Health Care Operations Management: Ambulance operations service

Amir A. Nasrollahzadeh

Clemson University

snasrol@g.clemson.edu

Fall 2018

Overview

1. Introduction
2. Ambulance operations management
   - Deterministic models
   - Stochastic models
3. Ambulance location and operation models
   - Maximal covering
   - Bernoulli trial/Binomial experiment
   - Maximal expected covering
   - M/M/m queues
   - Average response time model
4. Ambulance dynamic models
5. Exercises
   - Class works
   - Project

Introduction

Emergency medical services (EMS) provide out-of-hospital acute medical care and transport the sick or injured to hospitals.

Factors such as

- increased nonemergency calls, which by law require an ambulance be dispatched,
- insufficient funding,
- intense traffic, and growing densely populated areas

put pressure on EMS providers to use limited resources to achieve operation targets set by municipalities or contracts.

Introduction

Operations targets

- Response time: The amount of time that an ambulance takes to arrive at the scene of a call once the call is received.
- Coverage: In EMS literature, a call is covered if its response time is below a certain threshold.

For example,

- The U.S. National Fire Protection Association suggests that 90% of emergency medical calls must be reached by a first responder within four minutes, followed by an advanced life support response within eight minutes.

Ambulance operations management

The ambulance operations management problem consists of several different decision making processes:

- Location models: Where should ambulances be parked? Where are the optimal places for ambulance stations?
- Logistic models: Can we assign several calls to one ambulance? How the ambulance should reach its destination?
- Operation models: Which ambulance to dispatch? Where to redeploy the ambulance after serving the call?
- Scheduling models: How many ambulances should be available at different times of a day? What is the procedure after they receive a call?

Ambulance operations management

In all ambulance operations models, the objective function could be one of the following:

- Max. coverage level,
- Min. fraction of late calls,
- Min. response times,
- Min. the number of ambulances,
- Min. the number of stations,
- Max. ambulance utility.

Some of these objectives are positively correlated. For example, maximizing coverage level generally results in lowering fraction of late calls and response times. However, in minimizing the number of ambulances, maintaining a certain level of coverage or insuring a ceiling for response times may be as important.

Deterministic static models

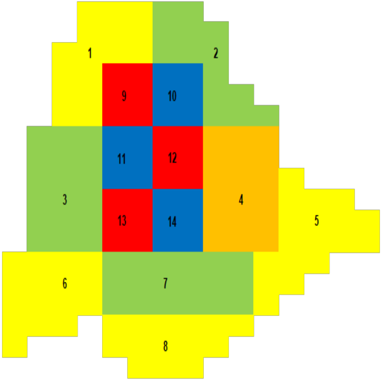
- Data and parameters are fixed, deterministic and given.
  - Ex. Demand is fixed and given.
- The procedures are fixed and static.
  - Ex. When a call arrives always dispatch the closest ambulance.
- The solutions (policies) are static.
  - Ex. The ambulances idle positions do not change with time.
- Easy to follow but not reliable in an ever changing environment.

Stochastic dynamic models

- Data and parameters are probabilistic and change over time.
  - Ex. Calls arrive according to a Poisson process.
- Procedures are dynamic and dependent on the state of system.
  - Ex. Allow both closest and not the closest dispatch as actions.
- The solutions (policies) are dynamic.
  - Ex. Sometime it is optimal to dispatch a farther ambulance.
- The model is able to consider future events and modify actions.
  - Ex. Maybe busy portions of days are anticipated and ambulances are relocated accordingly.

Maximal covering model

Mecklenburg County is divided into 14 zones for call arrivals and ambulance parkings. Call arrival in each zone is denoted by h. A call is covered if an ambulance can arrive at its location within 20 minutes of its reception. Find the optimal locations of ambulances such that the coverage level in the county is maximized. The county has only 5 ambulances.



Maximal covering model: Formal formulation

Parameters

Call zones: $C := \{1, 2, \ldots, 14\}$

Amb. zones: $A := \{1, 2, \ldots, 14\}$,

Demand: $h_i, \qquad \forall i \in C$,

Number of Ambs: 5,

Coverage: $a_{ij} = \begin{cases} 1 & \text{if call } i \text{ is covered by ambulance } j, \\ 0 & \text{otherwise}, \qquad \forall j \in A, \forall i \in C. \end{cases}$

Decision variables

$$x_j = \begin{cases} 1 & \text{if an ambulance is places in zone } j, \\ 0 & \text{otherwise.} \end{cases} \qquad \forall j \in A.$$

$$y_i = \begin{cases} 1 & \text{if call zone } i \text{ is covered by an amb.}, \\ 0 & \text{otherwise.} \end{cases} \qquad \forall i \in C.$$

Maximal covering model: Formal formulation

Model

Objective function: Maximize covered demand.

$$\max \quad \sum_{i \in D} h_i \, y_i,$$

Subject to,

$$\begin{array}{lll} \text{Limited resources:} & \sum_{j \in F} x_j \leq 5, & \\ \text{Coverage on zone i:} & \sum_{j \in F} a_{ij} x_j \geq y_i, & \forall i \in D, \\ & x_j \in \{0,1\} & \forall j \in F, \\ & y_i \in \{0,1\} & \forall i \in D. \end{array}$$

For more details, refer to the facility location slides.

Maximal covering model: AMPL

The AMPL code is discussed in facility location slides. Download the “max-covering-mecklenburg.txt” file from CANVAS, create AMPL model and data files, and run the code. Report the results.

Maximal covering model

- Q1: Suppose the goal is to keep a certain coverage level. How can one use this formulation to see how many ambulances are needed?
- Q2: What if one ambulance becomes busy serving a call? What happens to coverage? Is there a way to address this issue?

Bernoulli trial

The availability of an ambulance can be **approximated** by a Bernoulli trial.

Definition

A trial with exactly two possible outcomes, “Success” or “failure” in which the probability of success or failure remains the same every time the experiment is repeated.

Suppose “failure” denotes the event in which an ambulance is busy. The probability of failure is known and given, $q = 0.3$. Obviously, the probability of success is $1 - q = 0.7$.

Binomial experiment

The availability of k ambulances out of a total P is **approximated** by repeated Bernoulli trials, i.e., a Binomial experiment.

Definition

An experiment consisting of repeated independent Bernoulli trials. The probability of success and failure does not change during the experiment.

Binomial experiment

Parameters

- M: Total number of ambulances,
- k: Number of busy ambulances,
- q: probability that one ambulance is busy at any time.

Probabilities

- Probability that exactly k ambulances are busy

$$\underbrace{f(k, M, q)}_{\text{probability mass function (p.m.f.)}} = \binom{M}{k} q^k (1-q)^{M-k}$$

Binomial experiment

Probabilities

- Probability that at least k ambulances are busy

$$1 - prob.\{\text{at most } M-k \text{ ambulances are available}\}$$

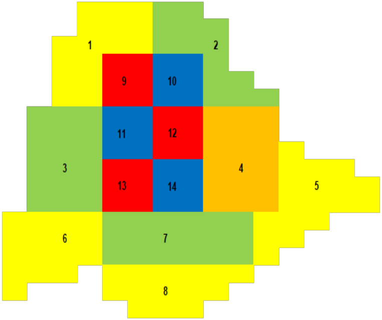
- Probability that at most $M-k$ ambulances are available

$$\boxed{F(M-k, M, q)} = \sum_{i=0}^{M-k} \binom{M}{i} q^{M-i} (1-q)^{i}$$

cumulative distribution function (c.d.f.)

Maximal expected covering model

Find the optimal locations of M ambulances such that expected demand coverage is maximized. An area is covered if and only if two conditions are satisfied: A number of ambulances should be located within 20 minutes of the demand, and at least one of these ambulances should be available. Demand (call arrival), probability that an ambulance is busy, travel time matrix, and max. number of ambulances are given. (Hint: To write the objective, evaluate the expected increase in coverage when an available ambulance is added to cover an area. Use Binomial distribution to model the availability of ambulances.



Maximal expected covering model

Data and parameters

- Demand zones: $D := \{1, 2, \ldots, 14\}$
- Ambulance locations: $A := \{1, 2, \ldots, 14\}$
- Max number of ambulances: M
- Call arrival rate at location $i \in D$: h_i
- Probability that an ambulance is busy: q
- Coverage matrix:

$$a_{i,j} = \begin{cases} 1 & \text{if demand } i \text{ is covered by ambulance } j, \\ 0 & \text{otherwise.} \end{cases}$$

Maximal expected covering model

Decision variables

The objective is to find optimal locations of ambulances such that the expected coverage level is maximized. Considering that several ambulances might cover an area, it is better to let each location to have more than one ambulance.

$$x_j := \text{Number of ambulances in location } j, \qquad \forall j \in A.$$

Similar to maximal covering model, in order to write the objective function, we need to evaluate if a demand zone is covered and if so by how many ambulances it is covered?

$$y_{ik} = \begin{cases} 1 & \text{if demand } i \text{ is covered by at least } k \text{ ambulances,} \\ 0 & \text{otherwise,} \end{cases}$$

$$\forall i \in D, \text{and } k = 1, \ldots, M.$$

Maximal expected covering model

Objective function

The objective is to find optimal locations of ambulances such that the expected coverage level is maximized. First, we need to calculate the expected increase in the coverage level when adding an available ambulance to demand zone i.

- Probability that k ambulances are available

$$\binom{M}{k}(1-q)^k q^{M-k}$$

- Probability that no ambulances are available

$$\binom{M}{M}(1-q)^0 q^{M-0} = q^M$$

Maximal expected covering model

Objective function (continued)

- Probability that at least one ambulance is available

$$1 - \textit{prob.}\{\text{no ambulance is availble}\} = 1 - q^M$$

- Expected coverage in demand location i

Define random variable $H_{i,k}$ to denote the covered demand at node i given that k out of M ambulances can potentially cover node i.

$$H_{i,k} = \begin{cases} h_i & \text{with probability } 1 - q^k, \\ 0 & \text{with probability } q^k. \end{cases}$$

The expected coverage is given by

$$\mathbb{E}(H_{i,k}) = h_i(1 - q^k), \qquad \forall i \in D, k = 1, \ldots, M.$$

Maximal expected covering model

Objective function (continued)

- The increase in expected coverage if an ambulance is added to cover node i which was potentially covered by $k-1$ ambulances

$$\begin{aligned}\Delta\mathbb{E}(H_{i,k}) &= \mathbb{E}(H_{i,k}) - \mathbb{E}(H_{i,k-1}) \\ &= h_i(1-q^k) - h_i(1-q^{k-1}) \\ &= h_i - h_iq^k - h_i + h_iq^{k-1} \\ &= h_i(q^{k-1} - q^k) \\ &= h_iq^{k-1}(1-q).\end{aligned}$$

Maximal expected covering model

Objective function (continued)

The objective is to find optimal locations of ambulances such that the expected coverage level is maximized. The expected coverage in demand i can be thought of the sum of expected increase in coverage when the number of ambulances increases from 0 to k. Total expected coverage is the demand-weighted sum of all demand areas.

$$\max \sum_{i \in D} \sum_{k=1}^{M} \boxed{h_i q^{k-1}(1-q) y_{ik}}$$

Let $y_{43} = 1$. The expected increase in coverage of demand zone 4 when the number of ambulances capable of covering it increases from 2 to 3.

Maximal expected covering model

Constraints

Recall the maximal covering model constraints in facility location models. One constraint limited the number of facilities to be built. Similarly, in maximal expected covering model, we only have M ambulances which can be assigned to demand zones.

$$M \text{ Ambulances:} \qquad \sum_{j \in A} x_j \leq M$$

Maximal expected covering model

Constraints (continued)

The other set of constraints in facility location maximal covering model insured that an area is covered if a facility which can cover it is already built. Here, a demand zone is covered by at least k ambulances if at least k ambulances are already located in places which can cover the demand zone.

$$\text{Demand zone } i: \qquad \sum_{j \in A} a_{ij} x_j \geq \sum_{k=1}^{M} y_{ik}.$$

Note that if $y_{25} = 1$, then $y_{24} = 1, y_{23} = 1, y_{22} = 1, y_{21} = 1$. The right hand side counts the largest number of ambulances that cover demand zone i.

Maximal expected covering model: Formal formulation

Model

Objective function: Maximize covered demand.

$$\max \quad \sum_{i\in D} \sum_{k=1}^{M} h_i q^{k-1}(1-q) y_{ik},$$

Subject to,

$$\begin{array}{lll}
\text{Limited resources:} & \sum_{j\in A} x_j \leq M, & \\
\text{Coverage on zone i:} & \sum_{j\in A} a_{ij} x_j \geq \sum_{k=1}^{M} y_{ik}, & \forall i \in D, \\
 & x_j \leq M, \text{integer} & \forall j \in A, \\
 & y_{ik} \in \{0,1\} & \forall i \in D, 1 \leq k \leq M.
\end{array}$$

Maximal expected covering model: AMPL

Find this model's data file from CANVAS\IE 4910\Files\Ch 02\max-expected-covering.dat, code the model file in AMPL, solve and report the results.

Maximal expected covering model

- Q1: Modify the maximal expected covering model so that coverage is satisfied when out of k ambulance which are located within the right distance, at least 2 are available.
- Q2: Why decision variable x_j has changed from a binary variable to an integer one?

M/M/m queues

The average time in queue for each emergency call is approximated by a M/M/m queueing system.

Queueing system

Naturally, queues are created in any system where service is not instantaneous.

- Queue notation

  Arrival/Service/Servers/Discipline/Capacity/Population

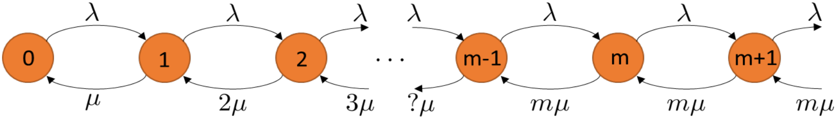
M denotes that interarrival and service times are independent and distributed according to an exponential distribution.

- Exponential distribution is used a lot in queueing theory because of its **memoryless** property.

M/M/m queues

Data and parameters

- Queue system: $M/M/m/FCFS/\infty/\infty$
- Arrival rate: λ, average arrivals per unit time
- Service rate: μ, average finished services per unit time
- Number of servers: m, number of ambulances



M/M/m queues

Target performance measures

- Expected time in queue, W_q
- Expected length of the queue, L_q

- If number of customers $n < m \rightarrow$ No queue!
- If number of customers $n \geq m \rightarrow$ May have a queue.

Therefore, to compute the expected values in target performance measures, we must know **steady-state** probabilities.

- When a queue becomes unstable?
- What are the steady-state probabilities of a $M/M/3$ queue with 3 servers, λ arrival rate, and μ service rate?

M/M/m queues

Formulas

- Little's law

$$L_q = \lambda W_q$$

- Expected length of the queue

$$L_q = \frac{\pi_0}{m!}(\frac{\lambda}{\mu})^m \frac{\rho}{(1-\rho)^2}, \qquad \rho = \frac{\lambda}{m\mu},$$

- Steady-state probability that no one is in the queue

$$\pi_0 = \frac{1}{1 + \sum_{n=1}^{m-1} (\frac{\lambda}{\mu})^n \frac{1}{n!} + \frac{1}{m!}(\frac{\lambda}{\mu})^m \frac{1-\rho^\infty}{1-\rho}}$$

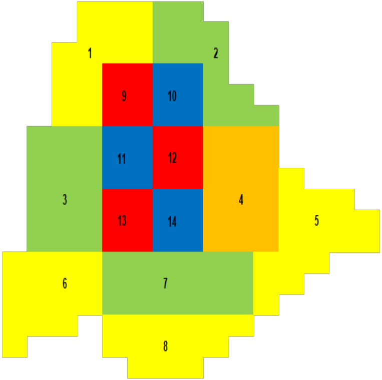
M/M/m queues

Approximating ambulance operations with queueing theory. Is it any good?

- Arrivals: Poisson process → Interarrival times: exponential ✓
- Service times do not usually follow exponential distributions ✗
- Every demand zone has the same call arrival rate **?**
- Every server has the same service time **?**
- Independent arrival ✗ Independent servers ✗ Dynamic service or arrival ✗ Queue discipline ✗

Average response time model

Find the optimal locations of P ambulances such that the demand-wighted total response time is minimized. Response time consists of the traveling time plus the average time in the queue for each call. Each zone has its own arrival rate, and a similar service rate applies to all ambulances. Approximate the average time in the queue by the $M/M/P$ queue assuming it is the same across all zones.



Average response time model

Data and parameters

- Demand zones: $D := \{1, 2, \ldots, 14\}$,
- Ambulance locations: $A := \{1, 2, \ldots, 14\}$,
- Max. # of ambulances: P,
- Service rate: μ,
- Arrival rates: $\lambda_i \qquad \forall i \in D$,
- Demand weights: $h_i \qquad \forall i \in D$,
- Travel times: $t_{ij} \qquad \forall i \in D, \forall j \in A$.

Average response time model

Decision variables

The objective is to find optimal locations of ambulances such that the demand-weighted total response time is minimized. Since we want to minimize the response time, it is better to spread the ambulances so that each zone has at most one ambulance.

$$x_j := \begin{cases} 1 & \text{if zone } j, \text{ has an ambulance,} \\ 0 & \text{otherwise,} \end{cases} \qquad \forall j \in A.$$

To write the objective function, we need to know which demand zone has been assigned to each ambulance.

$$y_{ij} = \begin{cases} 1 & \text{if demand } i \text{ is assigned to ambulance in location } j, \\ 0 & \text{otherwise,} \end{cases}$$

$$\forall i \in D, \text{and } \forall j \in A.$$

Average response time model

Objective function

The objective function is to minimize the demand-weighted total response times. Response times consist of travel times between an ambulance zone j to a demand zone i given by t_{ij} plus the average waiting time in the queue for each call.

- First, let us focus on the average waiting time in the queue, i.e., W_q.
  - Approximate the system by the $M/M/P$ queue.
  - Use the formula to calculate W_q.
  - There is a problem! We have multiple arrival rates.

We have to come up with a way to calculate the total arrival rate.

Average response time model

Objective function (continued)

- Total arrival rate

$$\lambda_t = \lambda_1 + \lambda_2 + \ldots + \lambda_{14},$$

- Average waiting time in the queue

$$W_q = \frac{L_q}{\lambda_t},$$

where $L_q = \frac{\pi_0}{P!}(\frac{\lambda_t}{\mu})^P \frac{\rho}{(1-\rho)^2}$, and $\rho = \frac{\lambda_t}{P\mu}$, and

$$\pi_0 = \frac{1}{1 + \sum_{n=1}^{P-1} (\frac{\lambda_t}{\mu})^n \frac{1}{n!} + \frac{1}{P!}(\frac{\lambda_t}{\mu})^P \frac{1-\rho^\infty}{1-\rho}}.$$

Average response time model

Objective function (continued)

To evaluate the demand-weighted total response time, add t_{ij} to W_q and multiply by the demand in zone i, i.e., h_i. Total response time is the demand-weighted sum of all response times.

$$\min \sum_{i \in D} \sum_{j \in A} \boxed{h_i(t_{ij} + W_q) y_{ij}}$$

Let $y_{43} = 1$. The travel time form zone i to ambulance location j is included in the objective with the demand weight h_i.

Average response time model

Constraints

- Observe the limited number of total ambulances

$$\sum_{j \in A} x_j = P,$$

- Identify the relationship between decision variables x and y

$$y_{ij} \leq x_j, \qquad \forall i \in D, \forall j \in A,$$

- Determine how demands are assigned to ambulances

$$\sum_{j \in A} y_{ij} = 1 \qquad \forall i \in D.$$

Average response time model: Formal formulation

Model

Objective function: Minimize total response time.

$$\min \quad \sum_{i \in D} \sum_{j \in A} h_i (t_{ij} + W_q) y_{ij},$$

Subject to,

$$\begin{array}{lll}
\text{Limited resources:} & \sum_{j \in A} x_j = P, & \\
\text{Relate x and y:} & y_{ij} \leq x_j, & \forall i \in D, \forall j \in A, \\
\text{Assignment:} & \sum_{j \in A} y_{ij} = 1, & \forall i \in D, \\
 & x_j \in \{0,1\} & \forall j \in A, \\
 & y_{ik} \in \{0,1\} & \forall i \in D, \forall j \in A.
\end{array}$$

Average response time model: AMPL

Code the model file in AMPL, solve and report the results. Use Excel to calculate W_q.

- The data to calculate W_q is given in CANVAS\ Files\Ch 02\queue-avg-waiting-time-data.txt
- The solution to W_q is given in CANVAS\Files\Ch 02\queue-avg-waiting-time-sol.
- The data file to AMPL model is given in CANVAS\IE 4910\Files\Ch 02\min-average-time.dat

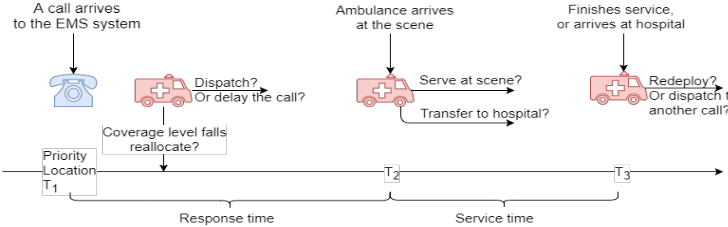
Average response time model

- Q1: Recall the double coverage concept in health care facility location. What was the purpose? Can we modify the average response time model to consider double assignment?
- Q2: What if service rates were not similar? How could we approximate the service rate?
- Q3: Assume different ambulances have the same service rates but we can not approximate the whole system with a single queueing model? Modify the average response time model such that each ambulance has its own queue. Notice that the modified model can not be solved by Gurobi anymore.

Ambulance dynamic models

A dynamic system

Suppose ambulance stations are located in some optimal way. An emergency call arrives into the EMS system. What are the decisions? What is the optimal decision?



Ambulance dynamic models

Dynamic models

Models of this type usually change during time, and thus require time-dependent decision making. Markov chains and Dynamic programming are used to model such problems. A dynamic model consists of five elements:

- Decision epochs
- State space
- Control/Action space
- Transition
- Reward/Cost function

Ambulance dynamic models

Decision epochs

These are discrete/continuous time periods which require the EMS decision maker to take an action

Ex. decision epochs are marked by two events: Arrival of a new call, or an ambulance finishing its service.

State space

This is the most important element in a dynamic programming framework. State variables summarize all the information in the system required to make decisions, predict transitions, and evaluate cost or rewards.

Ex. s :=(event, ambulance states, call states, time), where
ambulance states =(status, origin, destination, time of movement)
call states=(status, arrival time, priority)

Ambulance dynamic models

Action space

Actions are taken at decision epochs and are usually dependent on the state.

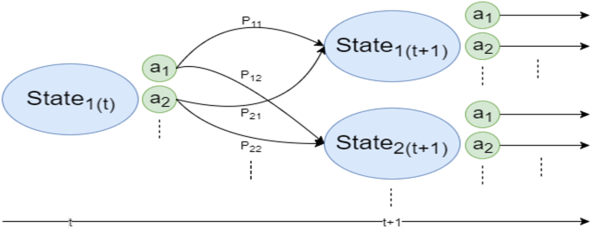
Ex. If the event is of the type “a new call arrives”, at least one ambulance is available, and no higher priority calls are in the queue, the decision maker may **dispatch** an ambulance or **queue** the call.

Ex. If the event is of the type “an ambulance finishes its service”, depending on whether any call is in the queue or not, the decision maker may decide to **return** the ambulance to a base or **redeploy** it to answer a call.

Ambulance dynamic models

Transitions

Transitions are modeled by Markov chains. At any cross section of time, the information is summarized by the state variable. The system dynamically moves to another state with a certain probability. If we take a particular action, the state evolves into another state with some probability.



Ambulance dynamic models

Reward/cost functions

Taking some actions or visiting some system in the evolution of the dynamic system may have some **immediate** cost/reward. They are usually defined to help determine the objective function.

Ex. Suppose we want to minimize the response time in this dynamic model. Upon visiting any state, if the event is of the type "an ambulance reaches the call scene", the immediate cost is the **response time**.

Ambulance dynamic models

Optimality equation

Dynamic systems are finite/infinite horizon, i.e., they evolve up to a certain point in time or they evolve indefinitely. Numerical solutions usually require finite horizon modeling. The total expected (maybe discounted) cost/reward is maximized or minimized in finite models.

Ex. The total expected discounted response time is

$$\min_{\pi}(s^0) = \mathbb{E}\Big\{ \sum_{t=0}^{T} \text{discount factor} * \text{response time of state}(s_t) \Big| s^0 \Big\}$$

π denotes a policy, which is a decision rule by which decisions are determined.

Ambulance dynamic models

Solutions

For most of dynamic problems, no simple analytical solution exists. Exact numerical solutions require **backward induction**, i.e., Enumerating the last period and working backwards. This is a problem for stochastic models since transitions are probabilistic. One way to address this issue is **forward simulation** to get an estimate of final states.

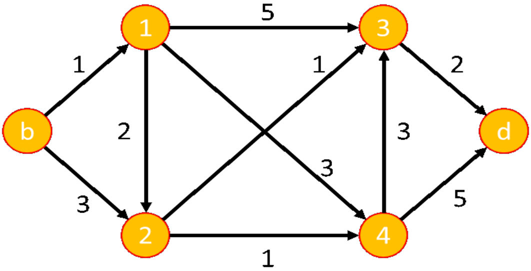
Even after all these methods, real world problems suffer from another disadvantage, **the curse of dimensionality**. If the number of states, or actions are huge, the number of transitions among them grows exponentially, and no machine is able to solve them just by brute force enumeration. Therefore, **Approximate dynamic programming** (ADP) methods are developed.

Class work

- 1: Availability. Code the average response time model in AMPL such that each demand zone is assigned to 2 ambulances. This way, some redundancy is built into the system by lowering the chance that both ambulances are busy at the same time.
- 2: Double coverage. Modify the maximum expected covering model so that coverage is satisfied only when at least 2 ambulances which are located within 20 minutes of demand are available. Code in AMPL.

Class work

- 3: Suppose the base of an ambulance, i.e., *b*, is given in a city. A call arrives at a particular location *d*. What is the shortest path to reach the destination in a network of roads in the city? The network is given below. The number on each edge is the travel time corresponding to that road.



Project

- Maximum availability location problem. In this model, prior to formulation, the number of ambulances b enough to insure coverage level α is determined with busy probability of an ambulance q. The value of b is given by $\lceil \frac{\log(1-\alpha)}{\log q} \rceil$. Formulate a model such that demand areas that α-coverage is guaranteed for them are maximized. Zones, demand, 0-1 coverage matrix and maximum number of ambulances is the same as the maximal expected covering problem. Note that α is 90%.

Sources:

- Daskin, M.S. 1983 A Maximum Expected Covering Location Model: Formulation, Properties and Heuristic Solution, Transportation Science 17(1):48-70
- van den Berg, P.L., van Essen, J.T., Harderwijk, E.J. Comparison of Static Ambulance Location Models, In Logistics Operations Management (GOL), 3rd International Conference on 2016 May 23 (pp. 1-10). IEEE.
- Nasrollahzadeh, A.A., Khademi, A. and Mayorga, M.E., 2018. Real-Time Ambulance Dispatching and Relocation. Manufacturing & Service Operations Management.