Robust Solutions for Multi-Defender Stackelberg Security Games[*]

Dolev Mutzari , **Yonatan Aumann** , **Sarit Kraus**
Department of Computer Science, Bar Ilan University, Ramat Gan, Israel
dolevmu@gmail.com, aumann@cs.biu.ac.il, sarit@cs.biu.ac.il

Abstract

Multi-defender Stackelberg Security Games (MSSG) have recently gained increasing attention in the literature. However, the solutions offered to date are highly sensitive, wherein even small perturbations in the attacker's utility or slight uncertainties thereof can dramatically change the defenders' resulting payoffs and alter the equilibrium. In this paper, we introduce a robust model for MSSGs, which admits solutions that are resistant to small perturbations or uncertainties in the game's parameters. First, we formally define the notion of robustness, as well as the robust MSSG model. Then, for the non-cooperative setting, we prove the existence of a robust approximate equilibrium in any such game, and provide an efficient construction thereof. For the cooperative setting, we show that any such game admits a robust approximate α-core, provide an efficient construction thereof, and prove that stronger types of the core may be empty. Interestingly, the robust solutions can substantially increase the defenders' utilities over those of the non-robust ones.

1 Introduction

Stackelberg Security Games (SSGs) have attracted much attention in the multi-agent community [Paruchuri *et al.*, 2008b; Tambe, 2011; Nguyen *et al.*, 2013; An and Tambe, 2017]. They were applied to a variety of security problems, including the allocation of security resources at the Los Angeles International Airport [Pita *et al.*, 2008], and protecting biodiversity in conservation areas [Basak *et al.*, 2016] (see [Sinha *et al.*, 2018] for an overview). The original SSG model postulates a single defender facing a single attacker. In many applications, however, different targets may be valuable — to varying extents — to multiple disparate parties, each of which is interested in defending its targets of interest. Accordingly, the model of Multi-defender Stackelberg Security Games (MSSGs), wherein multiple defenders (leaders) protect a set of targets against a strategic attacker (follower) who is their common enemy, has recently gained increasing attention. Solution concepts for this game were developed over time, starting from ζ-Nash Equilibrium (NE) [Gan *et al.*, 2018], coordination mechanism [Gan *et al.*, 2020], coalition formation [Mutzari *et al.*, 2021] and correlated equilibrium (CE) via negotiation [Castiglioni *et al.*, 2021].

However, for reasons we will detail shortly, to date, the solutions offered for MSSGs are highly sensitive, wherein small perturbations in the attacker's utility, or slight uncertainties thereof, can dramatically alter the defenders' resulting payoffs and alter the equilibrium.

Tie-Breaking and Discontinuity. The key, and somewhat subtle, reason for the high sensitivity of the previous solutions has to do with the issue of tie-breaking. Typically, in SSGs, at equilibrium, the attacker has several equally attractive targets. Indeed, otherwise the defender(s) could shift resources from the overly protected targets to increase the protection of the attacker's chosen target. So, as ties are ubiquitous, the attacker's tie-breaking behaviour plays a crucial role in the analysis. Different works have considered different tie-breaking policies. In the case of a single defender, it is commonly assumed that the attacker breaks ties in favor of the defender, the reason being that the defender can essentially enforce his choice by shifting an arbitrarily small amount of resources away from his desirable target, absorbing only an arbitrarily small utility loss.

In the case of multiple heterogeneous defenders, however, optimistic tie-breaking is not well-defined (and not justified), as the best target for each defender may be different. Hence, [Gan *et al.*, 2018] use *pessimistic tie-breaking*, wherein each defender acts as if (among its choices) the attacker will attack the target that is *worst* for the defender. So, different defenders act as if the attacker, deterministically, attacks a different target. This tie-breaking rule, however, is not only overly pessimistic and unrealistic (as the attacker cannot simultaneously attack multiple targets), but also results in a sharp non-continuity in the defenders' payoffs. This is because the pessimistic tie-breaking occurs only at the point of *exactly identical* utilities (of the attacker) for the disparate targets. If the attacker's utility is slightly perturbed (or if there is even a slight uncertainty with regards to the attacker's true utilities), then the pessimistic tie-breaking is no longer relevant and the defenders' payoff may change dramatically. This discontinu-

[*]This research has been partly supported by the Israel Science Foundation under grant 1958/20 and the EU Project TAILOR under grant 992215.

ity in payoff, in turn, shapes the structure of the solution concepts, which also tend to exhibit a discontinuity at the exact equilibrium point. Indeed, this sensitivity, and discontinuity, is the core reason why, in the existing MSSG models, exact-NE need not exist (see [Mutzari *et al.*, 2022] for an example).

In practice, however, such sharp discontinuity is often hard to justify. Indeed, while it is standard in game theory to assume that players' utilities are common knowledge, it is unrealistic to assume that these utilities are known accurately to any level of precision. Evidently, the defenders must somehow *infer* the attacker's utilities using some combination of data and reasoning, but this data (and reasoning) may be incomplete, imprecise, or noisy. Additionally, the attacker itself may have bounded computational resources, or be subject to other forms of noise, and thus not operate exactly as dictated by infinite precision calculations. Finally, the attacker's strategy is also determined by its knowledge/belief of the defender's coverage strategies, which the attacker needs to somehow infer from possibly incomplete and noisy data.

Robustness in MSSGs. Accordingly, as detailed in the related work, in the *single* defender setting, a whole line of research is devoted to addressing uncertainties (see [Nguyen *et al.*, 2014]). Most of these papers tackle uncertainty by applying robust optimization techniques ([Pita *et al.*, 2009; Jiang *et al.*, 2013; Qian *et al.*, 2015; Nguyen *et al.*, 2014]). Some works consider a Bayesian approach [Kiekintveld *et al.*, 2011; Yin and Tambe, 2012; Yang *et al.*, 2012; Amin *et al.*, 2016]. We believe, however, that this avenue is ill-suited for the multi-defender setting, with its tie-breaking subtlety. Robust optimization takes a *worst-case* approach: if some parameters are only known to lay within some set, then one assumes — pessimistically — that these parameters obtain values that minimize the objective function. In the multi-defender setting, however, there is no *one* worst case; the worst-case for one defender may well be good for another. So, worst-case robustness is not well defined in this setting. We note that one could possibly propose a multi-player worst case analysis wherein each defender acts as if the parameters are worst for itself. But, coupled with pessimistic tie-breaking, this would result in an unreasonably, doubly pessimistic and unrealistic perspective, wherein different defenders postulate attackers with different parameters, attacking different targets; as if they inhabit parallel universes. The multi-defender setting therefore calls for a different approach.

Our Contribution. In this paper, we offer a robust model for the analysis of MSSGs. First, we formally define the notion of a *robust solution*, which formalizes the notion that the game solutions (e.g. Nash equilibrium, core) remain valid even after small perturbations or uncertainties in the game's parameters. We then introduce a formal model wherein small perturbations in the attacker's utility result in only small changes in the attacker's expected behavior, and hence also in that of the defenders. Essentially, we model the attacker's behavior as being probabilistic, with a continuous distribution concentrated around the behavior dictated by the presumed (and possibly inaccurate) utility. Importantly, we do not suppose any specific form for this distribution, only that it is concentrated as stated. Thus, this one model captures and unifies many possible scenarios and sources of noise and uncertainty.

Once we have formally defined the notions of robustness and the robust MSSG model, we show that the robust model indeed offers robust solutions. For the non-cooperative setting, we provide an efficient algorithm for constructing a robust approximate NE. For the cooperative setting, we consider the core of the game, for different variants of the core (α, γ). For the α version, we prove that the robust approximate core is always non-empty, and give an efficient algorithm for finding solutions therein. Importantly, since the utility of the defenders is no longer determined by their pessimistic beliefs, the resulting core solutions allow greater utility for the defenders than in previous models. Finally, we show that the γ-core may be empty.

We note that an additional benefit of the model, besides robustness, is that it renders moot the entire issue of tie-breaking — optimistic or pessimistic. With a continuous probability distribution, there is zero probability for ties.

Full version. Due to space limitations, this version of the paper focuses on introducing the key concepts and proof ideas. Complete proofs can be found at [Mutzari *et al.*, 2022].

Selected Related Work. Robust analysis of the single-defender case has attracted considerable attention. Many of these employ robust optimization techniques to address uncertainties, while others model uncertainties using Bayesian models. Robust optimization is employed by: [Kiekintveld *et al.*, 2013] in a setting where the attacker's utilities are known to lay in some interval, but the exact value is not known; and by [Pita *et al.*, 2009] is a settings wherein the attacker may exhibit bounded rationality; by [Jiang *et al.*, 2013] in a setting wherein the attacker's type is only known to be monotone. [Qian *et al.*, 2015] consider a worst-case approach for studying risk-averse attackers, for which the level of risk-adverseness is unknown. Finally, [Nguyen *et al.*, 2014] offer a unified framework and methods for simultaneously addressing multiple types of uncertainties in single defender SSGs using robust optimization. As detailed in the introduction, we argue that the worst-case approach of robust optimization is ill-suited for the multi-defender setting, wherein there is no one worst-case for all defenders.

Other works, still in the single defender setting, employ a Bayesian approach to address uncertainties (as we do for the multi-defender setting). A general analysis of a single defender Bayesian SSG was introduced in [Paruchuri *et al.*, 2008a]. [Kiekintveld *et al.*, 2011] study Bayesian SSGs with a continuous payoff distribution for the attacker. They demonstrate scenarios where there are too many possible attacker types to consider, and known solutions for a finite number of attacker types don't scale well. [Yin and Tambe, 2012] provide a unified Bayesian approach to handling both discrete and continuous uncertainty in attacker types. [Yang *et al.*, 2012] model human adversaries with bounded rationality as quantal response adversaries, which with some probability do not choose the best response. They specifically assume that the attacker's strategy is determined by a scaled *soft-max* function of the expected utilities of the attacker on each target, and utilize the specific structure of the *soft-max* function to obtain a solution for the setting (see also [Amin *et al.*, 2016], who consider general SGs).

Our work differs from all of the above in that it considers the multi-defender case. The Bayesian modeling we introduce is also different. On the one hand, unlike the unrestricted, general Bayesian models (e.g. [Paruchuri *et al.*, 2008a; Kiekintveld *et al.*, 2011; Yin and Tambe, 2012], [Paruchuri *et al.*, 2008a; Yin and Tambe, 2012]), we assume a concentrated distribution (Definition 3.2). This, we believe, adequately models most frequent sources of uncertainty, including noisy and imprecise information, bounded computation power, and most incarnations of bounded rationality. On the other hand, we do not suppose any specific form for this concentrated distribution (e.g. *soft-max*), as such an assumption would fail to model many real-world uncertainties.

2 Defining Robustness

The main objective of this work is to develop a robust MSSG framework. We now formally define the notion of robustness. We deliberately do so in the most general terms, not confining the definition to the specific game, or solution concepts that we consider. We later show how the model we offer indeed exhibits robustness as defined here.

Solution Concepts. Consider the Nash equilibrium solution concept. Technically, given a game G, there is a set $\mathrm{NE}(G)$ of strategy profiles for which the NE property holds. Similarly, the core of G is a set of coalition structures *core*(G), for which the core properties hold. Thus, in the most general sense, a *solution concept* is a function $\mathcal{X}$ from the set of games (of some class) to some space S, which maps each game G (of the class) to the corresponding structures for which the specific property of the solution concept holds.

Nearness. Conceptually, robustness of a solution $\mathbf{x}$ states that $\mathbf{x}$ remains a solution even under "small" perturbations in game parameters. This requires a notion of *distance* amongst games. Given a distance function $d(\cdot,\cdot)$ over game pairs, say that games $G, \hat{G}$ are η*-near* if $d(G,\hat{G}) \leq \eta$ (later, d will be instantiated based on the MSSG's specific parameters).

Definition 2.1 (Robust Solution). *Let $\mathcal{X}$ be a solution concept, G a game, and $x \in \mathcal{X}(G)$. We say that x is an* η-robust $\mathcal{X}$ of G, *if $x \in \mathcal{X}(\hat{\mathcal{G}})$ for any $\hat{G}$ that is η-near to G.*

Thus, for example, a strategy profile $\mathbf{x}$ is said to be an η-robust ϵ-NE of G if it is an ϵ-NE of any $\hat{G}$ that is η-near to G (G itself included).

3 The Robust MSSG Model

The Standard model. In an MSSG, there is a set $\mathcal{N} = \{1,\ldots,n\}$ of *defenders*, a set $\mathcal{T} = \{t_1,\ldots,t_m\}$ of *targets* that the defenders wish to protect, and an attacker, who seeks to attack the targets. Each defender $i \in \mathcal{N}$ has $k_i \in \mathbb{N}$ *security resources*, each of which can be allocated to *protect* a target. The attacker chooses one target to attack. The attack is *successful* if the target is unprotected by any security resource. A successful (res. unsuccessful) attack at t yields utility $r^a(t)$ (res. $p^a(t)$) to the attacker and $p^d_i(t)$ (res. $r^d_i(t)$) to defender i, for all i ($r^d_i(t) > p^d_i(t)$, $r^a(t) > p^a(t)$).

An MSSG is thus a 5-tuple $G = (\mathcal{N},\mathcal{T},\mathcal{K},\mathcal{R},\mathcal{P})$, where $\mathcal{K} = (k_1,\ldots,k_n)$ are the numbers of security resources of the defenders, $\mathcal{R} = (r^a(t_1),\ldots,r^a(t_m), r^d_1(t_1),\ldots,r^d_n(t_m))$, is the sequence of rewards, and similarly $\mathcal{P}$ - the sequence of penalties.

The defenders' allocation of security resources to the targets may be randomized. Specifically, each defender i chooses a *coverage vector* $\mathbf{x}_i \in \mathcal{C}_{k_i}$ (where $\mathcal{C}_k = \{x \in [0,1]^m | \sum x_t \leq k\}$). Here, $x_{i,t}$ is the probability that target t is protected by one of defender i's security resources.[1] We denote by $\mathbf{X} = (\mathbf{x}_i)_{i\in\mathcal{N}}$.

If the defenders are uncoordinated, the probability that target $t \in \mathcal{T}$ is protected is

$$c_t = \mathsf{cov}_t(X) = 1 - \textstyle\prod_{i\in\mathcal{N}}(1 - x_{i,t}).$$

The *overall coverage vector* $\mathbf{c} = (c_t)_{t\in\mathcal{T}}$ is assumed to be known to the attacker, which chooses its action after the defenders have committed to their distribution. Thus, the attacker and defenders' utilities upon an attack at t are:

$$U^a(\mathbf{c},t) = U^a(c_t,t) = (1-c_t)\cdot r^a(t) + c_t \cdot p^a(t) \quad (1)$$

$$U^d_i(\mathbf{c},t) = U^d_i(c_t,t) = c_t \cdot r^d_i(t) + (1-c_t)\cdot p^d_i(t) \quad (2)$$

In the classic (non-robust) model, all players are assumed to be rational. As such, the attacker's best response is to attack a target in the set $\mathsf{BR}(\mathbf{c}) := \arg\max_{t\in\mathcal{T}} U^a(\mathbf{c},t)$, which maximizes its expected utility. However, as discussed in the introduction, $\mathsf{BR}(\mathbf{c})$ typically consists of multiple targets, which brings about the issues of tie-breaking and discontinuity.

The Robust Model. The non-robustness of the standard MSSG model arises from the assumption of exact deterministic behavior of the attacker, whereby it always plays the *exact* optimal play, even if the difference between the optimal play and the next in line is minuscule. Therefore, small changes in the attacker's utility function may lead to abrupt changes in the attacker's strategy, in turn causing abrupt changes in the defender's utilities. Accordingly, to obtain a robust model, we model the attacker's behavior as being *probabilistic*, with a continuous distribution concentrated around the deterministic behavior dictated by the presumed (possibly inaccurate) utility. Importantly, we seek a general model which can capture the multitude of possible reasons for the mentioned perturbations. The formal details follow.

Given a coverage vector $\mathbf{c}$, and the resultant attacker's utility vector $\mathbf{u} = \mathbf{u}(\mathbf{c}) = (U^a(\mathbf{c},t))_{t\in\mathcal{T}}$ (over the different targets), the attacker's actual behavior is assumed to be determined by a probability distribution $\omega(\mathbf{u}) \in [0,1]^m$, specifying the probability that the attacker will actually attack each target. The function ω is termed the *Attacker's Behavior Function*, and is assumed to have the following properties:

Definition 3.1 (Attacker's Behaviour Function (ABF)). *A continuously differentiable function $\omega : \mathbb{R}^m_+ \to \mathbb{R}^m_+$ is an* attacker's behaviour function *if the following axioms hold:*

1. *$\omega(\mathbf{u})$ is a probability distribution.*

[1] Provably, any such coverage vector can be implemented by a distribution over deterministic allocation strategies, each employing at most k resources.

2. *ω is monotone increasing at each coordinate.*
3. *For each $\emptyset \neq S \subseteq \mathcal{T}$ and $t \in S$: $\frac{\omega_t(\mathbf{u})}{\sum_{t' \in S} \omega_{t'}(\mathbf{u})}$ is independent of any $u_{t'}$, $t' \notin S$.*

Axiom 3 states that, given that the attack is within the set S, the conditional probability of an attack on any specific target $t \in S$ is only determined by the inter-relationships between the utilities of targets within S.

Thus, given a coverage vector $\mathbf{c}$, inducing attacker utilities $\mathbf{u}(\mathbf{c})$, and ABF ω, defender i's expected utility is given by:

$$U_i^d(\mathbf{c}) = \sum\nolimits_{t \in \mathcal{T}} U_i^d(\mathbf{c}, t) \cdot \omega_t(\mathbf{u}(\mathbf{c})) \qquad (3)$$

The following definition provides that the attacker's behavior is centered around the optimal deterministic one.

Definition 3.2 (**(δ, ϵ)-ABF**). *Let G be an (M)SSG, and $\delta, \epsilon > 0$. An ABF ω is a (δ, ϵ)-ABF if:*

$$u_{t'} < u_t - \delta \Rightarrow \omega_{t'}(\mathbf{u}) < \epsilon \qquad (4)$$

Thus, with a (δ, ϵ)-ABF, any target offering a utility δ less than the optimal one will be attacked with probability $< \epsilon$.

We believe this (δ, ϵ)-ABF formulation captures, under one unifying definition, the many possible sources for noise and uncertainty in the attacker's utility, as discussed in the introduction. With these definitions, a *robust MSSG* is a pair $\mathcal{G} = \langle G, \omega \rangle$, where G is a MSSG and ω is a (δ, ϵ)-ABF.[2]

4 Solution Concepts

For completeness, we review the MSSG solution concepts, as defined in previous works. Throughout, we consider the approximate versions.

ζ-Nash Equilibrium. Given the players' utility functions, the definitions of approximate NE is standard:

Definition 4.1 (**ζ-Nash Equilibrium**). *A strategy profile $\mathbf{X} = (\mathbf{x}_i)_{i \in \mathcal{N}}$ is an ζ-NE if for any defender $i \in \mathcal{N}$ and any strategy $\mathbf{x}_i'$ of i:*

$$U_i^d(\mathsf{cov}(X)) \geq U_i^d(\mathsf{cov}(\langle \mathbf{x}_i', X_{-i} \rangle)) - \zeta. \qquad (5)$$

Coalitions. When the defenders are uncoordinated, independent choices may result in inefficient resource use. Therefore, [Mutzari *et al.*, 2021] consider a model for coalition formation in MSSG, which we adopt. Any subset $P \subseteq \mathcal{N}$ of defenders may form a *coalition*, in which case they act as a single defender with $k_P = \sum_{i \in P} k_i$ resources. The coalition consisting of all of the defenders is called the *grand coalition*.

Coalitions partition the set of defenders: $\mathcal{N} = \{P_1, \ldots, P_\ell\}$. Each coalition P_i chooses a coverage vector $\mathbf{x}_j \in \mathcal{C}_{k_{P_j}}\}$. Denoting $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_\ell)$, the probability that target t is protected is:

$$c_t = \mathsf{cov}_t(\mathbf{X}) := 1 - \prod\nolimits_{i=1}^{\ell} (1 - x_{i,t}). \qquad (6)$$

Given a strategy profile $\mathbf{X}$ and a target t, the defenders' utilities are defined in the same way as in (3). A *coalition structure* is any such pair $\mathcal{CS} = \langle \mathcal{P}, \mathbf{X} \rangle$.

Since $\mathcal{CS}$ includes a coverage vector $\mathbf{X}$ that (by (6)) induces the overall coverage vector $\mathsf{cov}(X)$, for notational simplicity, the players' utilities can be viewed as a function of the coalition structure, which we simply denote $U_i^d(\mathcal{CS})$.

Deviations. Given a coalition structure $\mathcal{CS}$, a subset of defenders may choose to deviate and form a new coalition. In such a case, one must define the strategy played by all defenders. Following [Moulin and Peleg, 1982; Chalkiadakis *et al.*, 2011; Shapley, 1973], the assumption is that in order to protect the status quo, the remaining defenders take revenge against the deviators. The revenge is *ζ-successful* if it results in at least one deviator gaining no more than ζ from the deviation. It is assumed that this revenge strategy is chosen *after* the deviators choose their strategy. A deviation is *ζ-approximate successful* if it admits no ζ-successful revenge.

Definition 4.2 (**ζ-approximate core**). *A coalition structure $\mathcal{CS}$ is in the ζ*-approximate α-core *if it admits no ζ-approximate successful deviation.*

So, in the ζ-approximate core, a deviation cannot add more than ζ to the utility of at least one of the deviation's members.

5 The Robustness Theorem

We now provide the central robustness theorem, which essentially states that in the robust MSSG model, any approximate-NE or approximate-core automatically translates to their respective *robust* counterparts. We note that, technically, the theorem essentially follows directly from the continuity of ω, but the result is exactly what is necessary.

First, recall that the definition of a robust solution (Definition 2.1) requires a distance function between games. In MSSG, for $G = (\mathcal{N}, \mathcal{T}, \mathcal{K}, \mathcal{R}, \mathcal{P}), \hat{G} = (\mathcal{N}, \mathcal{T}, \mathcal{K}, \hat{\mathcal{R}}, \hat{\mathcal{P}})$, and $\mathcal{G} = (G, \omega), \hat{\mathcal{G}} = (\hat{G}, \hat{\omega})$, we define:

$$d(\mathcal{G}, \hat{\mathcal{G}}) = \quad \max\{\|\mathcal{R} - \hat{\mathcal{R}}\|_\infty, \|\mathcal{P} - \hat{\mathcal{P}}\|_\infty, \|\omega - \hat{\omega}\|_\infty\}$$

The following lemma bounds the change in utility arising from small changes in the game parameters.

Lemma 5.1. *Let $\mathcal{G}$ be a robust MSSG and $\eta > 0$. Then there exists a constant b such that, for any coverage vector $\mathbf{c}$, any defender $i \in \mathcal{N}$ and any game $\hat{\mathcal{G}}$ η-near to $\mathcal{G}$: $|\hat{U}_i^d(\mathbf{c}) - U_i^d(\mathbf{c})| \leq 0.5b\eta$ (where $\hat{U}_i^d(\mathbf{c})$, $U_i^d(\mathbf{c})$ are the utilities of defender i in games $\hat{\mathcal{G}}$, $\mathcal{G}$ respectively).*

The proof essentially follows from the continuity of the ABF (see full version, [Mutzari *et al.*, 2022], for details).

Theorem 5.2. *Let $\mathcal{G}$ be a MSSG. With b of Lemma 5.1, for any η:*

1. *any ζ-NE of $\mathcal{G}$ is also an η-robust $(\zeta + b\eta)$-NE of $\mathcal{G}$.*
2. *any ζ core of $\mathcal{G}$ is also an η-robust $(\zeta + b\eta)$-core of $\mathcal{G}$ (for all three variants of the core α, β, γ).*

Proof. We prove (1). The proof of (2) is analogous. Let $\mathbf{X}$ be a ζ-NE of $\mathcal{G}$, and $\hat{\mathcal{G}}$ η-near to $\mathcal{G}$. Then, for any possible $\mathbf{x}_i'$

[2] We note that inevitably the proper values for δ, ϵ are dependent on G. The reason is that δ must be small compared to the attacker's utilities, thus must be scaled in accordance, and ϵ must depend on the number of targets as it has to be small relative to $1/m$.

Algorithm 1 ALLOC

Input: value $c, \tilde{c} \in [0,1]$, $G = (\mathcal{N}, \mathcal{T}, \mathcal{K}, \mathcal{P}, \mathcal{R})$
Output: a strategy profile $\mathbf{X} = (x_{i,t})$

```
1: $c_t \leftarrow 0$, for all $t$
2: for $i \in \mathcal{N}$ do
3:     for $t \in \mathcal{T}$ by reverse $\preceq_i^{\tilde{c}}$ precedence order do
4:         $x_{i,t} \leftarrow \min\{1 - \frac{1-c}{1-c_t}, k_i\}$
5:         $k_i \leftarrow k_i - x_{i,t}$
6:         $c_t \leftarrow 1 - (1 - c_t)(1 - x_{i,t})$;
7:     end for
8: end for
```

of player i

$$\hat{U}_i^d(\mathsf{cov}(\langle \mathbf{x}'_i, \mathbf{X}_{-i} \rangle)) \geq U_i^d(\mathsf{cov}(\langle \mathbf{x}'_i, \mathbf{X}_{-i} \rangle)) - b\eta/2 \tag{7}$$

$$\geq U_i^d(\mathsf{cov}(\langle \mathbf{x}_i, \mathbf{X}_{-i} \rangle)) \text{ - } \zeta \text{ - } b\eta/2 \tag{8}$$

$$\geq \hat{U}_i^d(\mathsf{cov}(\langle \mathbf{x}_i, \mathbf{X}_{-i} \rangle)) - \zeta - b\eta \tag{9}$$

where we use Lemma 5.1 for (7) and (9), and (8) follows by the definition of ζ-NE. □

Accordingly, in order to find robust solutions in our model it suffices to find regular solutions in the model, and robustness then follows automatically. The remainder of the paper is thus dedicated to finding such solutions.

6 Nash Equilibrium in Robust MSSGs

We now show how to compute an approximate NE in robust MSSGs. For simplicity, we restrict the discussion to *non-saturated* games. A robust MSSG $\mathcal{G}$ is *α-saturated* if there exists a coverage $\mathbf{c} \in \mathcal{C}_{k_\mathcal{N}}$ and $t \in \mathcal{T}$ s.t. $c_t \geq 1 - \alpha$ and $\omega_t(\mathbf{c}) \geq \epsilon$. In a saturated game, there are sufficient resources to induce an attack that is almost surely caught. The results can be extended to saturated games using a similar approach to the one used by [Gan *et al.*, 2018].

Theorem 6.1. *There exists a polynomial algorithm, such that for any G there exist $A, B, C, \epsilon_0, \delta_0$ s.t. for any (δ, ϵ)-ABF ω with $\epsilon < \epsilon_0, \delta < \delta_0$, on input (G, ω) the algorithm outputs a strategy profile $\mathbf{X}$ that is a ζ-NE, for $\zeta = B\delta + C\epsilon$ (provided (G, ω) is not $A\delta$-saturated).*

The exact constants, together with a detailed proof, appear in the full version, [Mutzari *et al.*, 2022]. The proof builds upon the techniques of [Gan *et al.*, 2018], but requires significant adaptations for our setting. For simplicity, here, we consider the case where all targets carry identical penalties and identical rewards for the attacker (but not the defenders). This case allows us to explain the core elements of the construction and the proof, while avoiding the technicalities. Note that in the explanation we do not seek to obtain the best constants. Better bounds are offered in the complete proof.

For the case we are considering, we can further normalize the attacker's utilities so that $p^a(t) = 0$ and $r^a(t) = 1$ for all t. With this assumption, $U^a(c, t) = 1 - c_t$, for all c, t.

The core procedure underlying the algorithm is ALLOC (Algorithm 1). ALLOC gets as input a parameter c, and aims to have $c_t = c$, for all targets. To this end, the defenders — one by one in order — iterate through the targets, allocating resources until either (i) the c level is reached, or (ii) their resources are fully depleted. This allocation is performed in Line 4, where $x_{i,t} = 1 - \frac{1-c}{1-c_t}$ brings the coverage to exactly c, and $x_{i,t} = k_i$ depletes the defender's resources. Importantly, each defender considers the targets in *reverse preference order* of its expected utility from the target, assuming some identical coverage level $\tilde{c}$ (with $\tilde{c}$ possibly different from c) . Specifically, for targets t, t', we denote $t \preceq_i^{\tilde{c}} t'$ if $U_i^d(\tilde{c}, t) \leq U_i^d(\tilde{c}, t')$. In ALLOC, defenders iterate through the targets from the least to the highest in the $\preceq_i^{\tilde{c}}$ order.

Clearly, the actual coverage level obtained by ALLOC is determined by c: if c is too small then ALLOC may complete without depleting the defenders' resources, and if c is too big then some targets my have $c_t < c$. However, there exists a $\bar{c}$ for which the algorithm completes with $c_t = \bar{c}$ for all targets (assuming the game is non-saturated). This $\bar{c}$ can be approximated to within any level of accuracy by binary search, with multiple runs of ALLOC. Set $\beta = (1 - \bar{c})/2$.

Set $\check{c} = \bar{c} + m\delta/\beta$. For δ sufficiently small, $\check{c} < 1 - \beta$. Consider the outcome of running ALLOC with $c = \check{c}$ and $\tilde{c} = \bar{c}$. Let $\hat{\mathbf{X}}$ be the resultant allocation, and for each t, let $\hat{c}_t$ be the resultant coverage of target t. Since $\check{c} > \bar{c}$, there will necessarily be targets t for which $\hat{c}_t < \bar{c}$, but for δ sufficiently small there will be only one such target. Denote this target t^*. For all other t's, $\hat{c}_t = \check{c}$.

We now argue that $\hat{\mathbf{X}}$ is a ζ-NE for $\zeta = B\delta + C\epsilon$, for some constants B, C dependant only on G.

In the following, we focus on the attacker's behavior occurring with probability $\geq \epsilon$. For ease of exposition, we say that an event is *likely* if it happens with probability $\geq \epsilon$.

If all defenders play by $\hat{\mathbf{X}}$ then an attack is only likely at t^*. This is since $\hat{c}_t - \hat{c}_{t^*} \geq \delta$, and hence $\hat{u}_{t^*} - \hat{u}_t \geq \delta$, for all t, and ω is a (δ, ϵ)-ABF. Consider a deviation of defender i, and let $\mathbf{c}'$ be the coverage vector resulted from this deviation. We now explain why $\mathbf{c}'$ cannot "substantially" increase the utility of i. Let: L be the targets - aside from t^* to which i allocated resources, $L^+ = L \cup \{t^*\}$, and $H = \mathcal{T} \setminus L^+$. We consider what $\mathbf{c}'$ can do to attacks on H, L and t^*.

Attacks on H. We argue that under $\mathbf{c}'$ the attacker is unlikely to attack any target of H. The reason is that since i did not allocate resources to H, the only way that it can induce an attack on H is by increasing the coverage of all targets in L^+ to $\check{c} + \delta$. But, since $c_{t^*} < \bar{c} = \check{c} - m\delta/\beta$, bringing t^*'s coverage to $\check{c} + \delta$ can only be accomplished by taking coverage from the members of L. In doing so, at least one will result with coverage less than $\check{c} - \delta$ (the value of $\check{c}$ was so chosen).

Attacks on L. A deviation *can* induce attacks on L. However, it cannot provide i substantially more than it originally obtained, where *substantially* means adding more than δB utility. The reason is that since t^* is not $\check{c}$ covered, it must be that t^* was considered by i after all elements of L - or else i would either bring t^*'s coverage to $\check{c}$ or fully deplete its resources. So, $t \preceq_i^{\bar{c}} t^*$, for all $t \in L$. This means that - at coverage level $\bar{c}$ - attacks on targets of L offer i no more utility than attacks on t^*. So, if $\bar{c}$ where the coverage, then inducing an attack on L would offer no gain to i. In practice, the elements of $\mathbf{c}'$ and $\hat{\mathbf{c}}$ are not exactly $\bar{c}$, but they are $O(\delta)$ away. So, since $U_i^d(c,t)$ is linear in c, the differences in utility can also be only $O(\delta)$. So attacks on L cannot offer *substantially* more utility than what t^* initially offered.

Attack on t^*. A deviation can also possibly increase the coverage of t^*, thus offering more utility if and when attacked. However, one can only add $O(\delta)$ coverage (while keeping an attack on t^* likely), so that this addition cannot add more than $O(\delta)$ utility.

Putting It All Together. We obtain that with probability $1-\epsilon$ the added utility due to the deviation is at most δB, for some constant B. With probability ϵ the increase can be larger, but clearly bounded by $C = \max_{j,t} r_j^d(t)$. So, the utility increase is bounded by $\zeta = \delta B + \epsilon C$. This completes the construction of ζ-NE. By Theorem 5.2, this strategy is also an η-robust $(\zeta + b\eta)$-NE, for any $\eta > 0$.

7 The Robust Core

In this section we outline the algorithm to construct a robust ζ-approximate α-core, and correctness proof. The full details appear in the paper's full version [Mutzari *et al.*, 2022].

Resistance to Subset Deviations. With minor adaptations, the ALLOC procedure can be applied to the cooperative setting (wherein probabilities are additive rather than multiplicative). Also, the procedure can easily be configured to accept a target utility level u as input (rather than target coverage c) - see full version [Mutzari *et al.*, 2022]. As before, repeated calls to ALLOC allow us to find a $\bar{u}$ and strategy $\bar{\mathbf{X}}$ such that all targets offer utility $\bar{u}$ to the attacker (except for those with $r^a(t) < \bar{u}$), using all resources. Next, we re-run ALLOC with $\breve{u} = \bar{u} - mO(\delta)$. Let $\hat{\mathbf{X}}$ be the resulted strategy profile, $\hat{\mathbf{c}} = \mathsf{cov}(\hat{\mathbf{X}})$, and t^* the target not covered to $\breve{u}$. Then, $U^a(\hat{\mathbf{c}}, t^*) > \bar{u} + \delta$ and $U^a(\hat{\mathbf{c}}, t) = \breve{u}$ for all other t. So, the attacker is only likely to attack t^*.

We argue that $\hat{\mathbf{X}}$ is resilient to any deviation $\mathbf{x}_D$ of any proper subset $D \subset \mathcal{N}$ of defenders. Set $\mathbf{c}' = \mathsf{cov}(\hat{\mathbf{X}}_{-D}, \mathbf{x}_D)$. Let L be the targets, aside from t^*, to which D allocated resources in $\hat{\mathbf{X}}$, and $L^+ = L \cup \{t^*\}$. Then, analogously to the NE case, no deviation of D can induce a likely attack outside $L \cup \{t^*\}$. They may, however, be able to alter the attack probabilities within L^+. Let $\mathcal{A}$ be the targets that are likely to be attacked under $\mathbf{x}_D$. So, $|U^a(\mathbf{c}', t_i) - U^a(\mathbf{c}', t_j)| < \delta$, for any $t_i, t_j \in \mathcal{A}$. Now, recall that given D and $\mathbf{x}_D$, the remaining defenders - $\bar{D}$ - can change their strategy. So, using only $O(\delta)$ resources per target, $\bar{D}$ can raise the coverage of all but one target of $t_0 \in \mathcal{A}$, so that t_0 offers at least δ more utility than *all* other targets. Since $t_0 \in L^+$, there exists at least one $i \in D$ for which $t_0 \preceq_i^{\mathsf{cov}(\mathbf{X})} t^*$. So, for this i, the deviation does not offer any (substantial) gain.

Resistance to Grand Coalition Deviations. We now show how to transform the above strategy to one resistant to grand coalition deviations. Let $\hat{u}_i^d$ be the utility of defender i under $\hat{\mathbf{X}}$. Consider the following linear program (with variables p_t):

$$\begin{aligned} &\max \textstyle\sum_{t\in\mathcal{T}} \sum_{i\in\mathcal{N}} p_t \cdot U_i^d(\hat{\mathbf{c}}, t) && && (10)\\ &s.t. \textstyle\sum_{t\in\mathcal{T}} p_t \cdot U_i^d(\hat{\mathbf{c}}, t) \ge \hat{u}_i^d, && i = 1, \ldots, n \\ &\textstyle\sum_{t\in\mathcal{T}} p_t = 1; p_t \ge 0, && t = 1, \ldots, m \end{aligned}$$

Let $\mathbf{p}^* = (p_t^*)_{t=1}^m$ be a solution to LP 10 (there necessarily exists a solution, since $\mathbf{p} = \omega(\mathbf{u}(\hat{\mathbf{c}}))$ is a feasible one). Proposition 7.1 provides that with small changes to $\bar{\mathbf{u}} = (\bar{u}, \ldots, \bar{u})$ the distribution $\mathbf{p}^*$ can (essentially) be induced.

Proposition 7.1. *For any probability distribution* $\mathbf{p}$ *and any* $u > 0$*, there exists a utility vector* $\mathbf{u}$ *s.t.* $\|\omega(\mathbf{u}) - \mathbf{p}\|_\infty < m^2\epsilon$ *, and* $u \le u_t < u + 2m\delta$ *for all* $t \in \mathcal{T}$*.*

Let $\bar{\mathbf{u}}^*$ be the utility vector provided by invoking Proposition 7.1 with $u = \bar{u}$ and $\mathbf{p} = \mathbf{p}^*$. Since $\bar{\mathbf{u}}$ is implementable (by $\bar{\mathbf{X}}$) and $\bar{u}_t^* \ge \bar{u}$ for all t, then $\bar{\mathbf{u}}^*$ is also implementable, by some strategy $\bar{\mathbf{X}}^*$. We now argue that $\bar{\mathbf{X}}^*$ is an the ζ-approximate α-core, for $\zeta = A\epsilon + B\delta$ (for some constants A, B fully determined in the full proof). In the following arguments, interpret all statements "up to at most ζ gains".

To see that $\bar{\mathbf{X}}^*$ resists grand coalition deviations, first note that we may assume that in any such deviation $\mathbf{X}'$, all elements of $\mathbf{u}' = \mathbf{u}(\mathsf{cov}(\mathbf{X}'))$ are close (within $O(\delta)$) to $\bar{u}$. Otherwise, there must be targets that are more than δ apart in their utility, in which case it is possible to shift resources from the those with lesser u_t' to the higher ones, without substantially changing the attack distribution (by Proposition 7.1). Now, any target where $u_t' > \bar{u}$ is not substantially better for any deviator, and any target where $u_t' < \bar{u} - \delta$ is not likely to be attacked. By construction, $\mathbf{p}^*$ maximizes the sum of defender utilities, subject to each getting at least as in $\hat{\mathbf{X}}$. So, it is impossible that *all* defenders simultaneously get even more.

For subset deviations, consider $D \subset \mathcal{N}$ and suppose they have a deviation $\mathbf{x}_D$ that guarantees all members of D more than in $\bar{\mathbf{X}}^*$ (regardless of how the others play). But $\bar{\mathbf{X}}^*$ offers essentially as much as the output of the linear program, which, by its constraints, offers each defender as much as $\hat{\mathbf{X}}$. So, $\mathbf{x}_D$ would also constitute a successful deviation from $\hat{\mathbf{X}}$, which we proved cannot be. We thus obtain:

Theorem 7.2. *There exists a polynomial algorithm, such that for any* G *there exist* $A, B, C, \epsilon_0, \delta_0$ *s.t. for any* (δ, ϵ)*-ABF* ω *with* $\epsilon < \epsilon_0, \delta < \delta_0$*, on input* (G, ω) *the algorithm outputs a strategy profile* $\mathbf{X}$ *that is a* ζ*-approximate* α*-core, for* $\zeta = B\delta + C\epsilon$ *(provided that* (G, ω) *is not* $A\delta$*-saturated).*

By Theorem 5.2, for any $\eta > 0$, any such solution is also an η-robust $(\zeta + b\eta)$-approximate α-core.

γ-Core. Unlike the α-core, the approximate γ-core (see [Chander, 2010; Mutzari *et al.*, 2022]) may be empty, as in the following example. There are 6 targets and 4 defenders. Targets 1-3 yield a reward of 6 and penalty of 5 to defenders 1 and 2, as do targets 4-6 to defenders 3 and 4. All other targets yield a reward 1 and penalty 0 to all other players (including the attacker). Then defenders 1+2 can γ-deviate and get a utility ≥ 4 each, as can defenders 3+4, but no coalition structure can yield utility ≥ 4 to all defenders. See a full explanation in [Mutzari *et al.*, 2022].

References

[Amin *et al.*, 2016] K. Amin, S. Singh, and M. P Wellman. Gradient methods for stackelberg security games. In *Proc. of UAI*, pages 2–11, 2016.

[An and Tambe, 2017] B. An and M. Tambe. *Stackelberg Security Games (SSG) Basics and Application Overview*, pages 485–507. Cambridge U. Press, 2017.

[Basak *et al.*, 2016] A. Basak, F. Fang, T. H. Nguyen, and C. Kiekintveld. Abstraction methods for solving graph-based security games. In *AAMAS*, pages 13–33, 2016.

[Castiglioni *et al.*, 2021] M. Castiglioni, A. Marchesi, and N. Gatti. Committing to correlated strategies with multiple leaders. *Artificial Intelligence*, page 103549, 2021.

[Chalkiadakis *et al.*, 2011] G. Chalkiadakis, E. Elkind, and M. Wooldridge. Computational aspects of cooperative game theory. *Synthesis Lectures on AI and ML*, 5(6), 2011.

[Chander, 2010] Parkash Chander. Cores of games with positive externalities. *CORE DP*, 4:2010, 2010.

[Gan *et al.*, 2018] J. Gan, E. Elkind, and M. Wooldridge. Stackelberg security games with multiple uncoordinated defenders. In *AAMAS*. ACM Press, 2018.

[Gan *et al.*, 2020] J. Gan, E. Elkind, S. Kraus, and Mi. Wooldridge. Mechanism design for defense coordination in security games. In *AAMAS*, pages 402–410, 2020.

[Jiang *et al.*, 2013] A. X. Jiang, Thanh H. Nguyen, M. Tambe, and A. D Procaccia. Monotonic maximin: A robust stackelberg solution against boundedly rational followers. In Sajal K. Das, Cristina Nita-Rotaru, and Murat Kantarcioglu, editors, *Decision and Game Theory for Security*, pages 119–139, 2013.

[Kiekintveld *et al.*, 2011] C. Kiekintveld, J. Marecki, and M. Tambe. Approximation methods for infinite bayesian stackelberg games: Modeling distributional payoff uncertainty. In *AAMAS*, pages 1005–1012, 2011.

[Kiekintveld *et al.*, 2013] C. Kiekintveld, T. Islam, and Vl. Kreinovich. Security games with interval uncertainty. In *AAMAS*, page 231–238, 2013.

[Moulin and Peleg, 1982] H Moulin and B Peleg. Cores of effectivity functions and implementation theory. *Journal of Mathematical Economics*, 10(1):115 – 145, 1982.

[Mutzari *et al.*, 2021] D. Mutzari, J. Gan, and S. Kraus. Coalition formation in multi-defender security games. In *AAAI*, volume 35, pages 5603–5610, 2021.

[Mutzari *et al.*, 2022] Dolev Mutzari, Yonatan Aumann, and Sarit Kraus. Dolev mutzari yonatan aumann and sarit krausrobust solutions for multi-defender stackelberg security games. *arXiv:2204.14000*, 2022.

[Nguyen *et al.*, 2013] Thanh Hong Nguyen, Rong Yang, Amos Azaria, Sarit Kraus, and Milind Tambe. Analyzing the effectiveness of adversary modeling in security games. In *AAAI*, 2013.

[Nguyen *et al.*, 2014] Thanh Hong Nguyen, Albert Xin Jiang, and Milind Tambe. Stop the compartmentalization: unified robust algorithms for handling uncertainties in security games. In *AAMAS*, pages 317–324, 2014.

[Paruchuri *et al.*, 2008a] Praveen Paruchuri, Sarit Kraus, Jonathan P Pearce, Janusz Marecki, Milind Tambe, and Fernando Ordonez. Playing games for security: An efficient exact algorithm for solving bayesian stackelberg games. *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems*, 2008.

[Paruchuri *et al.*, 2008b] Praveen Paruchuri, Jonathan P Pearce, Janusz Marecki, Milind Tambe, Fernando Ordonez, and Sarit Kraus. Efficient algorithms to solve bayesian stackelberg games for security applications. In *AAAI*, pages 1559–1562, 2008.

[Pita *et al.*, 2008] James Pita, Manish Jain, Janusz Marecki, Fernando Ordóñez, Christopher Portway, Milind Tambe, Craig Western, Praveen Paruchuri, and Sarit Kraus. Deployed armor protection: the application of a game theoretic model for security at the los angeles international airport. In *Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems: industrial track*, pages 125–132, 2008.

[Pita *et al.*, 2009] James Pita, Manish Jain, Fernando Ordóñez, Milind Tambe, Sarit Kraus, and Reuma Magori-Cohen. Effective solutions for real-world stackelberg games: When agents must deal with human uncertainties. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 1*, AAMAS '09, page 369–376, 2009.

[Qian *et al.*, 2015] Yundi Qian, B William, and Milind Tambe. Robust strategy against unknown risk-averse attackers in security games. In *AAMAS*, pages 1341–1349, 2015.

[Shapley, 1973] Shubik Shapley. *Game Theory in Economics - Chapter 6: Characteristic Function, Core and Stable Set*. RAND Corporation, 1973.

[Sinha *et al.*, 2018] Arunesh Sinha, Fei Fang, Bo An, Christopher Kiekintveld, and Milind Tambe. Stackelberg security games: Looking beyond a decade of success. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence*. IJCAI, 2018.

[Tambe, 2011] Milind Tambe. *Security and game theory: algorithms, deployed systems, lessons learned*. Cambridge University Press, 2011.

[Yang *et al.*, 2012] Rong Yang, Fernando Ordonez, and Milind Tambe. Computing optimal strategy against quantal response in security games. In *AAMAS*, pages 847–854, 2012.

[Yin and Tambe, 2012] Zhengyu Yin and Milind Tambe. A unified method for handling discrete and continuous uncertainty in bayesian stackelberg games. In *Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems*, AAMAS '12, page 855–862, 2012.