Summary

July 2019

1 Epidemic Model

We adopt the ideas from the paper *A Data-Driven Air Transportation Delay Propagation Model Using Epidemic Process Models* (`https://www.hindawi.com/journals/ijae/2016/4836260/`) to build this model.

1.1 Model Setup

Asumme there are $n \in \mathbb{Z}^+$ airports from the dataset annually. Let $N_i \in \mathbb{Z}^+$ where $i = \{0, 1, ..., n\}$ as the total number of the flights in airport i. Let $N_i^*(t) \in \{0, 1, ..., N\}$ be the number of delayed flights in airport i at some time t.

Let $p_i = \frac{N_i^*(t)}{N_i}$ as the proportion of the number delayed flights in terms of the total total number of flights of airport i(delay rate). We then let δ_i be the recovery rate(the rate that an airport back to normal operation)and b_{ij} be the infection rate. If airport i is directly affected by airport j, then $b_{ij} \geq 0$; otherwise, $b_{ij} = 0$. As professor Heather's suggestion, b_{ij} can be considered as an adjacency matrix.

The dynamics of the epidemic process can be written as a differential equation:

$$\dot{p_i}(t) = -\delta_i p_i(t) + \sum_{j=1}^{n} b_{ij} p_j (1 - p_i)$$

If written in matrix form,

$$\dot{\vec{p}_i} = (\mathbf{B} - \mathbf{D})\vec{p} - \mathbf{B}\vec{p} \circ \vec{p}$$

where $\vec{p} = (p_1, p_2, ..., p_n)^\intercal$ is the state vector of the system, $\mathbf{D} = \text{diag}\{\delta_1, .., \delta_n\}$ as the matrix of recovery rate, $\mathbf{B}_{ij} = b_{ij}$ as the matrix of infection rate, and $\circ$ as Hadamard product.

1.2 Solving for Infection Rate

From the dateset, we can denote the number of flights from airport j to airport i as N_{ij}. We normalize the infection rate as $b_{ij} = \frac{N_{ij}}{\sum_{j=1}^{n} N_{ij}}$. Thus, $\sum_{j=1}^{n} b_{ij} = 1$.

We then use the equality to solve for the infection rate:

$$\frac{b_{ij}}{b_{ik}} = \frac{N_{ij}}{N_{ik}}$$

1.3 Solving for Recovery Rate

Denote m as the number of individual time(randomly picked). then we have m individual time t_1,.., t_m with time interval Δt(randomly picked). By definition of differentiation, we have equation

$$\frac{\vec{p}(t_k + \Delta t) - \vec{p}(t_k)}{\Delta t} = (\mathbf{B} - \mathbf{D_k})\vec{p}(t_k) + \mathbf{B}\vec{p}(t_k) \circ \vec{p}(t_k)$$

Simplify the equation, we can solve for $\mathbf{D_k}$, then we normalize the recovery rate by $\mathbf{D} = \frac{1}{m}\sum_{k=1}^{m} \mathbf{D_k}$.

1.4 Processing the Result

Since we are able to solve the system, we can make a delay rate vs. time plot for a specific airport i. We will compare this result with our other models to see if it is a good prediction. Nevertheless, we still need a new mathematical definition of vulnerability index for comparison.

2 Stochastic Epidemic Model

We further develop our dynamical system by adding an extra stochastic term,

$$\dot{p_i}(t) = -\delta_i p_i(t) + \sum_{j=1}^{n} b_{ij} p_j (1 - p_i) + \frac{\epsilon_i(t)}{N_i}$$

where $\epsilon_i(t)$ is some Poisson distribution.

3 Extra perturbation on Infection rate

$$\dot{p_i}(t) = -\delta_i p_i(t) + \sum_{j=1}^{n} (b_{ij} + \epsilon) p_j (1 - p_i)$$

where ϵ is an intrinsic perturbation term.

4 Vulnerability Index

Let $\alpha, \beta \in (0, 1)$ where α is threshold rate and β is delay rate. Denote the set of Airports by A. We define our vulnerability index

$$\begin{aligned} V_{\alpha,\beta} \coloneqq A &\longrightarrow \mathbb{R}. \\ i &\mapsto V_{\alpha,\beta}(i) \end{aligned}$$

Also, the delay rate over time $[T_0, T_1]$ as the output of our model is defined as

$$f \coloneqq A \longrightarrow L(T_0, T_1),$$

where $L(T_0, T_1)$ is the measurable functions over $[T_0, T_1]$. That is, for any airport i, we denote $f_i(t)$ as the output of our model, where $t \in [T_0, T_1]$.

And all of our models takes input of the data set of FAA flights in 2008 and an initial condition, which mathematically is an assignment of $I_i = f_i(0) \in (0, 1)$ for each airport i.

Mathematically, our vulnerability index for airport i can be written as

$$V_{\alpha,\beta}(i) \coloneqq \int_{T_0}^{T_1} \mathbb{1}_{\{f_i > \alpha\}}(t)dt$$

where $\mathbb{1}_{\{f_i>\alpha\}}(t)$ is the characteristic function of the set and $f_i(t)$ is the output of models with input data D initial values $I_{\beta,i}$.

Furthermore, we define two types of classical distributions for $I_{\beta,\imath}$ as airport based, and I_β^f as flight based. $I_\beta^\imath$ means $100\beta\%$ of airports have all flights departed being delay. I_β^f means $100\beta\%$ of flights are delayed.

The purpose for this definition is to capture the the concept of volubility in the sense of epidemiology. One use a a thresh hold α to determine if an airport i is in break down status at time t which is $f_i(t) > \alpha$. Consider the following two representative cases One is an sharp jump in the graph $f_i(t)$. One is an general increasing function in the graph $f_i(t)$. One is more tempt that to conclude the former region as a robust person, while the latter one as a potential patient.

If the model is a stochastic model, one may has a similar definition for $V_{\alpha,\beta}$. Instead of taking an initial value $I_{i,\beta}$, we interpret it as an initial distribution $I_{i,\beta}$. And $f_i(t)$ is a stochastic process for each i. And

$$V_{\alpha,\beta}(i) \coloneqq E\int_{T_0}^{T_1} \mathbb{1}_{\{f_i > \alpha\}}(t)dt$$

5 Formulation of SDE-SISN Model

This section will introduce the SDE-SISN model(stochastic differential equation SIS network-based model). The first subsection is devoted to the formulation of its discrete version DTMC-SISN(Discrete time Markov chain network-based model). The second subsection is to derive SDE version. The third subsection is to give a plan for numerical methods(TODO).

5.1 Formulation of DTMC-SISN Model

First, recall the deterministic epidemic model without considering network.

Let $N(t)$ be the total population, $I(t)$ be the infected total population, $N(t), I(t), S(t) \in C^1(= [0, +\infty)$ the classical epidemic model suggests that the following ODE with $N(0), I(0), S(0)$ given, $N(t) \equiv N$

$$N = S(t) + I(t)$$

$$\frac{dS}{dt} = -\frac{\beta}{N}SI + \delta I$$
$$\frac{dI}{dt} = \frac{\beta}{N}SI - \delta I$$

For its analog for the network case, denote A as the set of nodes in the network, G as an undirected network, $N_a(t), S_a(t), I_a(t) \in C^1[0, +\infty)$ as the number of total, susceptible, infectorous population respectively. We deduce an analogical deterministic SISN model:

$$N_a = S_a(t) + I_a(t)$$
$$\frac{dS_a}{dt} = -\sum_b \beta_a{}^b \frac{S_b}{N_b} + \delta_a I_a$$
$$\frac{dI_a}{dt} = \sum_b \beta_a{}^b \frac{S_b}{N_b} I_b - \delta_b I_b$$

where $N_a(t) \equiv N_a$, with $N(0), S(0), I(0)$ given and $\beta_a{}^b \neq 0$ if $b \notin Adj(a)$. We have the following matrix form.

$$\frac{dI}{dt} = \beta(E - I \circ N^{-1}) \circ I - \mathbf{D}I \tag{1}$$

To utilize the network information, with this understanding, one way is to consider a time homogeneous discrete time Markov chain model.

Back to single region case, now we assume time is discrete, and $S(t), I(t)$ are random variables with range in $\mathbb{N}$ for $t \in \mathbb{N}\Delta t \geqslant 0$, and $S(t) + I(t) = N$. We add two more assumptions:

(i) memorylessness: $E(I(t + \Delta t)|I(t_1), ..., I(t_k) \sim E(I(t + \Delta t)|I(t))$ for any $t_1, ..., t_k \leqslant t$.

(ii)local branching property: $\Pr(I(t + \Delta t) - I(t)) = 0$ for any $t \in \Delta t\mathbb{N}_{\geqslant 0}$.

The first assumption is for Markov chain model building, and the second assumption is to make the model easier to be adapted to continuous time version.

By the memorylessness property. Let $p_{ij} = P(I(t + \Delta t) = i|I(t) = j)$, then we have the following ODEs:

$$p_{ji}(\Delta t) = \begin{cases} \frac{\beta}{N} i(N - i), & \text{if } j = i + 1 \\ i\delta\Delta t, & \text{if } j = i - 1 \\ 1 - (\frac{\beta}{N} i(N - i) + \delta i)\Delta t, & \text{if } j = i. \\ 0 & \text{otherwise} \end{cases}$$

Moreover, Δt needs to be chosen small enough to ensure $p_{ji}(\Delta t) \in [0, 1]$.

Now let us generalize to the network case. The analog to p_{ij} is the following: Let $L = Hom(A, \{0, ..., N)) \simeq \{0, ..., N\}^{|A|}$, then

$$p_{fg}(\Delta t) : L \times L \times (0, +\infty) \longrightarrow \mathbb{R}$$
$$(f, g, \Delta t) \mapsto p_{fg}(\Delta t)$$

Now by considering the network case, one needs to find the distribution of $\{I_a(t)\}_{a\in A}$ with $t \in \Delta t\mathbb{N}_{\geqslant 0}$, $I_a(t) \in \mathbb{N}_{\geqslant 0}$ such that:

(i)$I_a(t)$ are all memory-less

(ii)local branching property $\Pr(I(t+\Delta t) - I(t)) = 0$ for any $a, b \in A, t \in \mathbb{N}\Delta t \geqslant 0$.

$$P_{f(a),g}(\Delta t) = \begin{cases} \sum_b \beta_a{}^b \frac{N_b - g(b)}{N_b} t g(b)\Delta t & f(a) = g(a) + 1 \\ \delta_a t g(a)\Delta t & f(a) = g(a) - 1 \\ 1 - \sum_b \beta_a^b \frac{N_b - g(b)}{N_b} g(b))\Delta t - \delta_a g(a)\Delta t & f(a) = g(a) \\ 0 & \text{otherwise} \end{cases}$$

To simplify our notation,we write:

$$P_{f(a),g}(\Delta t) = \begin{cases} B(g)\Delta t, & f(a) = g(a) + 1 \\ D(g)\Delta t, & f(a) = g(a) - 1 \\ 1 - (B(g) - D(g)) \cdot \Delta t & f(a) = g(a) \\ 0 & \text{otherwise} \end{cases}$$

We choose Δt such that $\sup_{g\in L}\{B(g) + D(g)\} \cdot \Delta t \leqslant 1$. Given $p_{f,g}(0)$, one may compute $p_{f,g}(t)$ for any $t \in \Delta t\mathbb{N}_{\geqslant 0}$.

This finishes the formulation of DTMC-SISN model.

However, considering in our case $|A|300$, even though we have an upper bound N for all $I_a(t)$, we still need to compute $(N^{|A|})^2$ for one iteration, which is exponential. For this reason, we formulate a continuous time model SDE-SISN in the next subsection.

5.2 SDE-SISN Model

We start again with one region case. Suppose $\{I(t)\}_{\mathbb{R}\geqslant 0}$ is a stochastic process with $I(t) \in \mathbb{R}_{\geq 0}$ for any $t \geqslant 0$ with the following properties:

(i) $\mathbb{E}(I(t+\Delta t)|I(t)) \sim \mathbb{E}(I(t+\Delta t)|I(t_1), ..., I(t_k))$ for any $t_1, ..., t_k \leqslant t, \Delta t \leq 0$

(ii) $I(t)$ is a continuous function of t.

Let $\Delta I = I(t + \Delta t) - I(t)$,

$$\begin{aligned} P(\Delta I = -1) &= b(i)\Delta t + o(\Delta t) \\ P(\Delta I = 0) &= 1 - (b(i) + d(i))\Delta t + o(\Delta t) \\ P(\Delta I = 1) &= d(i)\Delta t + o(\Delta t) \\ P(|\Delta I| \geqslant 2) &= o(\Delta t) \end{aligned} \tag{2}$$

Note that

$$\begin{aligned} \mathbb{E}(\Delta I) &= (b(I) - d(I))\Delta t + o \cdot \Delta t \\ Var(\Delta t) &= (b(I) + d(I))\Delta t + o \cdot \Delta t, \end{aligned}$$

where $b(i) = \frac{\beta}{N} i(N-i)$ and $d(i) = \delta_i$. Thus$I(t)$ is a solution of its SDE

$$dI = (b(I) - d(I))dt + \sqrt{b(I) + d(I)} + dW_t,$$

where W_t is the Wiener process.

For the network case, we have the following model:

Let $\Delta I_a = I_a(t + \Delta t) - I_a(t)$, $I(t) = (I_a(t))_{a \in A}$, then

$$\begin{aligned} \Pr(\Delta I_a = +1) &= B_a(I)\Delta t \\ \Pr(\Delta I_a = -1) &= D_a(I)\Delta t \\ \Pr(\Delta I_a = 0) &= (1 - B_a(I) - D_a(I))\Delta t \\ \Pr(|\Delta I_a| \geqslant 2) &= o(\Delta t) \end{aligned} \tag{3}$$

We have $\mathbb{E}(\Delta I) = (B_a(I) - D_a(I))\Delta t$ and $\mathrm{Cov}(\Delta I, \Delta I) = C\Delta t$.

Here we add one more assumption, which is questionable:

$\Delta I = (\Delta I_a)$ are independent.

Under this assumption, we have $C = diag\{B_a(I) + D_a(I)\}$. Thus, $I(t)$ is a solution of its SDE $dI = (B_a(I) - D_a(I))dt + C^{\frac{1}{2}} dW(t), where \mathrm{C} = \mathrm{C}_{ab}$ and $W(t)$ is the $|A|$ dimensional Wiener process.

Now we are ready to formulate our model:

$$dI = ((\beta(E - I \circ N^{-1}) \circ I - \mathbf{D}I)dt + ((\beta(E - I \circ N^{-1}) \circ I + \mathbf{D}I)^{\frac{1}{2}} \circ W(t),$$

where $\circ$ is the Hadamard product of the vector, with $N^{-1} \coloneqq (N_a^{-1})_{a \in A}$, matrix $\beta = {\beta_a}^b$, $\mathbf{D} = \mathrm{diag}\{\delta_1, .., \delta_{|A|}\}$ is a matrix of recovery rate, and $E = (1, 1, ..., 1)^\intercal$ is a vector with all entries being 1. $I(0)$ and $N = (N_a)$ given.

6 Infection Ratio and Vulnerability Index

This section is devoted to the definition of infection ratio and vulnerability index based on the given data set. We interpret our data as a set of three tuples $S = \{(a, b, t) | a, b \in A, t \in [T_0, T_1]\}$, where A is the set of airports, to represent a flight from airport b to airport a at time t. In our data set, for any airport a, and a time period $[T_0, T_1]$, we associate airport a with a time series $M(t)$. Mathematically,

$$M_a(t) = M_a(T_0) + \sum_{b \in A, s \in [T_0, T_1]} \mathbb{1}_S(a, b, s) - \sum_{c \in A, s \in [T_0, T_1]} \mathbb{1}_S(c, a, s).$$

We denote that a flight (a, b, s) is arrival delayed if $ArrDelay > 15$; (a, b, s) is delayed if $ArrDelay > 15$ or $DepDelay > 15$.

Let

$$\begin{aligned} D^{arr} &\coloneqq \{(a, b, t) \in S | \text{ArrDelay of} (a, b, t) > 15\}, \\ D^{dep} &\coloneqq \{(a, b, t) \in S | \text{DepDelay of} (a, b, t) > 15\}, \\ D^{total} &\coloneqq D^{arr} \cup D^{dep}. \end{aligned}$$

Then

$$I_a(t) \coloneqq I_a(T_0) + \sum \mathbb{1}_D(a,b,s) - \sum \mathbb{1}_D(c,a,s).$$

The delay rate $f_a(t) : A \times [T_0, T_1 - \Delta t] \longrightarrow \mathbb{R}$ is defined as

$$f_a(t) \coloneqq \frac{\int_t^{t+\Delta t} |I_a'(s)| ds}{\int_t^{t+\Delta t} |M_a'(s)| ds}.$$

Our model will produce an estimation $\hat{f}_a(t)$ for $f_a(t)$. For deterministic SISN model, we have

$$\hat{f}_a(t) \coloneqq \frac{I_a(t)}{TM_a}$$

as a function. For SDE-SISN model, we have

$$\hat{f}_a(t) \coloneqq \frac{I_a(t)}{TM_a}$$

as a random variable, where $TM_a \coloneqq \int_{T_0}^{T_1} |M_a'(s)| ds$. Now we are ready to define our Vulnerability index VI:

$$\begin{aligned} VI : A \times [0,1] \times \mathbf{D}^{|A|}([0,1]) &\longrightarrow \mathbb{R} \\ (a, \alpha, (p_a)_{a \in A}) &\mapsto VI_{\alpha,\beta}(a) \end{aligned} \tag{4}$$

where $\mathbf{D}(0,1)$ is the space of probability density function vanishing outside [0,1].

Let $\frac{I_a(0)}{TM_a}$ be a random variable with density function p_a, then for either model $\hat{f}_a(t)$ is a random variable. Thus, the vulnerability index is defined as

$$VI_{\alpha,p}(a) \coloneqq \mathbb{E} \int_{T_0}^{T_1} \mathbb{1}_{\{t | \hat{f}_a(t) \geqslant \alpha\}}(t) dt.$$

We introduce two classical initial distributions with parameter β: airport based distribution $p^{airport,\beta}$ and flight based distribution $p^{flight,\beta}$.

For $p^{airport,\beta}$, we let A^β be a random subset of A with $|A^\beta| = |A|\beta$.

$$\hat{f}_a(T_0) = \begin{cases} 1 & \text{if } a \in A^\beta \hat{f}_a(T_0) = 0 \\ 0 & \text{if } a \notin A^\beta \hat{f}_a(T_0) = 0. \end{cases} \tag{5}$$

For $p^{flight,\beta}$, we let S^β be a random subset of S with $|S|^\beta = |S|\beta$,

$$E^\beta \coloneqq \{(a,b,t) \in S^\beta | \text{ArrDelay of}(a,b,t) > 15 \text{ or DepDelay of}(a,b,t) > 15\}.$$

$$\hat{f}_a(T_0) \coloneqq \frac{\int_{T_0}^{T_0+\Delta t} |IN_a'(t)| dt}{\int_{T_0}^{T_0+\Delta t} |M_a'(t)| dt}.$$

where

$$IN_a(t) \coloneqq IN_a(T_0) + \sum_{b \in A, s \in [T_0, T_1]} \mathbb{1}_{E^\beta}(a,b,s) - \sum_{c \in A, s \in [T_0, T_1]} \mathbb{1}_{E^\beta}(c,a,s).$$

7 Parameter estimation for Deterministic SISN model

In this section, we use a statistic method to estimate $B = (b_{ab})$ and $D = diag\{\delta_1, \dots \delta_{|A|}\}$ in (1). Denote the (B, D) by θ, and the solution to (1) by $I(t, \theta)$. and we have data set $\{(t_i, I_i)\}_{i\in\mathbb{N}}$.

For a fixed $\xi > 0$ and $\sigma_0 > 0$, let

$$Y_i = I(t_i; \theta_0) + I(t_i; \theta_0)^{\xi}\epsilon_i,$$

where $\epsilon_i \sim N(0, \sigma_0^2)$ with each rare parameter θ_0. So Y_i is a random variable, we want to construct a weighted least square estimator $\hat{\theta}_{LS}$ for θ_0, such that it is strong consistent. That is $\hat{\theta}_{LS}$ is a random variable with $\hat{\theta}_{LS} \overset{\text{a.s.}}{=} \theta_0$.

The choice of ξ has a classical meaning. For $\xi = 0$, it gives the estimator for ordinary least square. For $\xi = 1$, it is a linear noise.

Let $L^{(n)}(\theta) = \sum_{i-1}^{n} w_i(Y_i - I(t_i, \theta))^2$ with weight $w_i = \frac{1}{I(t_i,\theta_0)^{2\xi}}$.

Let $\hat{\theta}_{LS}^{(n)} = \operatorname{argmin}_\theta L^{(n)}(\theta)$ and $\hat{\theta}_{LS} = \lim \hat{\theta}_{LS}^{(n)}$.

Consider an function $K^{(n)} = \sum_{i=1}^{n} Y_i I(t_i; \theta)^{-\xi}$, one can see that $\hat{\theta}_{LS}$ is the least square estimator for $Y_i I(t_i; \theta)^{-\xi}$. The least square estimator for the non-linear model $I(t_i; \theta)^{1-\xi}$ can be traced by asymptotic theory. We summarize some essential theorems in [1] as below:

(1) ϵ_i are $\overset{\text{i.t.d}}{\sim} N(0, \sigma^2)$.

(2) $f(t, \theta) \in C(\mathcal{L} \times \Theta)$ is continuous.

(3) $\mathcal{L}, \Theta$ compact.

(4)The empirical distribution function $H_n(t) := \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{s|t_i\leqslant s\}}(t)$ converges to a distribution function.

(5)$f(x; \theta) \overset{\text{a.s.}}{=} f(x, \theta_0) \Rightarrow \theta \overset{\text{a.s.}}{=} \theta_0$ where θ is a random variable, then $\hat{\theta}_{LS} \overset{\text{a.s.}}{=} \theta_0$.

Moreover, if the rest of the assumptions hold, we have:

(6)$\theta_0 \in \Theta$

(7)f is twice continuously differentiable with respect to θ

(8)$\Omega(\theta_0) = \int_T \nabla_\theta f(\nabla_\theta f)^t dH(t)$ is non-singular, then $\hat{\theta}_{LS} \sim N_p(\theta_s, \epsilon_0^2\Omega^{-1}(\theta_0))$ where $p = \dim \Theta$.

Now let us justify these eight assumptions for our case. Assumption (1), (2),(3),(6) satisfy automatically.

To satisfy (4), we pick data randomly in the range $[T_0, T_1]$; thus $dH(t) = \frac{1}{T_1 - T_0}dt$ where $T = [T_1, T_0]$.

To check (4), suppose we have two parameters $\theta = (B, D)$ and $\theta' = (B', D')$, such that ODE

$$\begin{aligned} \frac{dI}{dt} &= (B - D) - BN^{-1} \times I \times I, I(T_0) = I_0 \\ \frac{dI}{dt} &= (B' - D')I - B'N^{-1} \times I \times I, I(T_0) = I_0 \end{aligned} \tag{6}$$

has the same solution. If we assume I_0 has no entry equal to 0, then for $t \in [T_0, T_1]$, (

In general if we assume $I > 0$ and dim