attacker's pure strategies would now be based on the edges $(i, j)$, such that binary action variable $b_{ij}$ would now represent the attack, taking a value of one if the attackers wants to attack $j$ but only via a transfer from $i$.

**Multiple Attackers with Multiple Attacks.** While dealing with multiple attackers is outside the scope of this paper, we have in fact extended the model in a natural way in that direction. However, we were able to extend the representation results, but not the characterization or computational/algorithmic results. We leave that endeavor for future work. In principle, the best-response gradient dynamics can also be used as a heuristic in the multiple attackers' case.

**Open Problems.** A thorough characterization of the equilibria of interdependent defense games is lacking, specially for the case of multiple potential attacks by multiple aggressors. Also, we need a better understanding of the effect of network structure of the game and restrictions on the aggressors' available strategies on the equilibria of the game.

Many computational problems in the context of interdependent defense games remain open.

1. What is the computational complexity of the problem of computing equilibria of interdependent defense games with arbitrary transfer vulnerability? (e.g., a single, multiple or all MSNE? MSNE with particular properties?)

2. What is the computational complexity of the problem of identifying "influential" agents, in the sense of Irfan and Ortiz [2011] (see also, Kleinberg [2007] and the references therein)?

3. How is the complexity affected by network structure or restrictions on the aggressors' available strategies? For example, what if the network graph is some type of chain, cycle or tree?

# 8 Summary of Contributions

In this paper, we propose IDD games, an adaptation of IDS games to the setting in which the attack is deliberate and the attacker is explicitly modeled. We consider the special case of the single attack scenario as a way to limit the attackers power, and prove that no PSNE exists in such subclass of games. We then consider randomized strategies and derive the appropriate expressions for the expected costs of the internal players and the expected payoff of the attacker, and consequently their respective best-response correspondence.

We study in depth the case in which only one attack is possible and investment in security does nothing to protect the players from the transfer risk (which is the same implicit assumption made in the original IDS work). We completely characterize the MSNE of such a subclass of IDD games. We prove that, almost surely, every game in that subclass has a unique MSNE, which can be almost determined analytically.

That result immediately lead to a simple algorithm for computing the equilibrium that only requires a sorting of the cost-to-expected-loss ratio gain of the attacker for each player. Hence, the algorithm runs in $O(n \log n)$, where $n$ is the number of internal players.

We then discuss some corollaries of the characterization and highlight the connection between the network structure and the investment of players at equilibrium. In particular, we show how investment probabilities at equilibrium essentially reflect some degree of "cooperation" (in a fully non-cooperative setting), in that players want to protect their own *children* in the network graph (i.e., for each player $i$, the set of players to which player $i$ can *transfer*), and have no direct dependence on the player's *parents*, which are the true source of the risk to the player. In particular, we show how the probability of investment can *increase* with the number of children.

Finally, we built a generative model of single-attack IDD games based on real Internet graph (at the AS level) obtained from DIMES [Shavitt and Shir, 2005]. Using the graphs for March 2010 (the last ones available) and a simple best-response gradient heuristic from learning in games [Fudenberg

and Levine, 1999], we perform a series of experiments to both show the large-scale feasibility and scalability of the model and approach, and explore the behavior of the internal players and the attacker in the resulting equilibria and network-structure properties.

# References


D. Fudenberg and D. Levine. *The Theory of Learning in Games*. MIT Press, 1999.

Geoffrey Heal and Howard Kunreuther. IDS models of airline security. *Journal of Conflict Resolution*, 49(2):201–217, April 2005.

Geoffrey Heal and Howard Kunreuther. Modeling interdependent risks. *Risk Analysis*, 27:621–634, July 2007. doi: 10.1111/j.1539-6924.2007.00904.x. URL `http://dx.doi.org/10.1111/j.1539-6924.2007.00904.x`.

Mohammad T. Irfan and Luis E. Ortiz. A game-theoretic approach to inuence in networks. In *Proceedings of the Twenty-Fifth Conference on Artificial Intelligence (AAAI-11)*, August 2011. To appear.

M. Kearns, M. Littman, and S. Singh. Graphical models for game theory. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 253–260, 2001.

Michael Kearns. Economics, computer science, and policy. *Issues in Science and Technology*, Winter 2005.

Michael Kearns. Graphical games. In Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vaziran i, editors, *Algorithmic Game Theory*, chapter 7, pages 159–180. Cambridge University Press, 2007.

Michael Kearns and Luis E. Ortiz. Algorithms for interdependent security games. In *Neural Information Processing Systems (NIPS)*, 2003. URL `http://www.cs.sunysb.edu/~leortiz/papers/ids.pdf`.

Jon Kleinberg. Cascading behavior in networks: Algorithmic and economic issues. In Noam Nisan, Tim Roughgarden, Éva Tardos, and Vijay V. Vazirani, editors, *Algorithmic Game Theory*, chapter 24, pages 613–632. Cambridge University Press, 2007.

Howard Kunreuther and Geoffrey Heal. Interdependent security. *Journal of Risk and Uncertainty*, 26(2-3):231–249, March 2003. doi: 10.1023/A:1024119208153. URL http://www.springerlink.com/content/xr72631q57l8216q/fulltext.pdf.

Anahad O'Connor and Eric Schmitt. Terror attempt seen as man tries to ignite device on jet. The New York Times, 25 December 2009. Cited 31 August 2010. Avaliable from http://www.nytimes.com/2009/12/26/us/26plane.html.

Yuval Shavitt and Eran Shir. DIMES - Letting the internet measure itself. http://www.arxiv.org/abs/cs.NI/0506099. netdimes.org.

Yuval Shavitt and Eran Shir. DIMES: Let the Internet measure itself. *ACM SIGCOMM Computer Communication Review*, 35(5):71–74, October 2005.

Satinder P. Singh, Michael J. Kearns, and Yishay Mansour. Nash convergence of gradient dynamics in general-sum games. In *UAI*, pages 541–548, 2000.

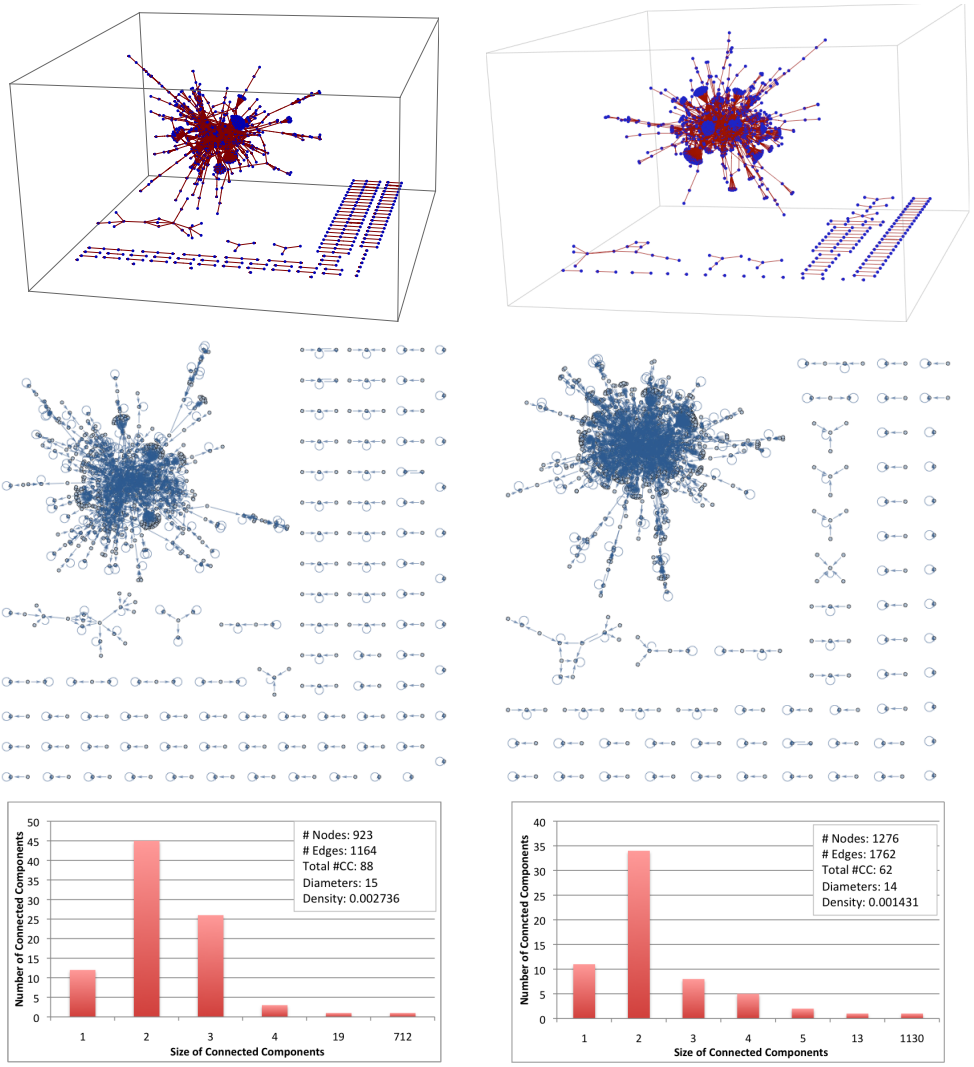


Figure 4: **The Structure of an Attack to the Internet.** The 3-d graphs (top) correspond to the top 360 Internet nodes most likely to be attacked (*left*) and to invest in defense/security measures (*right*), according to our model, and their neighbors (i.e., both parent and children family). The graphs in the middle are 2-d projections of the respective 3-d graphs on top. The self-loops mark the nodes that are actually attacked (left) and/or investing (right). Note that, for the most part, both graph structures have a very dense cluster, with the "highest defense/security investment" graph (right) being denser than the "most vulnerable" graph. This roughly suggests that, overall, the internal agents most likely to invest (and their neighbors) form a more tightly connected cluster within the network, than those more likely to be attacked. The bar graphs (bottom) correspond to the number of connected components of the top 360 Internet nodes most likely to be attacked (*left*) and to invest in defense/security measures (*right*)), according to our model. Some properties of the graph corresponding to the network structure are shown on the upper corner of the graphs.

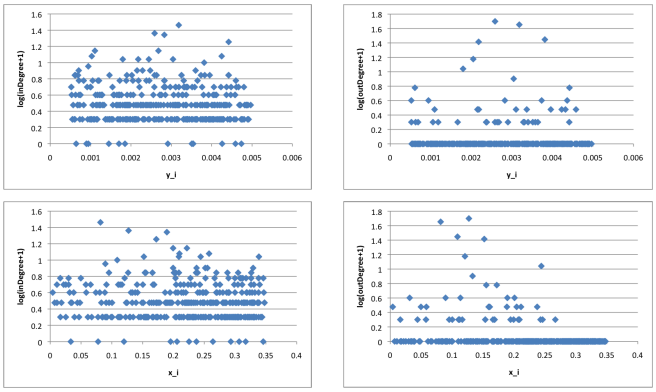


Figure 5: The two graphs on top show the corresponding $y_i$ (x-axis) and its in-degree and out-degree in log scale. Similarly, the two graphs at the bottom show the corresponding $x_i$ (x-axis) and its in-degree and out-degree in log scale.