Patrol Districting and Routing with Security Level Functions

Xu Chen*

Email: cx008@ie.cuhk.edu.hk
*Department of Information Engineering
The Chinese University of Hong Kong

Tak-Shing Peter Yum*†

Email: tsyum@ie.cuhk.edu.hk
†Hong Kong Applied Science and Technology
Research Institute Company Limited

***Abstract*—Public security is a key concern around the world. Efficient patrol strategy increases the effectiveness of police patrolling and improves public security. In this paper we propose a new general security measure by defining the security level function. Based on this, we present the balanced patrol districting solution for the multiple units assignment problem. For the patrol routing problem in a patrol district, we first formulate the patrol routing process as a graph-based Markov decision process, and then propose an $\epsilon-$optimal patrol routing strategy to deal with the curse of dimensionality. The strategy is derived based on the concept of $\epsilon-$optimal horizon approximation. Numerical studies demonstrate that the strategy is adaptive to the generalized security measure by security level function, and has significant performance improvement over the referenced strategies in previous works. In addition, as the randomness is an important factor for practices, we design the randomized patrol routing strategy on the basis of the randomized exploration method in the Reinforcement Learning.**

***Index Terms*—Public Security, Patrol Planning, Security Level Function, Patrol Districting, Patrol Routing, Markov Decision Process, Reinforcement Learning**

I. Introduction

Police patrolling is an important public service for security surveillance in an urban city. It helps to deter and prevent crimes, and creates a sense of public security. Police resources are limited, and there is an obvious public interest in developing low-cost approaches for improving the efficiency of police patrols. An effective solution for the patrol planning has other potential applications. It can be used in a military context, for example for information gathering by Unmanned Aerial Vehicles (UVAs), and in robotics, for intelligent monitoring critical infrastructures, such as electric power systems and water distribution networks.

In policing operations, a city is often divided into several police precincts. A precinct consists of multiple patrol regions. In a patrol region, patrol units are dispatched to conduct surveillance via patrolling in order to deter potential crimes and security threats by their presence.

For the patrolling problem, the first critical issue is to establish a proper measure that captures the dynamic security status of a location. Previous work such as [1] assumed the patrol environment is static and used the average crime arrival rate of a location as the security measure. This approach needs to be improved since it did not capture the dynamic behavior of crime arrivals [2]. As an alternative, the (weighted) patrol idling time of a location was used in [3], [4]. The idling time is defined as the time since the last visit to a location by a patrol unit. This measure did not describe the dynamic security status change explicitly. In this study, we propose a more general security measure by defining the security level function (SLF). Previous measures can be converted to a special case of the SLF. Moreover, for practical applications, the SLF can be obtained from the historical crime data by the statistics technique called survival analysis [7].

Based on the security measure by SLF, we propose the balanced districting solution to the multiple units assignment problem in a patrol region. Each patrol unit is assigned with a responsible district, with the goal to balance the workloads among the units and reduce the average response travel time.

In a patrol district, a patrol unit conducts security surveillance by touring different locations in the area. An efficient patrol routing strategy is therefore essential for improving the patrol effectiveness.

Several approaches have been proposed for the patrol routing problem. Sui et al. [1] assumed the crime arrival rate of a location to be a constant, modeled the patrolling process as a Markov decision process (MDP), and proposed a learning algorithm to maximize the long-term decision rewards approximately. Chevaleyre et al. [3] modeled the patrol area as a graph, and showed that the shortest Hamiltonian cycle is the optimal strategy for minimizing the average idling time of the locations, assuming that all the locations are equally important. This solution needs be generalized when there exist crime hotspots. Praveen et al. [6] modeled the interactions between the police and terrorists as a two-stage Stackelberg game, and derived an equilibrium strategy. This approach assumes that a player always predicts how his opponent will behave and chooses the best response.

Since the SLF is a generalized security measure, a more efficient patrol routing strategy is needed. In this paper, we formulate the patrol routing process in a patrol district as a graph-based Markov decision process. Due to the curse of the dimensions of the state space, the resulting MDP is computationally intractable. We thus propose the patrol routing strategy on the basis of $\epsilon-$optimal horizon approximation, and the strategy is shown to be an $\epsilon-$optimal policy. Numerical results

978-1-4244-6588-0/10/$25.00 ©2010 IEEE

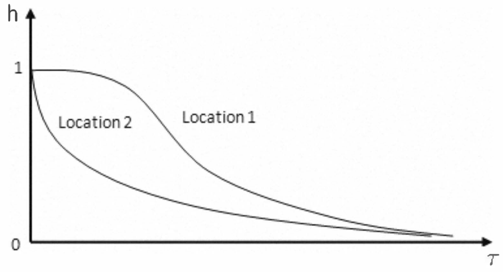

Fig. 1. Security Level Function

show that the proposed strategy has significant performance improvement over the referenced strategies in previous studies for the security measure by SLF. As the randomness is an important factor for practices, we design the randomized patrol routing strategy based on the randomized exploration method in the Reinforcement Learning.

The rest of the paper is organized as follows. We introduce the patrol model under the security measure by the SLF in Section II. In Section III, we address the patrol districting problem. In Section IV, we first formulate the patrol routing process as a MDP, and then propose the patrol routing strategy on the basis of the $\epsilon-$optimal horizon approximation. We introduce the randomized patrol routing strategy In Section V. In Section VI, we present the numerical results. In Section VII, we present our conclusion and discuss topics for future investigation.

II. Patrol Model

In this section, we build a mathematical model for the patrolling problem in a patrol region. We first introduce the generalized security measure by SLF, and then address the graphic model for the patrol region.

A. Security Level Function

As mentioned, previous works used a linear function of the idling time as the security measure. Let $\tau(t)$ be the idling time at time instance t. Then $\tau(t) = 0$ if the location is under patrolling, and $\tau(t + \triangle t) = \tau(t) + \triangle t$ if there are no patrol visits to the location in the time interval $(t, t + \triangle t)$. We generalize the security measure by defining the security level function (SLF). Intuitively, when the patrol unit dwells in a location, that location is in its best security status since the presence of the patrol unit will intercept or discourage potential crime. When the patrol unit leaves the location, the security status of the location would become worse as the idling time increases, i.e. the longer idling time the location has, the higher chances that an incident would occur. Formally, we define the SLF $h(\tau)$ as a function of the idling time, which gives the probability that location is secure without any incident when its idling time is τ. Thus the SLF $h(\tau)$ is a non-increasing function in the field $[1, 0]$ with $h(0) = 1$ and $h(\infty) = 0$. The Fig. 1 demonstrates the SLF's of two locations, with location 1 of lower crime rate.

TABLE I
KEY NOTATIONS

Symbol	Physical Meaning
$\tau_n(\cdot)$	Idling time of location n
$h_n(\cdot)$	Security level function of location n
$\mathcal{N} = \{1, \dots, N\}$	Set of locations in the patrol area
$\lambda_n(t)$	dynamic incident arrival rate function
λ_n	Average incident arrival rate of location n
$\mathcal{L}_z$	Workloads of the patrol district z
$\mathbf{B}_k$	Security state
a_k	Patrol pouting action
$R(\mathbf{B}_k, a_k)$	Immediate reward
$V^\pi(\cdot)$	Value function under policy π
$V^*_\epsilon(\cdot)$	Optimal Value by ϵ-optimal horizon approximation
ℓ_i	Action sequence
P^k	Transition Matrix

The security measure by SLF is more general. Previous measure by patrol idling time can be considered as a special case of SLF by just setting $h(\tau) = a^{w\tau}$, where $0 < a < 1$ is a normalized factor and w is the weight. If the dynamic crime arrival rate is concerned, the SLF can be defined as $h(\tau) = a^{\int_0^\tau \lambda(t)dt}$, where $\lambda_n(t)$ is the dynamic crime arrival rate function of a location. For practical applications, the SLF can be obtained by survival analysis using the historical crime data. The survival analysis is a branch of statistics involving the regression modeling of time to event data, which is widely applied in reliability analysis in engineering, diagnosis data analysis in medical research, and duration analysis in sociology. It has also been applied to criminology [7]. Under the framework of survival analysis, the SLF $h(\tau)$ is interpreted as the probability that the time to occurrence of an incident is larger than the given idling time τ. Formally, let T be the random variable denote the time to occurrence of an incident, then $h(\tau) = Pr\{T > \tau\}$.

B. Patrol Area Model

We model the patrol region as a graph $\mathcal{G}(\mathcal{N}, \mathcal{E})$. The graph is assumed to be connected, i.e. there is a route from a node to other nodes. For the case that the graph is not connected, it can be partitioned into multiple connected sub-graphs. We then introduce the patrol area model as follows:

- $\mathcal{N} = \{1, 2, \dots, N\}$ denotes a finite set of nodes, with a node $n \in \mathcal{N}$ representing a strategic patrol location in the area, such as a cluster of residential buildings or a crime hot-spot.
- $\mathcal{E}$ denotes the set of edges connecting the nodes, where an edge $e_{nm} \in \mathcal{E}$ between node n and a neighboring node m represents a road or street connecting two locations.
- t_{nm} denotes the average travel time for a patrol unit moving on the shortest route from node n to node m.
- w_n denotes the importance of node n in the patrol area.
- $h_n(\tau_n)$ denotes the SLF of node n.

III. Balanced Patrol Districting

In a patrol region, given Z patrol units available, we want to solve the patrol districting problem, i.e. the assignment of responsible district to each patrol unit. For the patrol districting, the balance of workloads among the patrol units is the primary concern in practices [8]. We use the average incident arrival rate $\bar{\lambda}_n$ as the expected workload requirement from node n. Since SLF can be considered as the probability distribution function of the time to occurrence of an incident, we obtain $\bar{\lambda}_n$ as

$$\bar{\lambda}_n = \frac{1}{\int_0^{\infty} h_n(\tau_n) d\tau_n}. \tag{1}$$

Let $\mathcal{N}_z$ be the set of nodes in the district z. Then workloads $\mathcal{L}_z$ in a district z is the sum of the workloads from all the nodes in the district, i.e. $\mathcal{L}_z = \sum_{n \in \mathcal{N}_z} \bar{\lambda}_n$.

In addition, each district must be contiguous and compact. A district is contiguous if it is possible to move from every node in the district to other node without crossing another district. Compactness means that the district is somewhat circular or square in shape rather than long and thin. Such shapes reduce average response travel time of a district. Formally, the patrol districting problem is described as follows: *Partition the patrol area $\mathcal{G}(\mathcal{N}, \mathcal{E})$ into Z districts, with the goal of minimizing the maximum workloads of the districts, subject to the constraints that each district is contiguous and compact.*

In fact, the patrol districting problem is similar as the balanced graph partitioning problem in graph theory, and the political districting problem in operation research. Solutions to these problems were extensively studied in [8], [9].

IV. Patrol Routing Strategy By ϵ-optimal horizon Approximation

After the stage of patrol districting, a patrol unit is dispatched in a district to conduct patrolling. An efficient patrol routing strategy is needed. The patrol routing problem is essentially a sequential decision making problem. This gives rise to the Markov Decision Process formulation. Due to the curse of dimensionality, we propose an efficient patrol routing strategy based on the $\epsilon-$optimal horizon approximation. For the resulting $\epsilon-$optimal horizon MDP, we introduce the Cross Entropy method to obtain the optimal solution.

A. MDP Formulation of the Patrol Routing Problem

The Patrol Routing MDP (PR-MDP) in a district is a 5-tuple $< \mathcal{G}', \mathcal{B}, \mathcal{A}, \Gamma, R >$, where

- $\mathcal{G}'(\mathcal{N}', \mathcal{E}')$ is the graph model for the patrol district.
- $\mathcal{B}$ is the set of security states. A security state $\mathbf{B} \in \mathcal{B}$ is a combination of the SLF's of all the nodes, which describes the security status of the patrol district. Formally, a security state $\mathbf{B} \triangleq (h_1(\tau_1), h_2(\tau_2), ..., h_{|\mathcal{N}'|}(\tau_{|\mathcal{N}'|}))$.
- $\mathcal{A}$ is the set of possible actions. An action $a \in \mathcal{A}$ is a node the unit chooses to patrol next. Due to topology constraint of the district, the patrol unit can only choose a neighboring node to patrol next. Moreover, when the patrol unit is currently locating in node n, the available actions $\mathcal{A}_n = \{m : e_{nm} \in \mathcal{E}', m \in \mathcal{N}'\}$.
- $\Gamma : \mathcal{B} \times \mathcal{A} \to \mathcal{B}$ is the state transition function. When given the current state $\mathbf{B}$, the state transition function Γ returns the new state $\mathbf{B}'$ after executing the action a, i.e. $\mathbf{B}' = \Gamma(\mathbf{B}, a)$. Given the security state $\mathbf{B} = (h_1(\tau_1), h_2(\tau_2), ..., h_{|\mathcal{N}'|}(\tau_{|\mathcal{N}'|}))$ and current locating node n, the patrol unit chooses a node $m \in A_n$ to patrol next. After executing the action, for node n and m, the updated idling times $\tau_n' = t_{nm}$ and $\tau_m' = 0$ respectively; for the other nodes $k \in \mathcal{N}' \backslash \{n, m\}$, the updated idling time $\tau_k' = \tau_k + t_{nm}$. In this case, the new security state $\mathbf{B}' = (h_1(\tau_1'), h_2(\tau_2'), ..., h_{|\mathcal{N}'|}(\tau_{|\mathcal{N}'|}'))$.
- $R : \mathcal{B} \times \mathcal{A} \to \mathbb{R}$ is the immediate reward function. When the patrol area is in security state $\mathbf{B}$ and an action a is executed, the reward $R(\mathbf{B}, a)$ is obtained. Let the weighting index vector $\mathbf{W} = (w_1, w_{2,...,} w_{|\mathcal{N}'|})$ denote the importance of the nodes in the district. Since the goal of patrolling is to make the district as secure as possible, the reward function is defined as the weighted sum of SLF's, i.e. $R(\mathbf{B}, a) \triangleq \mathbf{W}\mathbf{B}'^{\top} = \sum_{n=1}^{|\mathcal{N}'|} w_n h_n(\tau_n')$, where $\mathbf{B}' = \Gamma(\mathbf{B}, a)$ and $\top$ is the transpose operator.

Under the MDP framework, an policy π specifies an action $a = \pi(\mathbf{B})$ for any state $\mathbf{B} \in \mathcal{B}$. The objective of making the patrol district as secure as possible for the long run is equivalent to finding an optimal policy that maximizes the expected total discounted rewards over an infinite horizon. Suppose the sequential decision time epochs are $\{t_k\}_{k=0}^{\infty}$, the expected value for policy π starting from state $\mathbf{B}$ is defined as

$$V^{\pi}(\mathbf{B}) = E[\sum_{k=0}^{\infty} \gamma^k R(\mathbf{B}_k, a_k) | \mathbf{B}_0 = \mathbf{B}], \tag{2}$$

where $\mathbf{B}_k, a_k$ are the security state and action at k-th decision epoch respectively, and $0 < \gamma < 1$ is the discount factor balancing current reward and future return. The optimal policy π^* is obtained by maximizing $V^{\pi}(\mathbf{B})$, i.e.

$$\pi^* = \arg\max_{\pi} V^{\pi}(\mathbf{B}). \tag{3}$$

B. ϵ-Optimal Patrol Routing Strategy

It is known that under the optimal policy π^*, the value function $V^*(\mathbf{B}) = \max_{\pi} V^{\pi}(\mathbf{B})$ for each state $\mathbf{B} \in \mathcal{B}$ satisfies the Bellman optimality equation:

$$V^*(\mathbf{B}) = \max_a [R(\mathbf{B}, a) + \gamma V^*(\Gamma(\mathbf{B}, a))]. \tag{4}$$

When state space is finite, the number of Bellman optimality equations is finite. It is efficient to obtain the optimal policy π^* by the value iteration or policy iteration. However, for our case the state space $\mathcal{B}$ is high-dimensionally continuous. The number of resultant Bellman optimality equations is infinite, which makes the problem computationally intractable. To obtain efficient patrol routing strategy, we focus on developing approximate solutions. To proceed, we first establish some definitions.

Definition 1. *(ϵ-optimal policy) A policy π yields the value $V^{\pi}(\mathbf{B})$ that satisfies*

$$|V^{\pi}(\mathbf{B}) - V^*(\mathbf{B})| \leq \epsilon \tag{5}$$

for each $\mathbf{B} \in \mathcal{B}$ is called an ϵ-optimal policy.

The definition above gives error bound of performing approximate policies rather than the optimal policy. When ϵ is small, any ϵ-optimal policy is an efficient policy. For our case, the value function defined in (2) is a discounted function over the infinite horizon. As the stages of horizon k goes large, the value $\gamma^k R(\beta_k, a_k)$ becomes diminishing. The value function of PR-MDP over a sufficient long finite horizon is therefore a good approximation of the optimal value function $V^*(\mathbf{B})$.

Definition 2. *(ϵ-optimal horizon) Given an initial state $\mathbf{B}$, the ϵ-optimal horizon is a H-stages horizon over which the optimal value*

$$V_\epsilon^*(\mathbf{B}) \triangleq \max E[\sum_{k=0}^{H-1} \gamma^k R(\mathbf{B}_k, a_k)|\mathbf{B}_0 = \mathbf{B}] \tag{6}$$

satisfies

$$|V_\epsilon^*(\mathbf{B}) - V^*(\mathbf{B})| \leq \epsilon. \tag{7}$$

The ϵ-optimal horizon gives the required number of horizon stages to achieve the ϵ-optimal approximation. Let the immediate reward $R(\mathbf{B}, a)$ be bounded by R^*, i.e. $R^* = \sup_{\mathbf{B},a} R(\mathbf{B}, a)$. This bound can be obtained by setting the security state $\mathbf{B} = (1, 1, ..., 1)$ and evaluating all possible actions moving from a node to its neighboring node. We then have the following result:

Lemma 3. The number of horizon stages $H \geq \log_\gamma(\frac{\epsilon(1-\gamma)}{R^*})$ is an ϵ-optimal horizon.

Proof: For the optimal value over the infinite horizon, we have

$$\begin{aligned} V^*(\mathbf{B}) &= \max E[\sum_{k=0}^{\infty} \gamma^k R(\mathbf{B}_k, a_k)|\mathbf{B}_0 = \mathbf{B}] \\ &\leq \max E[\sum_{k=0}^{H-1} \gamma^k R(\mathbf{B}_k, a_k)|\mathbf{B}_0 = \mathbf{B}] + \sum_{k=H}^{\infty} \gamma^k R^* \\ &\leq V_\epsilon^*(\mathbf{B}) + \sum_{k=H}^{\infty} \gamma^k R^* \\ &= V_\epsilon^*(\mathbf{B}) + \frac{\gamma^H R^*}{1-\gamma}. \end{aligned} \tag{8}$$

The condition (7) holds when $\frac{\gamma^H R^*}{1-\gamma} \leq \epsilon$, i.e.

$$H \geq \log_\gamma(\frac{\epsilon(1-\gamma)}{R^*}). \tag{9}$$

■

We next propose the patrol routing strategy on the basis of the ϵ-optimal horizon approximation. For the sake of brevity, we first define the Q-function as:

$$Q(\mathbf{B}, a) = R(\mathbf{B}, a) + \gamma V_\epsilon^*(\Gamma(\mathbf{B}, a)) \tag{10}$$

where $V_\epsilon^*(\Gamma(\mathbf{B}, a))$ is the optimal value obtained by PR-MDP over an $\epsilon-$optimal horizon with the initial state $\Gamma(\mathbf{B}, a)$.

The optimal patrol routing strategy π_ϵ under the ϵ-optimal horizon approximation is then defined as: for each security state $\mathbf{B}$, the action $a^* = \arg\max_a Q(\mathbf{B}, a)$ is chosen. Given the initial state $\mathbf{B}_0$, the strategy π_ϵ yields the action $a_0 = \arg\max_a Q(\mathbf{B}_0, a)$ for the patrol unit. When the unit takes the action a_0, the new security state $\mathbf{B}_1 = \Gamma(\mathbf{B}_0, a_0)$ is realized. The patrol unit next takes the action $a_1 = \arg\max_a Q(\mathbf{B}_1, a)$, and the new state $B_2 = \Gamma(B_1, a_1)$ is realized. The procedure continues in this manner with the patrol route $\{a_k\}_{k=0}^{\infty}$ being generated. The performance of patrol routing strategy π_ϵ is guaranteed by the following theorem:

Theorem 4. *The patrol routing strategy π_ϵ is an ϵ-optimal policy.*

Proof: First, we denote the value by Q-function as $\bar{V}(\mathbf{B})$, i.e.

$$\begin{aligned} \bar{V}(\mathbf{B}) &= \max_a Q(\mathbf{B}, a) \\ &= \max_a [R(\mathbf{B}, a) + \gamma V_\epsilon^*(\Gamma(\mathbf{B}, a))]. \end{aligned} \tag{11}$$

$\bar{V}(\mathbf{B})$ is actually the optimal value of the PR-MDP over the $(H+1)$-stages horizon with initial state $\mathbf{B}$. It is always greater than the optimal value of the PR-MDP over the H-stages horizon with the same initial state, i.e.

$$V_\epsilon^*(\mathbf{B}) \leq \bar{V}(\mathbf{B}) = \max_a [R(\mathbf{B}, a) + \gamma V_\epsilon^*(\Gamma(\mathbf{B}, a))]. \tag{12}$$

Let $\hat{V}(\mathbf{B})$ be the value function by the patrol routing strategy π_ϵ for any $\mathbf{B} \in \mathcal{B}$. Suppose under the patrol strategy π_ϵ, a sequence of K state-action pairs $\{(\mathbf{B}_k, a_k)\}_{k=0}^{K}$ are generated, then

$$\hat{V}(\mathbf{B}_0) = \sum_{k=0}^{\infty} R(\mathbf{B}_k, a_k). \tag{13}$$

From (12), we have

$$\begin{aligned} V_\epsilon^*(\mathbf{B}_0) &\leq R(\mathbf{B}_0, a_0) + \gamma V_\epsilon^*(\mathbf{B}_1) \\ \gamma V_\epsilon^*(\mathbf{B}_1) &\leq \gamma R(\mathbf{B}_1, a_1) + \gamma^2 V_\epsilon^*(\mathbf{B}_2) \\ &\vdots \\ \gamma^K V_\epsilon^*(\mathbf{B}_K) &\leq \gamma^K R(\mathbf{B}_K, a_K) + \gamma^{K+1} V_\epsilon^*(\mathbf{B}_{K+1}), \end{aligned}$$

where $\mathbf{B}_{k+1} = \Gamma(\mathbf{B}_k, a_k)$, for $k = 0, ..., K$. Combining the inequalities above, we obtain

$$\begin{aligned} V_\epsilon^*(\mathbf{B}_0) \leq\ & R(\mathbf{B}_0, a_0) + \gamma R(B_1, a_1) + \cdots \\ & + \gamma^K R(B_K, a_K) + \gamma^{K+1} V_\epsilon^*(B_{K+1}). \end{aligned} \tag{14}$$

While $K \to \infty$, the last term in (14) $\gamma^{K+1} V_\epsilon^*(\beta_{K+1}) \to 0$, thus

$$V_\epsilon^*(\mathbf{B}_0) \leq \sum_{k=0}^{\infty} R(\mathbf{B}_k, a_k). \tag{15}$$

Since $\hat{V}(\mathbf{B}_0) = \sum_{k=0}^{\infty} R(\mathbf{B}_k, a_k)$, then

$$V_\epsilon^*(\mathbf{B}_0) \leq \hat{V}(\mathbf{B}_0), \tag{16}$$

From (7) we have

$$\begin{aligned} V^*(\mathbf{B}_0) - \hat{V}(\mathbf{B}_0) &\leq V^*(\mathbf{B}_0) - V_\epsilon^*(\mathbf{B}_0) \\ &\leq \epsilon. \end{aligned} \tag{17}$$

Similarly, we can prove for any $\mathbf{B}_k \in \mathcal{B}$, the following holds

$$V^*(\mathbf{B}_k) - \hat{V}(\mathbf{B}_k) \leq \epsilon.$$

which satisfies the definition of the ϵ-optimal policy. ■

C. Optimizing ϵ-Optimal Horizon PR-MDP by CE Method

To implement the proposed patrol routing strategy π_ϵ, the remaining problem is to obtain the optimal value $V_\epsilon^*(\mathbf{B})$ for a given state $\mathbf{B}$. It is equivalent to finding an optimal policy for the PR-MDP over a finite horizon. In this section, we propose the Cross Entropy (CE) method for optimizing this finite horizon PR-MDP.

1) Cross Entropy Method: The CE method was first used to estimate the probability of rare events. It was adapted for optimization with the observation that sampling around the optimum of a function is a rare event [10]. Due to its fast convergence and simplicity, the CE method has been successfully applied to a number of difficult optimization problems, including the maximal cut problem, the traveling salesman problem, and various kinds of scheduling problem [10]. To apply the CE method, the deterministic optimization problem must first be converted into a stochastic estimate problem. Then the following three steps are performed repeatedly until convergence:

1) Generate random data samples according to a parameterized random mechanism.
2) Select the elite samples according to a performance criterion.
3) Update the parameters of the random mechanism based on the selected elite samples for producing better new samples.

Formally, let $U(x)$ denotes the objective function. Suppose we wish to maximize $U(x)$ over the set $\mathcal{X}$ by the CE method. In Step 1, M random samples $X_1, ..., X_M$ are generated by the parameterized random mechanism $f(x, v)$, a probability density function with parameter v. In Step 2, the elite samples are selected according to the performance criterion, i.e. samples for which $U(x) \geq \eta$ are selected. In Step 3, the CE method uses minimization of the Cross Entropy (also known as Kullback-Leibler distance) between updated random mechanism and the important sampling distribution of the selected elite samples as the update criterion. According to [10], it is equivalent to solving:

$$\max_v \frac{1}{M} \sum_{i=1}^{M} I_{\{U(X_i) \geq \eta\}} \ln f(X_i, v), \tag{18}$$

where,

$$I_{\{U(X_i) \geq \eta\}} = \begin{cases} 1 & \text{If } U(X_i) \geq \eta \\ 0 & \text{Otherwise} \end{cases} \tag{19}$$

The CE method uses an iterative approach. In each iteration, the $\lfloor \rho M \rfloor$ best samples are selected as the elite samples, with a typical value of $0.01 \leq \rho \leq 0.03$. As shown in [11], the CE method guarantees convergence to the global optimum with proper parameters setting.

2) Cross Entropy Approach For Optimizing ϵ-Optimal Horizon PR-MDP: The state space $\mathcal{B}$ of the PR-MDP problem is high-dimensionally continuous. It is difficult to apply the CE method to find optimal policy mapping from states to actions directly. However, given a security state $\mathbf{B}$ and an action a, the transition state $\mathbf{B}' = \Gamma(\mathbf{B}, a)$ is deterministic. This immediately leads to:

Theorem 5. *For the PR-MDP over $H-$stages horizon with the initial state $\mathbf{B}$, the value $V_\epsilon^\pi(\mathbf{B}) \triangleq E[\sum_{k=0}^{H-1} \gamma^k R(\mathbf{B}_k, a_k) | \mathbf{B}_0 = \mathbf{B}]$ under a policy π can be achieved by acting according to the action sequence $\ell \triangleq \{a_0, a_1, ..., a_{H-1}\}$ only, regardless of the resultant states $\{\mathbf{B}_1, ..., \mathbf{B}_{H-1}\}$.*

This structure can be exploited to simplify the computation. To obtain the optimal value $V_\epsilon^*(\mathbf{B})$, we seek to find the optimal action sequence ℓ^* rather than derive the optimal policy π^*.

We next apply the CE method to obtain the optimal action sequence ℓ^*. To proceed, we first need to establish a random action sequence generation mechanism. Here we define H auxiliary $|\mathcal{N}'| \times |\mathcal{N}'|$ transition matrices $\{P^k : [p_{aa'}^k]\}_{k=0}^{H-1}$, where $p_{aa'}^k$ is the probability of choosing action a' in the $k-$th stage given the action in $(k-1)-$th stage is a. The action in (-1)-th stage is the initial locating node of the patrol unit. With the transition matrices, a random action sequence can be generated in the stage by stage manner. For initialization, we set each transition matrix as a uniform distribution among the available actions, i.e.

$$p_{aa'}^k = \begin{cases} \frac{1}{|\mathcal{A}_a|} & a' \in \mathcal{A}_a \\ 0 & a' \notin \mathcal{A}_a \end{cases} \tag{20}$$

This guarantees that all feasible action sequence can be generated.

After the sample action sequences are generated, the critical issue is the updating of the transition matrices, or solving the problem (18). Let $V(\ell)$ denote the value obtained by acting according to the action sequence ℓ. The following theorem gives the optimal update formula.

Theorem 6. *Given M sample action sequences $\ell_1, ..., \ell_M$, and the elite sample selection criterion as the indicator function $I_{\{V(\ell_i) \geq \eta\}}$, the optimal update formula minimizing the Cross Entropy between the updated transition matrices and the importance sampling distribution of the selected elite action sequences is*

$$p_{aa'}^k = \frac{\sum_{i=1}^{M} I_{\{V(\ell_i) \geq \eta\}} I_{\{g_{aa'}^k \in \ell_i\}}}{\sum_{i=1}^{M} I_{\{V(\ell_i) \geq \eta\}} \sum_{a'=1}^{|\mathcal{N}'|} I_{\{g_{aa'}^k \in \ell_i\}}}, \tag{21}$$

where $I_{\{g_{aa'}^k \in \ell_i\}}$ an indicator whether the action sequence ℓ_i contains the pattern that the action a' is chosen in the $k-$th stage given the action in previous stage is a.

Proof: For simplicity of notation, we denote the set of the transition matrices as $\mathbf{P} \triangleq \{P^k\}_{k=0}^{H-1}$. Using (18), the optimal update problem becomes:

$$\begin{aligned} \max_{\mathbf{P}} \quad & \tfrac{1}{M}\textstyle\sum_{i=1}^{M} I_{\{V(\ell_i)\geq\eta\}} \ln f(\ell_i,\mathbf{P}) && (22)\\ \text{s.t.} \quad & \textstyle\sum_{a'} p_{aa'}^{k} = 1 && \forall a = 1,...,|\mathcal{N}^{'}|;\\ & && k = 0,...,H-1. \end{aligned}$$

Since $f(\ell_i, \mathbf{P})$ is the probability distribution function that the action sequence ℓ_i is generated by the transition matrices $\mathbf{P}$, we have

$$\ln f(\ell_i,\mathbf{P}) = \sum_{k=0}^{H-1}\sum_{a=1}^{|\mathcal{N}^{'}|}\sum_{a'=1}^{|\mathcal{N}^{'}|} I_{\{g_{aa'}^{k}\in\ell_i\}} \ln p_{aa'}^{k}. \qquad (23)$$

This function is concave as it is a sum of logarithm functions. Therefore, the objective function in (22) is concave, and constraint set is convex. This guarantees that any local maximum is also the global maximum. Using the Lagrange multipliers $\{\mu_a^k\}_{a=1,...,|\mathcal{N}^{'}|}^{k=0,...,H-1}$ for solving the maximization problem in (22), we have

$$\max_{\mathbf{P}} \frac{1}{M}\sum_{i=1}^{M} I_{\{V(\ell_i)\geq\eta\}} \ln f(\ell_i,\mathbf{P}) + \sum_{k=0}^{H-1}\sum_{a=1}^{|\mathcal{N}^{'}|} \mu_a^k \Big(\sum_{a'=1}^{|\mathcal{N}^{'}|} p_{aa'}^{k} - 1\Big). \qquad (24)$$

Differentiating (24) with respect to $p_{aa'}^k$ yields

$$\frac{1}{M}\sum_{i=1}^{M} I_{\{V(\ell_i)\geq\eta\}} I_{\{g_{aa'}^{k}\in\ell_i\}} + \mu_a^k p_{aa'}^k = 0. \qquad (25)$$

Summing over $a' = 1, ..., |\mathcal{N}^{'}|$ gives us

$$\mu_a^k = -\frac{1}{M}\sum_{i=1}^{M} I_{\{V(\ell_i)\geq\eta\}} \sum_{a'=1}^{|\mathcal{N}^{'}|} I_{\{g_{aa'}^{k}\in\ell_i\}}, \qquad (26)$$

which substituting μ_a^k in (25) gives us the optimal solution (21). ■

We note that instead of updating the transition matrices with (21) directly, a smoothing procedure combining the updates between two successive iterations is more efficient, i.e.

$$\tilde{p}_{aa'}^{k}(i) = \alpha p_{aa'}^{k}(i) + (1-\alpha)\tilde{p}_{aa'}^{k}(i-1), \qquad (27)$$

where $p_{aa'}^k(i)$ is from (21), and $\tilde{p}_{aa'}^k(i-1)$ denotes the smoothing update in the last iteration. This step is necessary to prevent the algorithm from getting stuck in a local optimum when $p_{aa'}^k(i)$ is zero. Values of $0.7 \leq \alpha \leq 0.9$ are found to give the best results [10]. We conclude the CE method for optimizing ϵ-optimal horizon PR-MDP in the following algorithm:

Algorithm 1 CE Method for Optimizing $\epsilon-$Optimal Horizon PR-MDP

```
1) Set the initial transition matrices as the uniform distributions.
2) Generate $M$ sample action sequence $\ell_1, ..., \ell_M$ by random action sequence generation mechanism.
3) Determine $\lfloor \rho M \rfloor$ elite samples.
4) Use the selected elite samples to update the transition matrices according to formula (21).
5) Apply the equation (27) to smooth the transition matrices.
6) Repeat from step 2 until convergence.
```

To sum up, we describe $\epsilon-$optimal patrol routing strategy for the security measure by SLF in Algorithm 2.

Algorithm 2 $\epsilon-$Optimal Patrol Routing Strategy For SLF

```
/*Given the patrol district graph $\mathcal{G}^{'}$, the initial security state $\mathbf{B}_0$, the starting patrol node $n_0$ and the planning horizon $[T_0, T]$, the following two procedures are employed:
*/
1) Procedure Initialization:
    Obtain the reward upper bound $R^*$;
    Calculate the $\epsilon-$optimal horizon $H$;
    Initialize the parameter settings for CE method;
    Set running time $t = T_0$;
    Set sequence Index $k = 0$;
    Set initial action $a_{-1} = n_0$;
End Procedure

2) Procdure Patrol Routing:
    While $t \leq T$:
        For each available action $a \in \mathcal{A}_{a_{k-1}}$:
            Obtain the new transition state $\mathbf{B}^{'} = \Gamma(\mathbf{B}_k, a)$;
            Calculate the optimal PR-MDP value $V_\epsilon^*(\mathbf{B}^{'})$
            by the CE method;
        End For
        Choose the next action $a_k = \arg\max_a Q(\mathbf{B}_k, a)$;
        Update the security state $\mathbf{B}_{k+1} = \Gamma(\mathbf{B}_k, a_k)$;
        Update the running time $t = t + t_{a_{k-1},a_k}$;
        Update the sequence Index $k = k + 1$;
    End While
End Procedure
```

V. Randomized Patrol Strategy

The aim of police patrol is to provide the public with a sense of security and discourage potential crimes. One good way of discouraging crime is to adopt a randomized patrol strategy, so that criminals can never be sure when a patrol might arrive on the scene. The use of randomized police patrols has been one of the key factors in the significant fall in the crime rate in New York city in recent years [12].

Similar with the randomized exploration method in the Reinforcement Learning [13], we randomize the patrol action

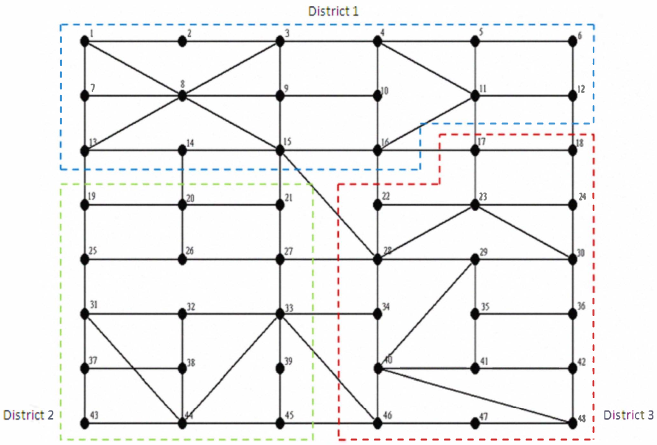

Fig. 2. Patrol Area Graph

on the basis of the Boltzmann distribution. Moreover, given the security state $\mathbf{B}$, the probability that an action $a \in \mathcal{A}$ is chosen is determined by

$$q(a|\mathbf{B}) = \frac{e^{\sigma Q(\mathbf{B},a)}}{\sum_{a' \in \mathcal{A}} e^{\sigma Q(\mathbf{B},a')}}, \tag{28}$$

where σ is the coefficient controls the greediness of the action selection procedure: if $\sigma = 0$, action selection is purely random, and if $\sigma \to \infty$ it approaches the greedy action selection, the action $a^* = \arg\max_a Q(\mathbf{B}, a)$ is always chosen.

VI. Numerical Results

A. Simulation Setup

We run the simulation experiments on a PC with a 3.2GHz CPU. The program is written in C language. The patrol area graph in the experiment is depicted in the Fig 2. For a node n, we let SLF $h_n(\tau_n) = a_n^{\tau_n}$ where the index τ_n is the idling time and the base a_n is chosen randomly from the set $\{0.1, 0.2, ..., 0.9\}$ and the weight $w_n = 1$. The travel time from a node to a neighboring node is assumed be 1 time unit. For the PR-MDP, we set the the discounted factor $\gamma = 0.8$ and the approximation horizon $H = 20$, which guarantees the error bound $\epsilon \leq 0.5$. For the CE method of optimizing $\epsilon-$optimal horizon PR-MDP, the number of sample action sequences $N = 5000$ are generated in each iteration, the fraction $\rho = 0.02$ of elite samples are chosen for updating, and the smoothing factor $\alpha = 0.7$ is used.

B. Patrol Districting

We apply the method addressed in [9] for the patrol districting. We partition the patrol area graph into $Z = 3$ districts. From (1), the workload from a location n is $\bar{\lambda}_n = \frac{1}{\int_0^\infty a_n^{\tau_n} d\tau_n} = -\ln a_n$. Subject to the contiguity and compactness requirement, the balanced workload partitioning solution is shown in Fig. 2. To evaluate the solution, we define the workload deviation in a district z as $\Delta_z = \frac{|\mathcal{L}_z - \bar{\mathcal{L}}|}{\bar{\mathcal{L}}}$ where $\bar{\mathcal{L}} = \frac{\sum_{n=1}^{|\mathcal{N}|} \bar{\lambda}_n}{Z}$ is the mean workloads. The table in sequel lists the workloads deviations of the districts, which shows the small deviation property of the solution.

District z	1	2	3
Workload Deviation Δ_z	1.2%	0.9%	2.1%

C. Patrol Routing

We implement the patrol routing strategy π_ϵ under $\epsilon-$optimal horizon approximation in the district 1. The patrolling is assumed to start from node 1 with the initial security state $\mathbf{B}_0 = (1, 1, ..., 1)$. We run the patrol routing strategy π_ϵ with the number of total horizon stages equals to 50,100,150,200,250 and 300 time units respectively, and calculate the accumulated rewards obtained. To evaluate the performance of the strategy, we compare it with the following two strategies:

- Greedy strategy: the patrol unit always choose the next node with the highest patrol reward. This is the patrol strategy with short-term reward maximization.
- Hamilton strategy: the patrol unit visits all the nodes according to the shortest Hamilton path. This strategy was proposed in [3] and [4], with the goal to minimize average idling time of the nodes.

We plot the rewards obtained by the three strategies in Fig 3. We find that the rewards grow linearly as the running horizon increases. Among the strategies, the patrol routing strategy π_ϵ is the best performing one. Its performance gain is $13\% \sim 18\%$ over the Hamilton strategy, and $25\% \sim 40\%$ over the Greedy Strategy. Since the objective of the strategy π_ϵ is to maximize the rewards obtained for the long run, the performance gain grows as the total running time units increase.

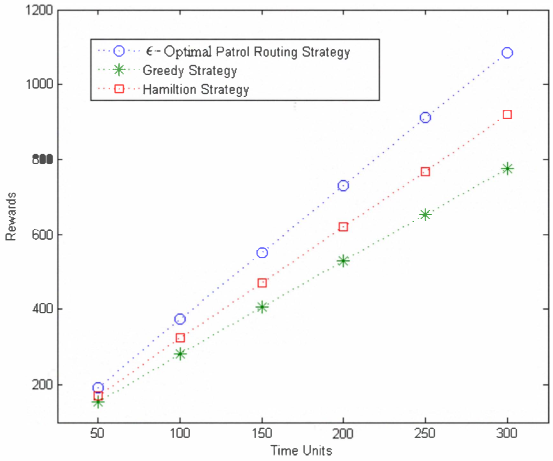

Fig. 3. Different Patrol Routing Strategies

VII. Conclusion

In this paper we presented a new general security measure by defining the security level function. Based on this, we proposed the balanced patrol districting solution to the multiple units assignment problem. For the patrol routing problem, we first formulated the patrol routing process as a Markov decision process, and proposed an efficient patrol routing strategy under the $\epsilon-$optimal horizon approximation. Numerical studies demonstrated that the strategy has significant performance improvement over the Hamilton strategy and the Greedy Strategy for the security measure by SLF. In addition, as the randomness is an important factor for practices, we designed the randomized patrol routing strategy on the basis of the randomized exploration method in the Reinforcement Learning.

An interesting and challenging extension of our approach is to consider the case where there are multiple patrol units in a single patrol district. The cooperation among the patrol units would be a fascinating topic for future study.

References

[1] S. Ruan, C. Meirina, F. Yu and K.R. Pattipati, Patrolling in a stochastic environment. *10th Intl. Command and Control Research and Tech*, 2005.

[2] Tony H. Grubesic, and Elizabeth A. Mack. Spatio-Temporal Interaction of Urban Crime. *The Journal of the Quantative Criminology*, vol. 24, pp.285šC306, 2008.

[3] Chevaleyre Y., Sempe F. and Ramalho, A theoretical analysis of multiagent patrolling strategy. *Proc. of AAMAS*, 2004.

[4] T. Sak, J Wainer and S.K. Goldenstein, Probabilistic multiagent patrolling. *Proc. of 19th Brazilian Symposium on Artificial Intelligence*, 2008.

[5] K Andreev and H Racke. Balanced graph partitioning. *Theory of Computing Systems*, 2006.

[6] P. Paruchuri, J.P. Pearce, J. Marecki and M. Tambe, Playing games for security: An efficient exact algorithm for solving bayesian stackelberg games. *Proc. of the 7th Intl. Joint Conference*, 2008.

[7] Ching-Fan Chung, Peter Schmidt and Ana D. Witte. Survival analysis: A survey. *Journal of Quantitative Criminology*, Vol. 7, No. 1, 1991.

[8] Steven J., Shoou-Jiun Wangb, Rajan Battac and Christopher M. Rump. A simulated annealing approach to police district design. *Computer and Operation Research.* Vol. 29, pp. 196-213, 2002.

[9] R.S. Garfinkel and G.L. Nemhauster. Optimal Political Districting by Implicityly Enumeration Techniques. *Management Science*, Vol. 16, pp. 495-508, 1970.

[10] R. Rubinstein and D. P. Kroese. The Cross-Entropy method: An unified approach to Combinatorial Optimization, Monte-Carlo Simulation, and Machine Learning. *Information Science & Statistics, Springer*, 2004.

[11] L. Margolin, On the Convergence of the Cross Entropy method. *Annals of Operations Research, Springer*, 2005.

[12] N. Billante, The Beat Goes On: Policing for Crime Prevention. *http://www.cis.org.au/IssueAnalysis/ia38/ia38.htm,* 2003.

[13] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. *The MIT Press*, 1998.