Dynamically Maintaining Infrastructure Networks

Jefferson Huang
Assistant Professor
Operations Research Department
Naval Postgraduate School

In Collaboration With:
Vincent Wickel (Lieutenant, German Army)
Daniel Eisenberg (Assistant Prof. & Director, NPS Center for Infrastructure Defense)
David Alderson (Professor & Executive Director, NPS CID)

INFORMS Annual Meeting
Seattle, WA
23 October, 2024

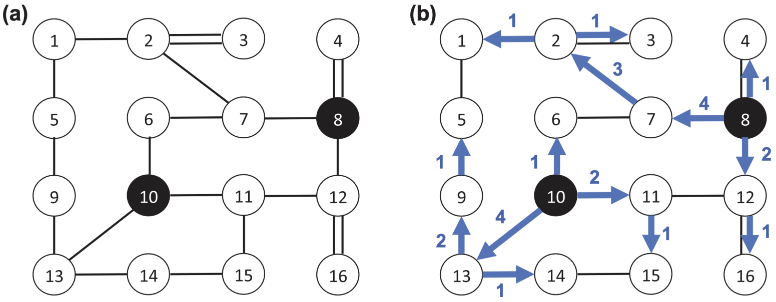


Fig. 1. A notional infrastructure system. (a) A white circle (node) represents a location with demand equal to one barrel of fuel. A black circle (node) represents a location with supply equal to 10 barrels. Each link is bidirectional, has a fuel flow capacity of 15 barrels, and has per-barrel transit cost of $1. The penalty for unsatisfied demand per node is $10 per barrel. Nodes 3, 4, and 16 each have two (parallel, redundant) connections to the rest of the network. This network has been built to be N−1 reliable, meaning that the loss of any single link does not disconnect any node. (b) Shows baseline flows corresponding to a minimum-cost flow solution, which results in a total cost of $25.

Source: D. L. Alderson, G. G. Brown, and W. M. Carlyle, Operational Models of Infrastructure Resilience, *Risk Analysis* 35(4), 2015.

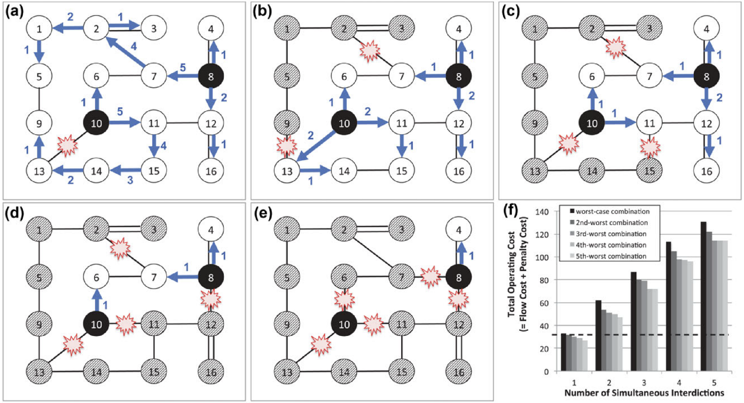


Fig. 3. Worst-case simultaneous interdictions. (a) The worst-case single interdiction is of link [10, 13], resulting in a total cost of 33. In this case, the flow cost increases but all nodes are still served. (b) The worst-case simultaneous two-link interdiction is of links [2, 7] and [9, 13], which denies nodes 1, 2, 3, 5, and 9 (now shaded) any flow. The total cost is 62 (=12 + 50), most of which is unmet demand penalty cost. (c) The worst-case simultaneous three-link interdiction is of links [2, 7], [10, 13], and [11, 15], resulting in a total cost of 87 (=7 + 80). (d) The worst-case simultaneous four-link interdiction is of links [2, 7], [8, 12], [10, 11], and [10, 13], resulting in a total cost of 113 (=3 + 110). (e) The worst-case simultaneous five-link interdiction is of links [6, 10], [7, 8], [8, 12], [10, 11], and [10, 13], resulting in a total cost of 131 (=1 + 130). (f) The worst-case (rank 1) attack for 1–5 simultaneous interdictions increases approximately linearly. The second-worst (rank 2) through fifth-worst (rank 5) attacks do less damage, but all are significantly worse than the baseline (no interdiction) case that has operating cost 25.

Source: D. L. Alderson, G. G. Brown, and W. M. Carlyle, Operational Models of Infrastructure Resilience, *Risk Analysis* 35(4), 2015.

Markov Decision Process (MDP) Model Description

Infrastructure Network Data

$\mathcal{N}$ = node set $\qquad$ $\mathcal{A}$ = arc set

d_n = demand at $n \in \mathcal{N}$

p_n = per-unit demand shortfall cost at n

c_{ij} = unit flow cost for $(i,j) \in \mathcal{A}$

q_{ij} = unit penalty for flow on (i,j) if broken

u_{ij} = flow capacity for (i,j)

r_{ij} = cost to replace (i,j)

ω_{ij} = prob. that (i,j) will break by next time step, if just replaced

ϕ_{ij} = increase in break prob. of (i,j) per time step not broken

K = max. number of edges that can be replaced at once

MDP Model

$$\textit{state} = \mathbf{s} = (\mathbf{x}, \mathbf{b}) \in \{0,1\}^{|\mathcal{A}|} \times [0,1]^{|\mathcal{A}|} =: \mathbb{S}$$

$$\textit{action} \in \left\{ \mathbf{a} \in \{0,1\}^{|\mathcal{A}|} \;\middle|\; \textstyle\sum_{(i,j)} a_{ij} \leqslant K \right\} =: \mathbb{A}$$

The *one-step costs* have the form

$$c(\mathbf{s}, \mathbf{a}) = f(\mathbf{x}) + r(\mathbf{a})$$

The *transition probabilities* $p(\mathbf{s}'|\mathbf{s}, \mathbf{a})$ are products of transition probabilities for individual arcs.

One-Step Cost Function

The *one-step cost function* is

$$c(\mathbf{s},\mathbf{a}) = c((\mathbf{x},\mathbf{b}),\mathbf{a}) = f(\mathbf{x}) + r(\mathbf{a})$$

The **flow cost**, given the current broken/non-broken status $\mathbf{x} \in \{0,1\}^{|\mathcal{A}|}$ of the arcs, is

$$f(\mathbf{x}) := \min_{Y,U} \left\{ \sum_{(i,j)\in\mathcal{A}} \left[(c_{ij} + q_{ij}x_{ij})\, Y_{ij} + (c_{ji} + q_{ji}x_{ij})\, Y_{ji} \right] + \sum_{n\in\mathcal{N}} p_n U_n \;\middle|\; \begin{array}{l} \displaystyle\sum_{(n,j)\in\mathcal{A}} Y_{nj} - \sum_{(j,n)\in\mathcal{A}} Y_{in} - U_n \leqslant d_n \forall n \in \mathcal{N} \\ 0 \leqslant Y_{ij} + Y_{ji} \leqslant u_{ij} \quad \forall (i,j) \in \mathcal{A} \\ U_n \geqslant 0 \quad \forall n \in \mathcal{N} \end{array} \right\}$$

The **replacement cost**, given the indicator vector $\mathbf{a} \in \{0,1\}^{|\mathcal{A}|}$ of arcs to be replaced, is

$$r(\mathbf{a}) := \sum_{(i,j)\in\mathcal{A}} r_{ij} a_{ij}$$

Transition Probabilities

The *transition probabilities* have the form

$$p(\mathbf{s}'|\mathbf{s},\mathbf{a}) = \prod_{(i,j)\in\mathcal{A}} \rho\left((x'_{ij}, b'_{ij}) \mid (x_{ij}, b_{ij}), a_{ij}\right)$$

where the arc-wise transition probabilities $\rho\left((x'_{ij}, b'_{ij}) \mid (x_{ij}, b_{ij}), a_{ij}\right)$ are defined by:

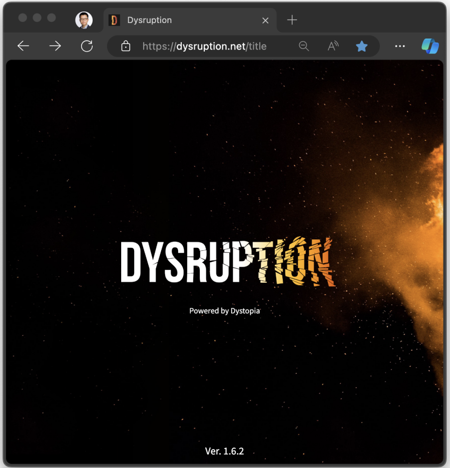
Current State		Action	Next State		Transition Probability
x_{ij}	b_{ij}	a_{ij}	x'_{ij}	b'_{ij}	$\rho\left((x'_{ij}, b'_{ij})\vert(x_{ij}, b_{ij}), a_{ij}\right)$
1	1	0	1	1	1
1	1	1	0	ω_{ij}	1
0	b_{ij}	0	1	1	b_{ij}
0	b_{ij}	0	0	$\min\{b_{ij} + \phi_{ij}, 1\}$	$1 - b_{ij}$
0	b_{ij}	1	0	ω_{ij}	1

Some Observations

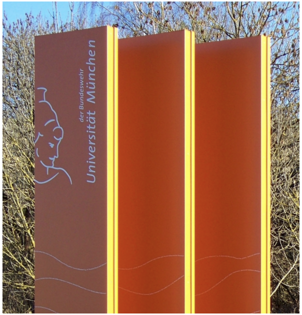
- For a single arc, this is a classic sequential replacement problem; see e.g., Derman & Sacks (1960), Derman (1963), and Taylor (1965).
- The arcs deteriorate according to independent Markov chains.
- The one-step costs

$$c((\mathbf{x}, \mathbf{b}), \mathbf{a}) = f(\mathbf{x}) + r(\mathbf{a})$$

  depend non-linearly on the indicator vector $\mathbf{x} \in \{0, 1\}^{|\mathcal{A}|}$ of broken arcs.



Game developed by the NPS Center for Infrastructure Defense (Documentation)



Our Master's thesis advisee, LT Vincent Wickel (University of the German Armed Forces, Graduated Dec 2023), applied Q-Learning to find approximately optimal policies for the MDP.

Q-Functions

For each state $\mathbf{s}$, let $v_*(\mathbf{s})$ be the optimal infinite-horizon expected discounted cost with discount factor $\gamma \in [0, 1)$. The *Q-function* is

$$Q(\mathbf{s}, \mathbf{a}) := c(\mathbf{s}, \mathbf{a}) + \gamma \sum_{\mathbf{s}' \in \mathbb{S}} p(\mathbf{s}'|\mathbf{s}, \mathbf{a}) v_*(\mathbf{s}') \qquad \text{for} \quad (\mathbf{s}, \mathbf{a}) \in \mathbb{S} \times \mathbb{A}.$$

Given the Q-function, any policy $\varphi : \mathbb{S} \to \mathbb{A}$ satisfying

$$\varphi(\mathbf{s}) \in \arg\min_{a \in \mathbb{A}} Q(\mathbf{s}, \mathbf{a}) \qquad \forall\ \mathbf{s} \in \mathbb{S}$$

is optimal.

Q-Learning approximates the Q-function via sampled states and actions.

Tabular Q-Learning (Watkins, 1989)

Algorithm 1 Tabular Q-Learning

Require: Learning Rate Schedule $\alpha_1, \alpha_2, \cdots \in (0,1]$, Number of Episodes N, Episode Length T, $\epsilon \in [0,1]$

```
Initialize Q̂(s, a) = 0 for each state s ∈ 𝕊 and a ∈ 𝔸.
for each episode n = 1, ..., N do
    Select an initial state s_1 ∈ 𝕊.
    for each decision epoch t = 1, ..., T do
        Select an ε-greedy action a_t ∈ 𝔸.
        Select the next state s_{t+1} according to the probability distribution p(·|s_t, a_t)
        Set Q̂(s_t, a_t) = (1 − α_t)Q̂(s_t, a_t) + α_t [c(s_t, a_t) + γ min_{a∈𝔸} Q̂(s_{t+1}, a)]
    end for
end for
```

$$\hat{Q}(\mathbf{s}_t,\mathbf{a}_t) = (1-\alpha_t)\hat{Q}(\mathbf{s}_t,\mathbf{a}_t) + \alpha_t\left[c(\mathbf{s}_t,\mathbf{a}_t) + \gamma \min_{\mathbf{a}\in\mathbb{A}} \hat{Q}(\mathbf{s}_{t+1},\mathbf{a})\right]$$

- For our MDP, the number of state-action pairs grows exponentially with the number of arcs.

Fitted Q-Learning (Gordon, 1995)

Algorithm 2 Fitted Q-Learning

Require: Learning Rate Schedule $\alpha_1, \alpha_2, \cdots \in (0, 1]$, Number of Episodes N, Episode Length T, $\epsilon \in [0, 1]$

Initialize the parameters to θ so that $\hat{Q}(\mathbf{s}, \mathbf{a}; \theta_0) \approx 0$ for all $(\mathbf{s}, \mathbf{a}) \in \mathbb{S} \times \mathbb{A}$.
for each episode $n = 1, \ldots, N$ **do**
    Select an initial state $\mathbf{s}_1 \in \mathbb{S}$.
    for each decision epoch $t = 1, \ldots, T$ **do**
        Select an ϵ-greedy action $\mathbf{a}_t \in \mathbb{A}$.
        Select the next state $\mathbf{s}_{t+1}$ according to the probability distribution $p(\cdot|\mathbf{s}_t, \mathbf{a}_t)$
        Set $\theta = \theta - \alpha_t \left(\hat{Q}(\mathbf{s}_t, \mathbf{a}_t; \theta) - c(\mathbf{s}_t, \mathbf{a}_t) - \gamma \min_{\mathbf{a} \in \mathbb{A}} \hat{Q}(\mathbf{s}_{t+1}, \mathbf{a}; \theta) \right) \nabla_\theta \hat{Q}(\mathbf{s}_t, \mathbf{a}_t; \theta)$.
    end for
end for

- We used feed-forward neural networks (Riedmiller, 2005) for the Q-function approximation $\hat{Q}(\mathbf{s}, \mathbf{a}; \theta)$.

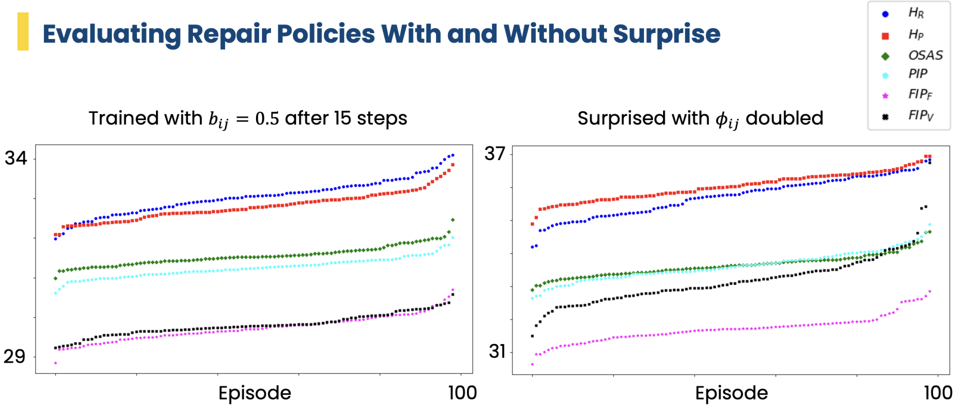
Contributions

- Implemented simulation environment for the notional infrastructure system from Alderson, Brown, and Carlyle (2015).
- Using this simulation environment, applied neural fitted Q-learning to compute an approximately optimal policy.
- Compared the Q-learning-based policy with some baselines:
  1. **Random:** Pick a random subset of broken arcs to replace.
  2. **Flow-Based:** Replace the arcs that carry the most flow under min-cost flows for current network.
  3. **One-Step Improvement:** Perform an approximate one-step policy improvement on greedy policy.
- Evaluated the robustness of these policies to “surprise” increases in failure rates (e.g., due to climate change).

Results

- The Q-learning-based policy selects groups of arcs to replace that are consistent with worst-case simultaneous interdiction solutions.
- There is a clear separation in performance between the heuristic (Random, Flow-Based) policies, the policy based on one-step policy improvement, and the Q-learning-based policy.
- Knowing the (deterioration) state of the arcs can make a big difference when the failure rates change unexpectedly.

Evaluating Repair Policies With and Without Surprise



Policy performance given trained failure rates

- Heuristic policies perform the same
- $OSAS$ and PIP perform the same
- FIP_F and FIP_V perform the same

Policy performance given surprise failure rates

- $OSAS$, PIP, and FIP_V have similar performance
- FIP_F outperforms all other policies

Source: D. A. Eisenberg, *Towards Models for Managing (Climate) Surprise in Infrastructure Systems*, Applied Math Colloquium, University of Arizona, Feb 2024.

Next Steps

- More computational studies (e.g., try different neural network architectures).
- Rewrite the simulation environment, e.g., as a Gymnasium environment.
- Try more modern (e.g., robust, risk-sensitive) reinforcement learning methods
- Study other types of infrastructure networks (e.g., water distribution networks)
- Identify useful structural properties of the MDP.