# Playing Repeated Network Interdiction Games with Semi-Bandit Feedback

**Qingyu Guo**[1], **Bo An**[2], **Long Tran-Thanh**[3]
[1]Joint NTU-UBC Research Centre of Excellence in Active Living for the Elderly, NTU, Singapore
[2]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[3]Electronics and Computer Science, University of Southampton, UK
[1,2]{qguo005,boan}@ntu.edu.sg,[3]ltt08r@ecs.soton.ac.uk

## Abstract

We study repeated network interdiction games with no prior knowledge of the adversary and the environment, which can model many real world network security domains. Existing works often require plenty of available information for the defender and neglect the frequent interactions between both players, which are unrealistic and impractical, and thus, are not suitable for our settings. As such, we provide the first defender strategy, that enjoys nice theoretical and practical performance guarantees, by applying the adversarial online learning approach. In particular, we model the repeated network interdiction game with no prior knowledge as an online linear optimization problem, for which a novel and efficient online learning algorithm, SBGA, is proposed, which exploits the unique semi-bandit feedback in network security domains. We prove that SBGA achieves sublinear regret against adaptive adversary, compared with both the best fixed strategy in hindsight and a near optimal adaptive strategy. Extensive experiments also show that SBGA significantly outperforms existing approaches with fast convergence rate.

## 1 Introduction

Many security domains involve interdiction of adversarial behaviors on networks, including infectious disease control, disruption of enemy's supply chain, and prevention of illegal drug smuggling [Assimakopoulos, 1987; Guo *et al.*, 2016b; Wood, 1993; Yin and An, 2016; Guo *et al.*, 2016a; Wang *et al.*, 2016; 2017]. Checkpoints, such as physical roadblocks and inspection stations, are placed on different positions in order to defend the physical networks. In order to optimally utilize limited security resources on checkpoints, earlier works model network interdiction as a one-shot leader-follower game, where the defender randomly allocates the resources first and the adversary, knowing the defender strategy with extensive surveillance, chooses an optimal action to respond. However, such models cannot capture many realistic scenarios, such as illegal drug interdiction, due to several unreasonable assumptions. First, it is well understood that complete rationality, unlimited observation and high level computational ability are not ideal for modeling human adversaries [An *et al.*, 2013; Camerer, 2003]. Second, the defender's prior knowledge of the adversary and the environment is extremely limited in practice, while existing works in network interdiction assume full information on the defender side. Finally, the one-shot game model neglects the repeated interactions between the defender and the adversary, which are common in many network interdiction domains. For example, it is reported that over one thousand tons of drugs are seized on land in 2009 within the United States [CTR., 2010], and the U.S. border patrol agency has to alter the checkpoint operation policy frequently against unknown and fickle smugglers [Office, 2009].

Existing works typically deal with part of these challenges, either the bounded rationality of the attacker or the repeated interactions, as we discuss in the next section. We are the first to study the network interdiction from a perspective that aims to address all these challenges. In particular, inspired by the work of Xu *et al.* [2016], we apply the adversarial online learning approaches to tackle the network interdiction problem. However, we have to emphasize that this application is in fact non-trivial, since the repeated network interdiction problem is formulated as an online submodular maximization problem due to the network structure, for which state-of-the-art learning algorithms fail due to their poor performance. To overcome this issue, we linearize our learning problem by using a novel decision space transformation. The cost of this linearization transformation, however, is the exponentially increased size of the decision space. As such, a direct application of existing online linear learning methods, such as geometric hedge [Bartlett *et al.*, 2008], FPL [Neu and Bartók, 2015] and Exp3 [Auer *et al.*, 2002], will fail, due to the large decision space. Thus, we propose a new online linear learning algorithm, which exploits the semi-bandit style feedback in the network interdiction domain, called *Semi-Bandit style Geometric decision algorithm against an Adaptive adversary (SBGA)* to provide online decisions for the defender. We formally show that SBGA achieves $\mathcal{O}(T^{2/3})$ regret bound against *adaptive* adversary compared with the best fixed strategy on hindsight, and a low regret compared with the near optimal adaptive defender strategy. Finally, extensive numerical evaluations are conducted to demonstrate that our approach outperforms existing online learning approaches significantly and achieves much faster convergence rate. In particular, SBGA requires significantly fewer rounds (typically 50 rounds) to obtain satisfactory low average regret solution against various adversarial behavior models, and can efficiently scale up to realistic sized game instances.

## 2 Related Work

Existing works in the security game literature typically focus on bounded rationality of the attacker by following the concept of Quantal Response [McFadden, 1976], in order to predict the future move of the attacker. Such models include, but are not limited to, e.g., Subjective Utility Quantal Response (SUQR) [Nguyen *et al.*, 2013], Bayesian SUQR [Yang *et al.*, 2014] and Robust SUQR [Haskell *et al.*, 2014]. However, those models suffer from a number of limitations [Kar *et al.*, 2015]. In particular, they failed to capture the attacker's adaptive behavior of changing the attacking strategy based on defender's past moves. Besides, all these approaches only focus on simple target protection scenarios.

Another line of related work applies online learning approaches on repeated security games [Blum *et al.*, 2014; Xu *et al.*, 2016]. Unfortunately, the scope of all these works is limited to the target protection scenarios and cannot be applied to the network security domain, where the defender utility is a complex submodular function (we will show later). The straightforward approaches for online submodular maximization have poor performance in network security domain. In particular, the state-of-the-art algorithm for the online submodular maximization problem is an online version of the greedy algorithm [Streeter and Golovin, 2008], which only achieves a low $(1 - \frac{1}{e})$-regret and the performance guarantee only holds for the *oblivious* adversary.

Although we provide a novel transformation from online submodular maximization to online linear optimization, existing approaches for online linear optimization cannot address all the brutal challenges that we face: *scalability*, *limited feedback* and *adaptive adversary*. Most works related to online linear optimization typically focus on the *oblivious* adversary [Kakade *et al.*, 2009; Awerbuch and Kleinberg, 2004; Abernethy *et al.*, 2008]. Bartlett *et al.* [2008] provided the Geometric Hedge algorithm with $\mathcal{O}(\sqrt{T})$ regret against adaptive adversary with high probability. However, their algorithm needs to store and compute on the entire exponentially large decision space. The most relevant approach is *Bandit-style Geometric decision algorithm against an Adaptive adversary (BGA)* which achieves sublinear regret against adaptive adversary with limited feedback [McMahan and Blum, 2004; Dani and Hayes, 2006]. However, the regret convergence rate is extremely low in practice since BGA fails to exploit the unique semi-bandit feedback in our problem.

## 3 Repeated *Network Interdiction Game (NIG)*

```
For all t=1,2,...,T, repeat
1. The defender chooses an allocation $S^t \in \mathcal{S}$ to play.
2. The attacker chooses a network flow $\boldsymbol{f}^t \in \mathcal{F}$ to play.
3. The defender observes the amount of interdicted flow at each
   operated checkpoint in $S$ and receives the utility $U_d(S^t, \boldsymbol{f}^t)$.
```

Figure 1: The Repeated NIG Procedure.

We first briefly describe the *Network Interdiction Game (NIG)* model which is suitable for various security domains, including escaping path interdiction [Jain *et al.*, 2011] and network flow interdiction [Guo *et al.*, 2016b]. NIG models an attacker and a defender who take actions on a capacitated graph $G = (\mathcal{V}, \mathcal{E})$, with node set $\mathcal{V}$ and edge set $\mathcal{E}$, and a capacity vector $\mathbf{c}$, where capacity $c_e$ represents the maximum amount of adversary flow passing through edge $e$. The graph is assumed to have a unique *source node* $s \in \mathcal{V}$ and a unique *sink node* $t \in \mathcal{V}$. The unique source/sink assumption is no loss of generality: a graph with multiple source nodes and sink nodes can be transformed into a single-source-sink graph by adding two new nodes $s$ and $t$ as the new unique source and sink nodes respectively, connecting s to each source node and t to each sink node with proper capacitated edges. Let $\mathcal{P}$ denote the set of candidate $s$-$t$ paths in $G$ for the attacker, such as the eight drug transportation corridors in the United States [CTR., 2010]. We denote by $m$ the number of candidate paths, i.e., $m = |\mathcal{P}|$. Let $\mathcal{I}$ denote the set of all *inspection stations (checkpoints)* and the defender can operate at most $k$ inspection stations at same time due to limited security resources. Let $n$ denote the number of checkpoints, i.e., $n = |\mathcal{I}|$. Each station $i \in \mathcal{I}$ is characterized by a location, either a node or an edge in the graph, and a constant parameter $\tau_i \in [0, 1]$ denoting the proportion of adversary flow interdicted at $i$ when operated, i.e., *inspection probability*[1]. We say $i \in p$ if the location of checkpoint $i$ is on path $p$. We assume that the inspection probability can be estimated by the defender from historical data, which is reasonable since the defender fully controls all inspection facilities [Office, 2009]. Let $S = \langle S_i \rangle$ denote the allocation of $k$ resources for the defender, i.e., $\sum_{i \in \mathcal{I}} S_i = k$, where $S_i \in \{0, 1\}$ and $S_i = 1$ indicates that the inspection station $i$ is operated. The set of all possible allocations is denoted by

$$\mathcal{S} = \{S \in \{0, 1\}^{|\mathcal{I}|} : \sum\nolimits_{i \in \mathcal{I}} S_i = k\}.$$

With little abuse of notation, let $S$ be exchangeable with the set of operated allocations, i.e., $S \subseteq \mathcal{I}$. The adversarial flow is represented by $\mathbf{f}$ where $f_p$ denotes the amount of adversary flow passing along path $p \in \mathcal{P}$. Let $\mathcal{F}$ denote the set of all feasible attacker strategies, i.e.,

$$\mathcal{F} = \{\mathbf{f} \geq \mathbf{0} : \sum\nolimits_{p \in \mathcal{P} : e \in p} f_p \leq c_e, \forall e \in \mathcal{E}; \sum\nolimits_{p \in \mathcal{P}} f_p \leq 1\}$$

where, w.l.o.g., the total amount of network flow is upper bounded by 1, for normalization reason. Given a defender resource allocation $S \in \mathcal{S}$ and an adversarial flow $\mathbf{f} \in \mathcal{F}$, the defender's utility is the sum of interdicted flows on all paths, i.e., $U_d(S, \mathbf{f}) = \sum_{p \in \mathcal{P}} (1 - \Phi(S, p)) f_p$, while the attacker's utility is the sum of successful flows on all paths, i.e., $U_a(S, \mathbf{f}) = \sum_{p \in \mathcal{P}} \Phi(S, p) f_p$ where $\Phi(S, p)$ represents the proportion of adversary flow on path $p$ not interdicted by

[1] Although we assume that the proportion $\tau$ of the flow is interdicted, our approach and theoretical analysis also apply to the stochastic interdiction where the flow is fully interdicted with probability $\tau$.

the operated inspection stations given $S$:

$$\Phi(S,p)=\prod\nolimits_{i\in p}(1-\tau_i)^{S_i}.$$

**Semi-Bandit Feedback:** In this paper, we study the repeated NIG where the game is played for $T$ rounds (see Figure 1). Unlike existing works which assume plenty of available information for both players, the defender in repeated NIG has no prior knowledge of the adversarial behavior and only knows $\mathcal{P}$ and a rough estimation of the upper bound of the adversarial flow's total amount. Regarding the feedback of past plays, a natural assumption for the repeated NIG is that the defender only knows the amount of flow interdicted by each operated checkpoint. Such information on each operated checkpoint is called the *semi-bandit feedback*. On the other side, the *only* requirement for the attacker is that he cannot observe the real-time defense $S^t$ before round $t$ starts since the game is simultaneously played at each round. Furthermore, we assume that the defender is an expected utility maximizer, while no behavior model for the attacker is required. In other words, the attacker could be an expected utility maximizer or irrational to any extent; could be fully adversarial, or a random player.

The above repeated NIG model is general for many realistic network security domains. For example, in the illegal drug trafficking scenario, the drug smuggling activity on the transportation network can be formulated as an adversarial network flow. The adversarial behavior, on the other hand, is hard to capture as the defender is facing several sophisticated drug trafficking cartels who hire people with unknown backgrounds to smuggle illegal drug [Beittel, 2015]. Besides, as pointed out by the government report, the checkpoint operation policy almost changes daily to deal with the fickle smugglers [Office, 2009], which is captured by the repeated manner of our model. Another example is the escaping path interdiction problem [Jain *et al.*, 2011], where the escaping path of the attacker can be modeled by a unit flow on the path. The roadblock placed by the defender can be treated as a checkpoint with interdiction probability 1.

**Regret:** Let $\mathcal{H}_t$ denote the history information of the game by time $t$ (inclusive), and $\mathcal{H}_0$ denote no history information at all. Given a sequence of adversarial flows $\mathbf{f}^1,...,\mathbf{f}^T$, where $\mathbf{f}^t$ may be adaptive depending on $\mathcal{H}_t$, we are interested in designing an *online* policy $S^1(\mathcal{H}_0),...,S^T(\mathcal{H}_{T-1})$ (possibly randomized) that maximizes the defender's expected utility $\mathbb{E}[\sum_{t=1}^T U_d(S^t,\mathbf{f}^t)]$, where the randomization is taken over the randomness of the policy and the environment. Alternatively, we aim at minimizing the defender's *regret*:

$$R_T=\max_{S\in\mathcal{S}}\sum\nolimits_{t=1}^T U_d(S,\mathbf{f}^t)-\mathbb{E}[\sum\nolimits_{t=1}^T U_d(S^t,\mathbf{f}^t)].$$

The first term $\max_{S\in\mathcal{S}}\sum_{t=1}^T U_d(S,\mathbf{f}^t)$ is the utility of the best hindsight allocation, which serves as the benchmark. Without any prior knowledge of the adversarial behavior, we take the worst case analysis and focus on the largest regret against all possible adaptive adversaries. This type of regret notion is common in the online learning theory literature [Cesa-Bianchi and Lugosi, 2006].

**Online Submodular Maximization:** As shown by Proposition 1, the defender utility function is submodular with allocation $S$. At the first glance, we may relate the repeated NIG with the online submodular maximization problem where the decision-maker faces a sequence of submodular reward function and chooses a fix sized subset to play at each round. However, existing approaches for online submodular maximization cannot perform well in repeated NIG, as we discussed in Section 2. As such, we propose a novel approach to solve the repeated NIG efficiently with low regret solution. In particular, we first provide a non-trivial transformation from the repeated NIG as an online submodular maximization problem to an online linear optimization problem. To solve the transformed online linear optimization problem, we propose a novel algorithm called SBGA which exploits the semi-bandit feedback on operated checkpoints and obtains low regret solutions.

**Proposition 1.** *The defender utility function $U_d(S,\mathbf{f})$ is submodular with allocation $S$.*

*Proof.* For each $S\subseteq\mathcal{I}$ and $i\in\mathcal{I}\setminus S$, we have:

$$U_d(S\cup\{i\},\mathbf{f})-U_d(S,\mathbf{f})=\sum\nolimits_{p\in\mathcal{P}:i\in p}\tau_i\cdot\Phi(S,p)f_p.$$

Since $\Phi(S',p)\geq\Phi(S'',p)$ for any pair of $S'\subseteq S''$, $U_d(S,\mathbf{f})$ is submodular with allocation $S$ when $\mathbf{f}$ is fixed. □

## 4 From Online Submodular Maximization to Online Linear Optimization

In this section, we show a transformation of repeated NIG to an online linear optimization problem for which efficient low regret solution is possible. The intuition is as follows: if we look at the defender utility function $U_d(S,\mathbf{f})$ in details, we can find that the only parameter unknown is the adversarial flow $\mathbf{f}$. Fortunately, $U_d(S,\mathbf{f})$ is actually linear with $\mathbf{f}$. In other words, we can write down $U_d(S,\mathbf{f})$ as $\mathbf{w}\cdot\mathbf{f}$ where the coefficient $\mathbf{w}$, depending on the allocation $S$, is actually known to the defender since the interdiction probability can be estimated. As such, we define a mapping $\phi:\mathcal{S}\rightarrow\mathbb{R}^m$ such that for each allocation $S\in\mathcal{S}$, $\phi(S)=\mathbf{w}=\langle w_p\rangle$ with $w_p=1-\Phi(S,p), \forall p\in\mathcal{P}$. Given an allocation $S$ and $\mathbf{w}=\phi(S)$, we have: $U_d(S,\mathbf{f})=\mathbf{w}\cdot\mathbf{f}$.

Therefore, instead of regarding the repeated NIG as allocating resources for a sequence of unknown submodular functions, we can treat it as selecting decision points $\mathbf{w}$ for the coming unknown linear functions. The later one is an online linear optimization problem, as shown in Figure 2, where $\phi(\mathcal{S})=\{\mathbf{w}|\exists S\in\mathcal{S}:\mathbf{w}=\phi(S)\}$ denotes the set of points mapped from $\mathcal{S}$ with $\phi$. However, as we discussed in Section 2, the most relevant existing approach BGA fails to solve such an online linear optimization problem due to the slow regret convergence rate. The main drawback of BGA is that once an allocation $S$ is played, only the total amount of interdicted flow of $S$ is utilized to learn the unknown adversarial flow. In this case, BGA has to put a lot of effort in exploration of estimating the adversarial flow. However, as we will show later, by further exploiting the semi-bandit feedback on each operated checkpoint of $S$, the exploration efficiency can be improved significantly. Based on this idea, we propose the SBGA algorithm with nice and provable guarantees and satisfactory practical performance.

**For all** t=1,2,...,T, **repeat**
1. The defender chooses a point $\boldsymbol{w^t} \in \phi(\mathcal{S})$.
2. The attacker chooses a network flow $\boldsymbol{f^t} \in \mathcal{F}$ to play.
3. The defender receives the utility $\boldsymbol{w^t} \cdot \boldsymbol{f^t}$.

Figure 2: Online Linear Optimization Problem.

## 5 SBGA

As depicted in Algorithm 1, the intuition of SBGA is as follows: suppose that the full information of the adversarial flows of past plays can be accessed in repeated NIG, then an efficient online algorithm with low regret against adaptive adversary is available, denoted as *Geometric EXperts algorithm (GEX)* (Algorithm 2). Unfortunately, we only have semi-bandit feedback in repeated NIG. As such, SBGA samples from an *exploration basis* $\mathcal{S}^{eb}$, exploits the received semi-bandit feedback on operated checkpoints, and builds an "imaginary" full information game where the unknown adversarial flows (of past plays) are replaced by their unbiased estimators $\hat{\mathbf{f}}$. The algorithm GEX is then applied on the "imaginary" repeated NIG to provide online decisions. We now describe in details each components of SBGA.

**GEX:** Suppose that the adversarial flows $\mathbf{f}^1, ..., \mathbf{f}^{t-1}$ of previous rounds are known. We adopt the algorithm proposed by Kalai and Vempala [2003] to serve as GEX subroutine, shown in Algorithm 2. In particular, GEX returns the optimal decision $\mathbf{w}$ in $\phi(\mathcal{S})$ against the cumulative flow of previous rounds perturbed by a random noise vector $\mathbf{z}$. The scale of $\mathbf{z}$ is controlled by a learning parameter $\epsilon$, which balances the tradeoff between *exploitation* of playing the best hindsight decision so far, and *exploration* of adding perturbations to make the algorithm less predictable, especially for the adaptive adversary. Such tradeoff between exploitation and exploration is essential in the online learning theory, and GEX successfully addresses it by setting $\epsilon$ to a delicate value and achieves sublinear $\mathcal{O}(\sqrt{T})$ regret [Kalai and Vempala, 2003].

**Exploration and Unbiased Estimator:** As we mentioned before, at round $t$, SBGA creates an unbiased estimator $\hat{\mathbf{f}}^t$ of the unknown adversarial flow $\mathbf{f}^t$ by sampling from an exploration basis (with some probability), and hence an "imaginary" game, for which GEX is applicable, is constructed. To do so, we first show that the semi-bandit feedback on each operated checkpoint is linear with the unknown adversarial flow. Specifically, for each allocation $S$, let $r^{S,i,\mathbf{f}}$ denote the amount of interdicted flow on operated checkpoint $i \in S$ against adversarial flow $\mathbf{f}$ when $S$ is played. We have:

$$r^{S,i,\mathbf{f}} = \sum\nolimits_{p\in\mathcal{P}} \prod\nolimits_{j\in S(p,i)} (1-\tau_j)\tau_i \alpha_{p,i} \cdot f_p \quad \forall i \in S$$

where $S(p,i)$ denotes the set of checkpoints in $S$ such that path $p$ passes through before checkpoint $i$, and $\alpha_{p,i}$ is the indicator which takes value 1 if $i$ is on path $p$ and 0 otherwise. Let $\mathbf{w}^{S,i}$ denote the vector such that:

$$w_p^{S,i} = \prod\nolimits_{j\in S(p,i)} (1-\tau_j)\tau_i \cdot \alpha_{p,i} \quad \forall p \in \mathcal{P}. \qquad (1)$$

$$\underbrace{\begin{bmatrix} w_1^{S',1} & \cdots & w_m^{S',1} \\ \vdots & & \vdots \\ w_1^{S',k} & \cdots & w_m^{S',k} \\ w_1^{S'',k+1} & \cdots & w_m^{S'',k+1} \\ \vdots & & \vdots \\ w_1^{S'',m} & \cdots & w_m^{S'',m} \end{bmatrix}}_{W^\dagger} \underbrace{\begin{bmatrix} f_1{}^t \\ f_2{}^t \\ \vdots \\ f_m{}^t \end{bmatrix}}_{\boldsymbol{f}^t} = \underbrace{\begin{bmatrix} r^{S',1,\boldsymbol{f}^t} \\ \vdots \\ r^{S',k,\boldsymbol{f}^t} \\ r^{S'',k+1,\boldsymbol{f}^t} \\ \vdots \\ r^{S'',m,\boldsymbol{f}^t} \end{bmatrix}}_{\boldsymbol{r}^t}$$

(a) True adversarial flow

$$\underbrace{\begin{bmatrix} w_1^{S',1} & \cdots & w_m^{S',1} \\ \vdots & & \vdots \\ w_1^{S',k} & \cdots & w_m^{S',k} \\ w_1^{S'',k+1} & \cdots & w_m^{S'',k+1} \\ \vdots & & \vdots \\ w_1^{S'',m} & \cdots & w_m^{S'',m} \end{bmatrix}}_{W^\dagger} \underbrace{\begin{bmatrix} \hat{f}_1{}^t \\ \hat{f}_2{}^t \\ \vdots \\ \hat{f}_m{}^t \end{bmatrix}}_{\hat{\boldsymbol{f}}^t} = \underbrace{\begin{bmatrix} 2r^{S',1,\boldsymbol{f}^t} \\ \vdots \\ 2r^{S',k,\boldsymbol{f}^t} \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{\hat{\boldsymbol{r}}^t}$$

(b) Estimation when $S'$ is drawn

Figure 3: Exploration with $\mathcal{S}^{eb} = \{S', S''\}$.

We have: $r^{S,i,\mathbf{f}} = \mathbf{w}^{S,i} \cdot \mathbf{f}$ for each checkpoint $i \in S$, and $\mathbf{w}^{S,i}$ is known to the defender. Once $S$ is played, $r^{S,i,\mathbf{f}}$ is revealed for each $i \in S$. Let $m \times k$ matrix $W^S = \bigcup_{i\in S} \mathbf{w}^{S,i}$.

To illustrate the exploration in SBGA, consider an example where $m = 2k$ and let $\mathcal{S}^{eb} = \{S', S''\}$, as shown in Figure 3. To simplify the explanation, suppose that the checkpoints set $\mathcal{I} = \{1, ..., n\}$ and $n \geq m$. Allocation $S'$ operates the first $k$ checkpoints in $\mathcal{I}$ and $S''$ operates checkpoints $\{k+1, ..., m\}$. Let $m \times m$ matrix $W = [W^{S'}, W^{S''}]$, whose transpose $W^\dagger$ is illustrated in the figure. Each element of $W^\dagger$ is defined in (1). Assume that $W$ is full rank. At round $t$, let $\mathbf{r}^t$ be the vector of semi-bandit feedbacks as shown in Figure 3(a), and the true adversarial flow satisfies $W^\dagger \mathbf{f}^t = \mathbf{r}^t$. However, $\mathbf{r}^t$ is not available since only one allocation is played at round $t$, which only reveals half of $\mathbf{r}^t$. Thus, we randomly pick one allocation in $\mathcal{S}^{eb}$ to play and estimate $\mathbf{r}^t$ with $\hat{\mathbf{r}}^t$, as shown in Figure 3(b) where $S'$ is drawn and semi-bandit feedbacks of $S'$ are doubled in $\hat{\mathbf{r}}^t$. We estimate the flow with $\hat{\mathbf{f}}^t = (W^\dagger)^{-1}\hat{\mathbf{r}}^t$ and it is easy to see that $\mathbb{E}[\hat{\mathbf{f}}^t] = \mathbf{f}^t$ since $\mathbb{E}[\hat{\mathbf{r}}^t] = \mathbf{r}^t$.

In general, assume that $m$ is dividable by $k$ and let $\bar{m} = \frac{m}{k}$. Let $\mathcal{S}^{eb}$ be an *exploration basis* of $\bar{m}$ allocations, and let $m \times m$ matrix $W = \bigcup_{S\in\mathcal{S}^{eb}} W^S = [\mathbf{b}^1, ..., \mathbf{b}^m]$ where the $l$-th column is denoted as $\mathbf{b}^l$. W.l.o.g., suppose $W$ to be full rank.[2] At round $t$, as illustrated in Algorithm 1, with probability $\lambda$, SBGA uniformly samples one allocation $S$ in $\mathcal{S}^{eb}$ to play (line 10), and observes the interdicted flow for each operated checkpoint $i$ with amount $r^{S,i,\mathbf{f}^t} = \mathbf{w}^{S,i} \cdot \mathbf{f}^t$ (line 13). SBGA estimates flow $\mathbf{f}^t$ according to the idea mentioned before. The estimation $\hat{\mathbf{f}}^t = (W^\dagger)^{-1}\hat{\mathbf{r}}^t$ where the vector $\hat{\mathbf{r}}^t$ is defined as follows: $\hat{r}_l^t = (\bar{m}/\gamma)\mathbf{b}^l \cdot \mathbf{f}^t$ if $\mathbf{b}^l \in W^S$ and $\hat{r}_l^t = 0$ otherwise (lines 14 & 15). Proposition 2 shows that $\mathbb{E}[\hat{\mathbf{f}}^t] = \mathbf{f}^t$.

**Proposition 2.** $\hat{\mathbf{f}}^t$ *in SBGA is an unbiased estimator of* $\mathbf{f}^t$.

*Proof.* Let $\hat{\mathbf{r}}^{S,t}$ denote the vector $\hat{\mathbf{r}}^t$ when $S \in \mathcal{S}^{eb}$ is drawn.

[2] These assumptions are only for simplifying the analysis. The algorithm and analysis work for general cases where usually $\bar{m} = \lceil \frac{m}{k} \rceil$. In practice, the full rank assumption of $W$ may not hold true for any set of $\bar{m}$ allocations. Thus, the allocation with larger rank of $W^S$ will be selected into $\mathcal{S}^{eb}$ to reduce the size of exploration basis.

**Algorithm 1:** SBGA

```
1  Parameter: $\gamma$ and $\epsilon$, where $\epsilon$ is a parameter of GEX
2  Initialize $\mathcal{S}^{eb}$, let $W = \bigcup_{S \in \mathcal{S}^{eb}} W^S = [\mathbf{b}^1, .., \mathbf{b}^m]$
3  for $t = 1, ..., T$ do
4      Let $\chi^t = 1$ with probability $\gamma$ and $\chi^t = 0$ otherwise
5      if $\chi^t = 0$ then
6          Select $\mathbf{w}^t$ from the distribution GEX($\hat{\mathbf{f}}^1, ..., \hat{\mathbf{f}}^{t-1}$)
7          Receive utility $u^t = \mathbf{w}^t \cdot \mathbf{f}^t$
8          $\hat{\mathbf{f}}^t = \mathbf{0} \in \mathbb{R}^m$
9      else
10         Draw $S$ uniformly at random from $\mathcal{S}^{eb}$
11         $\mathbf{w}^t = \phi(S)$
12         Receive utility $u^t = \mathbf{w}^t \cdot \mathbf{f}^t$
13         Observe interdicted flow at each $i \in S$: $r^{S,i,\mathbf{f}^t}$
14         Let $\hat{\mathbf{r}}^t \in \mathbb{R}^m$ by $\hat{r}^t_l = 0$ for $\mathbf{b}^l \notin W^S$ and
           $\hat{r}^t_l = (\bar{m}/\gamma)\mathbf{b}^l \cdot \mathbf{f}^t$ otherwise
15         $\hat{\mathbf{f}}^t = (W^\dagger)^{-1}\hat{\mathbf{r}}^t$
```

**Algorithm 2:** GEX

```
Parameter: $\epsilon$, network flows $\mathbf{f}^1, ..., \mathbf{f}^{t-1}$ of previous rounds
Define a noise vector $\mathbf{z} \in \mathbb{R}^m$ such that $z_i$ is uniformly drawn
  from $[0, \frac{1}{\epsilon}]$
$\mathbf{w}^t = \arg\max_{\mathbf{w} \in \phi(\mathcal{S})} \mathbf{w} \cdot (\mathbf{f}^1 + ... + \mathbf{f}^{t-1} + \mathbf{z})$
return $\mathbf{w}^t$
```

$\hat{r}^{S,t}_l = (\bar{m}/\gamma)\mathbf{b}^l \cdot \mathbf{f}^t$ if $\mathbf{b}^l \in W^S$, and we have:

$$\sum\nolimits_{S \in \mathcal{S}^{eb}} \hat{\mathbf{r}}^{S,t} = (\bar{m}/\gamma) W^\dagger \mathbf{f}^t. \tag{2}$$

$$\begin{aligned}\mathbb{E}[\hat{\mathbf{f}}^t] &= (W^\dagger)^{-1}\mathbb{E}[\hat{\mathbf{r}}^t] = (W^\dagger)^{-1}\frac{\gamma}{\bar{m}}\sum\nolimits_{S \in \mathcal{S}^{eb}} \hat{\mathbf{r}}^{S,t} \\ &= (W^\dagger)^{-1} W^\dagger \mathbf{f}^t = \mathbf{f}^t.\end{aligned}$$

□

**Exploitation with GEX:** At this point, the unbiased estimator of the adversarial flow is obtained. At each round $t$, with probability $1 - \lambda$, SBGA applies GEX routine to *exploit* the "imaginary" game, where the adversarial flows are replaced by their unbiased estimators $\hat{\mathbf{f}}^1, ..., \hat{\mathbf{f}}^{t-1}$, to generate decision $\mathbf{w}^t$ and the corresponding allocation $S^t$ to play (line 6).[3] Parameter $\lambda$ balances the tradeoff between exploitation of playing decisions provided by GEX and exploration of sampling from $\mathcal{S}^{eb}$ to obtain an low-variance estimation of adversarial flow. However, it is not clear yet whether SBGA can achieve acceptable sublinear regret against adaptive adversary since the analysis of existing approaches aimed at cost minimization [McMahan and Blum, 2004], while the repeated NIG is formulated as a utility maximization problem. The following section shows that the answer is affirmative.

## 6 Theoretical Analysis

We first show that SBGA achieves a sublinear regret $\mathcal{O}(T^{2/3})$ against the adaptive adversary, compared with the best fixed allocation on hindsight. The proof of Theorem 1 is provided in Appendix A.

**Theorem 1.** *If $\bar{m} = 1$, set $\gamma = T^{-1/3}$ and $\epsilon = \sqrt{\frac{m}{T}}$ (roughly optimal), we have: $R_T(SBGA) \le 5m^{1/2}T^{2/3}$; Otherwise, set $\gamma = \bar{m}T^{-1/3}$ and $\epsilon = \frac{1}{m}\sqrt{\gamma/T}$ (roughly optimal), we have: $R_T(SBGA) = \mathcal{O}((\beta_\infty + C)m\sqrt{\bar{m}}T^{2/3})$, where $\beta_\infty = \|(W^\dagger)^{-1}\|_\infty$ and $C$ is the spanning ratio of $W$ with respect to $\phi(\mathcal{S})$ defined below (Definition 1).*

**Definition 1.** *Given a basis of $\mathbb{R}^m$: $W = \{\mathbf{b}^1, ..., \mathbf{b}^m\}$, the spanning ratio of $W$ with respect to the subset $H \subseteq \mathbb{R}^m$ is the minimal value of $C$ such that: for any point $\mathbf{w} \in H$, we can write $\mathbf{w} = \sum_{l=1}^m \alpha_l \mathbf{b}^l$ where the coefficient $\alpha_l \in [-C, C]$.*

Besides the best fixed allocation on hindsight, we also set the optimal *adaptive* strategy as benchmark. Specifically, the defender applying the optimal adaptive strategy plays the optimal allocation against the adversarial flow at each round. Let $OPT$ be the cumulative defender utility of the optimal adaptive strategy. Our next result shows that SBGA achieves a low regret compared with a constant fraction $\delta$ of $OPT$.

**Theorem 2.** *Let $\tau_{min} = \min_{i \in \mathcal{I}} \tau_i$ and $\tau_{max} = \max_{i \in \mathcal{I}} \tau_i$, we have $\delta \cdot OPT - Reward(SBGA) \le \mathcal{O}(T^{2/3})$ where $\delta = \min\{1, \frac{k}{m}\}\frac{\tau_{min}}{1-(1-\tau_{max})^k}$.*

*Proof.* Suppose $k \le m$ and let $\delta = k/m$. Consider the fixed allocation which allocates the $k$ checkpoints on $k$ pathes with maximal cumulative flows in $\sum_{t=1}^T \mathbf{f}^t$:

$$\max\nolimits_{S \in \mathcal{S}} U_d(S, \sum\nolimits_{t=1}^T \mathbf{f}^t) \ge \frac{k}{m} \cdot \tau_{min} \sum\nolimits_{p \in \mathcal{P}} \sum\nolimits_{t=1}^T f^t_p.$$

For the optimal adaptive defender strategy,

$$\max\nolimits_{S \in \mathcal{S}} U_d(S, \mathbf{f}^t) \le [1 - (1 - \tau_{max})^k] \sum\nolimits_{p \in \mathcal{P}} f^t_p.$$

Therefore, we have:

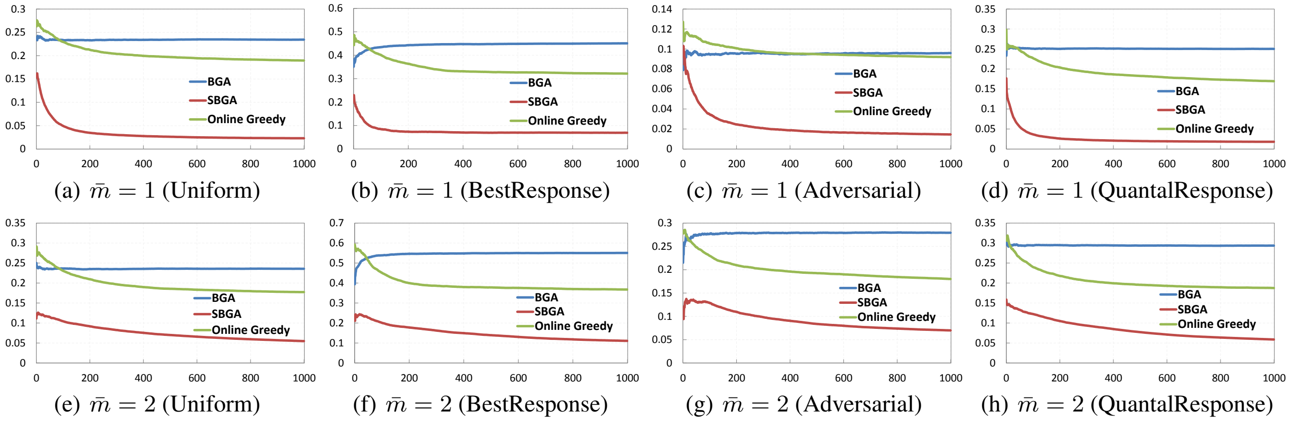
$$\frac{\max_{S \in \mathcal{S}} U_d(S, \sum_{t=1}^T \mathbf{f}^t)}{OPT(F)} \ge \frac{k}{m}\frac{\tau_{min}}{1 - (1 - \tau_{max})^k} = \delta$$

□

## 7 Experimental Evaluation

To demonstrate the practical applicability of our approach, we evaluate its performance through extensive experiments. All computations were performed on a 64-bit PC with 16 GB RAM and a quad-core 3.4 GHz processor. The tested networks are random planar graphs generated by the Waxman geographical model (*WG*) suitable for modeling transportation networks [Waxman, 1988]. By default, the instances are parameterized as follows: the number of nodes $|\mathcal{N}| = 200$ and the average degree is 3.0. The number of inspection stations $n = 100$. The number of candidate paths for adversarial flow $m = 20$, and the number of resources $k \in \{10, 20\}$, corresponding with $\bar{m} \in \{1, 2\}$. It is worthwhile to point out

[3] In order to apply the approach of Kalai and Vempala [2003], the estimated flows need to be transformed to meet its requirements, which is omitted in this paper for the ease of reading, and the readers can refer to appendix A of [McMahan and Blum, 2004] for details.



(a) $\bar{m} = 1$ (Uniform) (b) $\bar{m} = 1$ (BestResponse) (c) $\bar{m} = 1$ (Adversarial) (d) $\bar{m} = 1$ (QuantalResponse)

(e) $\bar{m} = 2$ (Uniform) (f) $\bar{m} = 2$ (BestResponse) (g) $\bar{m} = 2$ (Adversarial) (h) $\bar{m} = 2$ (QuantalResponse)

Figure 4: Experimental Evaluation (horizontal axis: round $t$, vertical axis: average regret per round).

that $\bar{m}$ is usually not large (around 2) in practice. Take the illegal drug trafficking scenario as an example. It is reported that 88 percent of all drug shipment is transported to the major drug markets within the United States through *eight* principal corridors [CTR., 2010]. Meanwhile, there are 39 tactical checkpoints owned by the Border Patrol in 2009 and the average operation percentage is about 8% for traffic jam concern [Office, 2009]. Thus, the settings here are realistic. All inspection stations are randomly placed on the graph and all candidate paths for the adversarial flow are randomly generated. The edge capacity $c_e$ is randomly chosen in $[0.5, 1.0]$, and inspection probability $\tau_i \sim [0.2, 0.6]$. The total number of rounds $T = 1000$. The results are averaged on 100 runs on *one* randomly generated instance. However, we do emphasize that the convergence trend is almost the same across simulated instances except in the initial rounds.

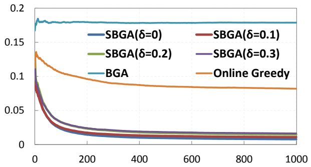
We compare SBGA against two benchmark algorithms: i) *BGA* [McMahan and Blum, 2004], and ii) *Online Greedy* algorithm for online submodular maximization [Streeter and Golovin, 2008]. The submodular maximization problem of GEX subroutine in SBGA is easily formulated as a convex integer program, which is solved by KNITRO (version 9.0.0) efficiently. The learning parameters in all algorithms are set to their optimal or rough optimal values w.r.t. the theoretical regret bounds. The other candidate benchmarks, including FPL-UE [Xu *et al.*, 2016], the variant of FPL algorithm for semi-bandit optimization [Neu and Bartók, 2015], and the Geometric Hedge algorithm [Dani *et al.*, 2007], are not tested here since they cannot scale up to instances with 100 checkpoints and over 10 resources. However, we do test their performance on small instances which is significantly worse than SBGA. Please see Appendix B for details.

We test our algorithm against four types of attackers, which together represent the majority of typical attacking models: i) *Uniform*: The attacker with uniform network flow at each round; ii) *BestResponse*: At round $t$, the attacker best responds to the empirical mixed defender strategy $\mathbf{x}$ of history with $x_S = t_S/(t-1)$ where $t_S$ represents the number of times that defender plays allocation $S$ in first $t-1$ rounds; iii) *Adversarial*: The adversarial type attacker aims at minimizing the defender's utility. In particular, at round $t$, the attacker plays the network flow which minimizes the defender utility of the best fixed allocation on hindsight, under the restriction that the amount of flow is no smaller than certain threshold, set to be half of the maximal flow; and iv) *QuantalResponse (QR)*: It is non-trivial to apply QR model in network security domain. Thus, we provide a simple implementation where we first uniformly generate a set of 50 network flows $\mathcal{F}^{qr}$. At round $t$, the attacker responds to the empirical mixed defender strategy $\mathbf{x}$ by playing a probability distribution over $\mathcal{F}^{qr}$ such that a network flow $\mathbf{f} \in \mathcal{F}^{qr}$ is chosen with probability $p_{\mathbf{f}} = \frac{e^{\lambda U_a(\mathbf{x},\mathbf{f})}}{\sum_{\mathbf{f'} \in \mathcal{F}^{qr}} e^{\lambda U_a(\mathbf{x},\mathbf{f'})}}$. The parameter $\lambda$ controls the rationality level of the attacker. When $\lambda = 0$, the attacker uniformly chooses the network flow in $\mathcal{F}^{qr}$ to play; while when $\lambda = \infty$, the attacker plays the best response flow in $\mathcal{F}^{qr}$.

**Solution Quality:** The performance of tested algorithms is depicted in Figure 4. We can observe that: i) the convergence rate of the average regret for SBGA is extremely fast and solutions with low enough regret (less than 20% of the average reward of the best fixed hindsight allocation) are obtained after about 50 rounds; ii) SBGA outperforms BGA and online greedy algorithm significantly in both convergence rate and average regret per round; iii) the convergence of the average regret for BGA cannot be observed due to the extensive exploration. In fact, according to the optimal learning parameter, for $T = 1000$ and $m = 20$, the exploration probability $\gamma$ is 1; and iv) it is obvious to see that the convergence rate of the average regret for online greedy algorithm slows down at a high regret level, which is reasonable since it only achieves low approximate regret.

Regarding the adversarial types, we can observe that the scale of regret against the BestResponse type adversary is significantly larger than others, which is reasonable since the BestResponse adversary will maximize the overall successful flows, and the corresponding network flow should have a larger amount. Even though, the average regret of SBGA is still low enough (around 0.1) for practical use. All these results support out intuition of exploiting the semi-bandit feedback and show that SBGA achieves low regret solutions with fast convergence rate against various realistic adaptive adversaries, whose practical applicability is acknowledged.

**Robustness:** Since SBGA requires a good estimation of the interdiction probability on each checkpoint, we evaluate the robustness of our algorithm under the uncertainty of the estimation of interdiction probability. In particular, let $\tilde{\tau}_i$ denote the defender's estimation of the interdiction probability of checkpoint $i$, which is generated in such way: $\tilde{\tau}_i \sim \tau_i \cdot [1-\delta, 1+\delta]$ where $\tau_i$ is the true interdiction probability and $\delta$ measures the defender's uncertainty on the estimation. We measure the performance of SBGA in games with $n = 100$, $k = m = 20$, and varying values of $\delta \in \{0, 0.1, 0.2, 0.3\}$, compared with BGA and online greedy algorithm, and the result is shown in the right figure. We can see that with higher uncertainty, the regret is larger, which is reasonable. Even though, the average regret of SBGA still outperforms BGA and online greedy algorithm significantly and its convergence rate did not decrease a lot.



## 8 Conclusions

This paper provides the first defender strategy in repeated network security domains with no prior knowledge of the adversarial behavior model and the environment. In particular, we proposed an adversarial online learning approach and non-trivially modeled the repeated network interdiction game as an online linear optimization problem, for which we provided a novel online learning algorithm, SBGA, to exploit the unique semi-bandit feedback in network interdiction domains. We formally proved that SGBA achieves low regret bounds compared both with best fixed strategy on hindsight, and the near optimal adaptive strategy. We have also run extensive experiments to show that SBGA efficiently obtains robust solutions with fast convergence rate against various realistic adversarial types. In addition, it significantly outperforms the existing methods. These imply the usefulness of our algorithm in many practical scenarios.

## Acknowledgements

This research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its IDM Futures Funding Initiative and NRF2015 NCR-NCR003-004. Tran-Thanh was supported by the EPSRC funded project EP/N02026X/1.

## A Proof of Theorem 1

We first explain some necessary notations. Let $\mathbf{r}^t \in [0,1]^m$ denote the vector where $r^t_l = \mathbf{b}^l \cdot \mathbf{f}^t$ for $\mathbf{b}^l \in W$. $\hat{\mathbf{r}}^t \in \mathbb{R}^m$ is SBGA's estimation of $\mathbf{r}^t$ in Algorithm 1. $\mathbf{w}^t \in \phi(\mathcal{S})$ is the decision made by SBGA in round $t$. Let $\hat{\mathbf{w}}^t \in \phi(\mathcal{S})$ be the decision recommended by GEX in round $t$, and $\hat{\mathbf{w}}^t = \mathbf{w}^t$ if SBGA does not sample the exploration basis in round $t$. $u^t \in [0,1]$ is SBGA's utility in round $t$, i.e., $u^t = \mathbf{w}^t \cdot \mathbf{f}^t$. Let $\hat{u}^t \in \mathbb{R}$ denote the utility of GEX in the "imaginary" game, i.e., $\hat{u}^t = \hat{\mathbf{w}}^t \cdot \hat{\mathbf{f}}^t$. Let $\mathcal{G}^t$ denote the history of the game by time $t$ (inclusive), which consists of all valid information of the game played so far. In particular, if we condition on a history $\mathcal{G}^t$, the random variables $\mathbf{f}^1, ..., \mathbf{f}^t, \hat{\mathbf{r}}^1, ..., \hat{\mathbf{r}}^t, \hat{\mathbf{f}}^1, ..., \hat{\mathbf{f}}^t, \mathbf{w}^1, ..., \mathbf{w}^t$ and $\chi^1, ..., \chi^t$ are fully determined. Let $\mathbf{w}^{1:t}$ be a sequence of decisions $\{\mathbf{w}^1, ..., \mathbf{w}^t\}$ and let $\mathbf{f}^{1:t}$ be a sequence of adversarial flows $\{\mathbf{f}^1, ..., \mathbf{f}^t\}$. We denote

$$utility(\mathbf{w}^{1:T}, \mathbf{f}^{1:T}) = \sum\nolimits_{t=1}^{T} \mathbf{w}^t \cdot \mathbf{f}^t$$

$$best(\mathbf{f}^{1:T}) = \arg\max\nolimits_{\mathbf{w} \in \phi(\mathcal{S})} \sum\nolimits_{t=1}^{T} \mathbf{w} \cdot \mathbf{f}^t$$

$$opt(\mathbf{f}^{1:T}) = best(\mathbf{f}^{1:T}) \cdot \sum\nolimits_{t=1}^{T} \mathbf{f}^t.$$

Theorem 1 can be directly derived from four inequalities:

$$\mathbb{E}[opt(\mathbf{f}^{\mathbf{1:T}})] \leq T \qquad \text{since } \|\mathbf{f}^t\|_1 \leq 1 \tag{3}$$

$$(1-\gamma)\mathbb{E}[utility(\hat{\mathbf{w}}^{1:T}, \hat{\mathbf{f}}^{1:T})] - \mathbb{E}[utility(\mathbf{w}^{1:T}, \mathbf{f}^{1:T})] \leq \gamma T \tag{4}$$

$$\mathbb{E}[opt(\hat{\mathbf{f}}^{1:T})] - \mathbb{E}[utility(\hat{\mathbf{w}}^{1:T}, \hat{\mathbf{f}}^{1:T})] \leq \begin{cases} (8m+2)C\sqrt{T/\gamma}, & \text{if } \bar{m} \geq 2; \\ 2\sqrt{mT}, & \text{if } \bar{m} = 1. \end{cases} \tag{5}$$

$$\mathbb{E}[opt(\mathbf{f}^{\mathbf{1:T}})] - \mathbb{E}[opt(\hat{\mathbf{f}}^{1:T})] \leq \begin{cases} m\bar{m}(\beta_\infty + 1)\sqrt{T/\gamma}, & \text{if } \bar{m} \geq 2; \\ \sqrt{mT/\gamma}, & \text{if } \bar{m} = 1. \end{cases} \tag{6}$$

*Proof of* (4). We follow the similar procedure to prove (4) as the proof of Theorem 3 in [McMahan and Blum, 2004].

$$\bar{\mathbf{w}}^t = \sum\nolimits_{\hat{\mathbf{w}}^t \in \phi(\mathcal{S})} Pr(\hat{\mathbf{w}}^t | \mathcal{G}^{t-1}) \hat{\mathbf{w}}^t.$$

Let $\hat{\mathbf{r}}^{t,S}$ and $\hat{\mathbf{f}}^{t,S}$ be the estimators of $\mathbf{r}^t$ and $\mathbf{f}^t$ when $S \in \mathcal{S}^{eb}$ is sampled in round $t$, according to (2):

$$\sum\nolimits_{S \in \mathcal{S}^{eb}} \hat{\mathbf{f}}^{t,S} = (W^\dagger)^{-1} \sum\nolimits_{S \in \mathcal{S}^{eb}} \hat{\mathbf{r}}^{t,S} = \frac{\bar{m}}{\gamma} \mathbf{f}^t. \tag{7}$$

$$\begin{aligned} &\mathbb{E}[\hat{u}^t | \mathcal{G}^{t-1}] \\ &= \sum\nolimits_{S \in \mathcal{S}^{eb}} \frac{\gamma}{\bar{m}} \sum\nolimits_{\hat{\mathbf{w}}^t \in \phi(\mathcal{S})} Pr(\hat{\mathbf{w}}^t | \mathcal{G}^{t-1}) (\hat{\mathbf{f}}^{t,S} \cdot \hat{\mathbf{w}}^t) \\ &= \gamma [\sum\nolimits_{S \in \mathcal{S}} \frac{1}{\bar{m}} \hat{\mathbf{f}}^{t,S}] \cdot \bar{\mathbf{w}}^t = \frac{\gamma}{\bar{m}} [\sum\nolimits_{S \in \mathcal{S}^{eb}} \hat{\mathbf{f}}^{t,S}] \cdot \bar{\mathbf{w}}^t \\ &= \mathbf{f}^t \cdot \bar{\mathbf{w}}^t \qquad \text{according to (7).} \end{aligned}$$

$$\begin{aligned} \mathbb{E}[u^t | \mathcal{G}^{t-1}] &= (1-\gamma)(\mathbf{f}^t \cdot \bar{\mathbf{w}}^t) + \gamma \sum\nolimits_{S \in \mathcal{S}^{eb}} \frac{1}{\bar{m}} (\mathbf{f}^t \cdot \phi(S)) \\ &\geq (1-\gamma)\mathbb{E}[\hat{u}^t | \mathcal{G}^{t-1}] - \gamma. \quad \text{since } \|\mathbf{f}^t\|_1 \leq 1 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[u^t] &= \mathbb{E}[\mathbb{E}[u^t | \mathcal{G}^{t-1}]] \geq \mathbb{E}[(1-\gamma)\mathbb{E}[\hat{u}^t | \mathcal{G}^{t-1}] - \gamma] \\ &= (1-\gamma)\mathbb{E}[\mathbb{E}[\hat{u}^t | \mathcal{G}^{t-1}]] - \gamma = (1-\gamma)\mathbb{E}[\hat{u}^t] - \gamma. \end{aligned}$$

Sum it up over $t = 1, ..., T$, and we will get the inequality (4). □

*Proof of Inequality* (5). We follow the sketch of proof of (2) in [Dani and Hayes, 2006]: Let $nonzero(\hat{\mathbf{f}}^{1:T})$ denote the sequence of non-zero flows in $\hat{\mathbf{f}}^{1:T}$ and let $\xi$ denote the length of $nonzero(\hat{\mathbf{f}}^{1:T})$. Let $regret(GEX, \hat{\mathbf{f}}^{1:T})$ denote the regret of GEX in the "imaginary" game.

**Case 1 [$\bar{m} \geq 2$]:** If $\bar{m} \geq 2$, $\hat{\mathbf{f}}^t$ can be negative, and the approach by Kalai and Vempala [2003] has to be adapted to serve as the GEX, which has been illustrated in appendix A of [McMahan and Blum, 2004]:

$$\mathbb{E}[regret(GEX, \hat{\mathbf{f}}^{1:T})|\xi] \leq \epsilon(4m+2)R^2\xi + 4m/\epsilon, \tag{8}$$

where $R$ is an upper bound of $|\hat{\mathbf{f}}^t \cdot \mathbf{w}|$ for $\mathbf{w} \in \phi(\mathcal{S})$:

$$\begin{aligned}|\hat{\mathbf{f}}^t \cdot \mathbf{w}| &= |(W^\dagger)^{-1}\hat{\mathbf{r}}^t \cdot W\boldsymbol{\alpha}| = |\hat{\mathbf{r}}^t \cdot \boldsymbol{\alpha}| \leq \|\hat{\mathbf{r}}^t\|_1 \|\boldsymbol{\alpha}\|_\infty \\ &\leq Cm/\gamma \qquad\qquad \text{since } \|\hat{\mathbf{r}}^t\|_1 \leq k\bar{m}/\gamma = m/\gamma,\end{aligned}$$

where $\mathbf{w} = W\boldsymbol{\alpha}$ with $\|\boldsymbol{\alpha}\|_\infty \leq C$ and $C$ is the spanning ratio of $W$ w.r.t. $\phi(\mathcal{S})$. Substitute to Eq.(8) with $\mathbb{E}[\xi] = \gamma T$:

$$\begin{aligned}&\mathbb{E}[regret(GEX, \hat{\mathbf{f}}^{1:T})] = \mathbb{E}[\mathbb{E}[regret(GEX, \hat{\mathbf{f}}^{1:T})|\xi]] \\ &\leq \epsilon(4m+2)C^2m^2T/\gamma + 4m/\epsilon.\end{aligned}$$

Let $\epsilon = \sqrt{\gamma/T}/m$, we have:

$$\mathbb{E}[regret(GEX, \hat{\mathbf{f}}^{1:T})] \leq (8m+2)C\sqrt{T/\gamma}.$$

**Case 2 [$\bar{m} = 1$]:** In this case, $\hat{\mathbf{f}}^t = \frac{1}{\gamma}\mathbf{f}^t$. The algorithm by Kalai and Vempala [2003] can be directly applied for GEX:

$$\begin{aligned}\mathbb{E}[regret(GEX, \hat{\mathbf{f}}^{1:T})|\xi] &\leq \epsilon\xi/\gamma + m/\epsilon \\ \mathbb{E}[regret(GEX, \hat{\mathbf{f}}^{1:T})] &\leq \epsilon T + m/\epsilon.\end{aligned}$$

Set $\epsilon = \sqrt{m/T}$, we get:

$$\mathbb{E}[regret(GEX, \hat{\mathbf{f}}^{1:T})] \leq 2\sqrt{mT}.$$

□

*Proof of Inequality* (6). We follow the sketch of proof of (3) in [Dani and Hayes, 2006], since $\|\mathbf{w}\|_2 \leq \sqrt{m}$ for $\mathbf{w} \in \phi(\mathcal{S})$:

$$|opt(\hat{\mathbf{f}}^{1:T}) - opt(\mathbf{f}^{1:T})| \leq \sqrt{m}\|\sum_{t=1}^{T}(\hat{\mathbf{f}}^t - \mathbf{f}^t)\|_2. \tag{9}$$

Define $\Delta^t = \hat{\mathbf{f}}^t - \mathbf{f}^t$, according to [Dani and Hayes, 2006]:

$$\mathbb{E}[\|\sum_{t=1}^{T}\Delta^t\|_2]^2 \leq \sum_{t=1}^{T}\mathbb{E}[\|\Delta^t\|_2^2]. \tag{10}$$

**Case 1 [$\bar{m} \geq 2$]:**

$$\begin{aligned}&\|\hat{\mathbf{f}}^t\|_2 \leq \sqrt{m}\|(W^\dagger)^{-1}\|_\infty\|\hat{\mathbf{r}}^t\|_\infty \leq \bar{m}\sqrt{m}\beta_\infty/\gamma \\ &\|\mathbf{f}^t\|_2\|\mathbf{f}^t\|_1 \leq 1.\end{aligned}$$

$$\|\Delta^t\|_2 \leq \|\hat{\mathbf{f}}^t\|_2 + \|\mathbf{f}^t\|_2 \leq \begin{cases} 1, & \text{w.p. } 1-\gamma; \\ \frac{\bar{m}\sqrt{m}\beta_\infty}{\gamma} + 1, & \text{w.p. } \gamma. \end{cases}$$

$$\mathbb{E}[\|\Delta^t\|_2^2] \leq (\bar{m}\sqrt{m}\beta_\infty + 1)^2/\gamma.$$

Substitute to (9) & (10):

$$\mathbb{E}[|opt(\hat{\mathbf{f}}^{1:T}) - opt(\mathbf{f}^{1:T})|] \leq \bar{m}m(\beta_\infty + 1)\sqrt{T/\gamma}$$

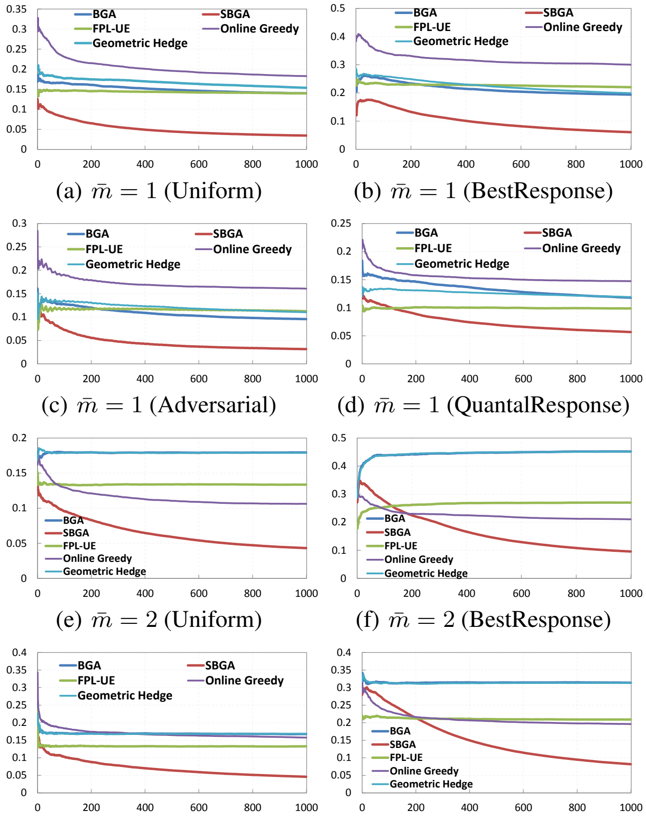
**Case 2 [$\bar{m} = 1$]:** Similarly, since $\|\hat{\mathbf{f}}^t\|_2 = \|\frac{1}{\gamma}\mathbf{f}^t\|_2 \leq \frac{1}{\gamma}$:

$$\mathbb{E}[|opt(\hat{\mathbf{f}}^{1:T}) - opt(\mathbf{f}^{1:T})|] \leq \sqrt{mT/\gamma}.$$

□

## B Supplementary Experimental Evaluation

We also conduct experimental evaluation on small instances with 20 checkpoints and 5 resources. Except BGA and online greedy algorithm, other benchmarks are also tested, including FPL-UE [Xu *et al.*, 2016], the variant of FPL algorithm for semi-bandit optimization [Neu and Bartók, 2015], and the Geometric Hedge algorithm [Dani *et al.*, 2007]. The results are depicted in Figure 5, from which we can see that even in the small instances, SBGA still outperforms all other benchmarks significantly w.r.t. both the convergence rate and low regret.



(a) $\bar{m} = 1$ (Uniform) (b) $\bar{m} = 1$ (BestResponse)

(c) $\bar{m} = 1$ (Adversarial) (d) $\bar{m} = 1$ (QuantalResponse)

(e) $\bar{m} = 2$ (Uniform) (f) $\bar{m} = 2$ (BestResponse)

(g) $\bar{m} = 2$ (Adversarial) (h) $\bar{m} = 2$ (QuantalResponse)

Figure 5: Supplementary Experimental Evaluation.

## References

[Abernethy *et al.*, 2008] Jacob D. Abernethy, Elad Hazan, and Alexander Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *COLT*, pages 263–274, 2008.

[An *et al.*, 2013] Bo An, Matthew Brown, Yevgeniy Vorobeychik, and Milind Tambe. Security games with surveillance cost and optimal timing of attack execution. In *AAMAS*, pages 223–230, 2013.

[Assimakopoulos, 1987] Nikitas Assimakopoulos. A network interdiction model for hospital infection control. *Computers in Biology and Medicine*, 17(6):413–422, 1987.

[Auer *et al.*, 2002] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. The nonstochastic multiarmed bandit problem. *SIAM Journal on Computing*, 32(1):48–77, 2002.

[Awerbuch and Kleinberg, 2004] Baruch Awerbuch and Robert D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *STOC*, pages 45–53, 2004.

[Bartlett *et al.*, 2008] Peter L. Bartlett, Varsha Dani, Thomas P. Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *COLT*, pages 335–342, 2008.

[Beittel, 2015] June S. Beittel. *Mexico: Organzied Crime and Drug Trafficking Organizations*. Congressional Research Service, 2015.

[Blum *et al.*, 2014] Avrim Blum, Nika Haghtalab, and Ariel D. Procaccia. Learning optimal commitment to overcome insecurity. In *NIPS*, pages 1826–1834, 2014.

[Camerer, 2003] Colin Camerer. *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, 2003.

[Cesa-Bianchi and Lugosi, 2006] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[CTR., 2010] US DEPT OF JUSTICE NATL DRUG INTELLIGENCE CTR. National drug threat assessment, 2010.

[Dani and Hayes, 2006] Varsha Dani and Thomas P Hayes. Robbing the bandit: Less regret in online geometric optimization against an adaptive adversary. In *ACM-SIAM/SODA*, pages 937–943. Society for Industrial and Applied Mathematics, 2006.

[Dani *et al.*, 2007] Varsha Dani, Thomas P. Hayes, and Sham Kakade. The price of bandit information for online optimization. In *NIPS*, pages 345–352, 2007.

[Guo *et al.*, 2016a] Qingyu Guo, Bo An, Yevgeniy Vorobeychik, Long Tran-Thanh, Jiarui Gan, and Chunyan Miao. Coalitional security games. In *AAMAS*, pages 159–167, 2016.

[Guo *et al.*, 2016b] Qingyu Guo, Bo An, Yair Zick, and Chunyan Miao. Optimal interdiction of illegal network flow. In *IJCAI*, pages 2507–2513, 2016.

[Haskell *et al.*, 2014] William B. Haskell, Debarun Kar, Fei Fang, Milind Tambe, Sam Cheung, and Elizabeth Denicola. Robust protection of fisheries with compass. In *AAAI*, pages 2978–2983, 2014.

[Jain *et al.*, 2011] Manish Jain, Dmytro Korzhyk, Ondrej Vanek, Vincent Conitzer, Michal Pechoucek, and Milind Tambe. A double oracle algorithm for zero-sum security games on graphs. In *AAMAS*, pages 327–334, 2011.

[Kakade *et al.*, 2009] Sham M. Kakade, Adam Tauman Kalai, and Katrina Ligett. Playing games with approximation algorithms. *SIAM J. Comput.*, 39(3):1088–1106, 2009.

[Kalai and Vempala, 2003] Adam Kalai and Santosh Vempala. Efficient algorithms for online decision problems. In *COLT*, pages 26–40, 2003.

[Kar *et al.*, 2015] Debarun Kar, Fei Fang, Francesco Maria Delle Fave, Nicole Sintov, and Milind Tambe. "a game of thrones": When human behavior models compete in repeated Stackelberg security games. In *AAMAS*, pages 1381–1390, 2015.

[McFadden, 1976] Daniel L McFadden. Quantal choice analaysis: A survey. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pages 363–390. 1976.

[McMahan and Blum, 2004] H. Brendan McMahan and Avrim Blum. Online geometric optimization in the bandit setting against an adaptive adversary. In *COLT*, pages 109–123, 2004.

[Neu and Bartók, 2015] Gergely Neu and Gábor Bartók. Importance weighting without importance weights: An efficient algorithm for combinatorial semi-bandits. *CoRR*, abs/1503.05087, 2015.

[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.

[Office, 2009] United States Government Accountability Office. Border patrol: Checkpoints contribute to border patrols mission, but more consistent data collection and performance measurement could improve effectiveness, 2009.

[Streeter and Golovin, 2008] Matthew J. Streeter and Daniel Golovin. An online algorithm for maximizing submodular functions. In *NIPS*, pages 1577–1584, 2008.

[Wang *et al.*, 2016] Zhen Wang, Yue Yin, and Bo An. Computing optimal monitoring strategy for detecting terrorist plots. In *AAAI*, pages 637–643, 2016.

[Wang *et al.*, 2017] Xinrun Wang, Qingyu Guo, and Bo An. Stop nuclear smuggling through efficient container inspection. In *AAMAS*, pages 669–677, 2017.

[Waxman, 1988] Bernard M Waxman. Routing of multipoint connections. *IEEE Journal on Selected Areas in Communications*, 6(9):1617–1622, 1988.

[Wood, 1993] R Kevin Wood. Deterministic network interdiction. *Mathematical and Computer Modelling*, 17(2):1–18, 1993.

[Xu *et al.*, 2016] Haifeng Xu, Long Tran-Thanh, and Nicholas R. Jennings. Playing repeated security games with no prior knowledge. In *AAMAS*, pages 104–112, 2016.

[Yang *et al.*, 2014] Rong Yang, Benjamin J. Ford, Milind Tambe, and Andrew Lemieux. Adaptive resource allocation for wildlife protection against illegal poachers. In *AAMAS*, pages 453–460, 2014.

[Yin and An, 2016] Yue Yin and Bo An. Efficient resource allocation for protecting coral reef ecosystems. In *IJCAI*, pages 531–537, 2016.