Infectious Disease Spread through Indian Transportation Network

A Thesis

submitted to
Indian Institute of Science Education and Research Pune
in partial fulfillment of the requirements for the
BS-MS Dual Degree Programme

by

Onkar Sadekar

Indian Institute of Science Education and Research Pune
Dr. Homi Bhabha Road,
Pashan, Pune 411008, INDIA.

June, 2021

Supervisor: Prof. M.S. Santhanam

Certificate

This is to certify that this dissertation entitled 'Infectious Disease Spread through Indian Transportation Network ' towards the partial fulfillment of the BS-MS dual degree programme at the Indian Institute of Science Education and Research, Pune represents study/work carried out by Onkar Sadekar at Indian Institute of Science Education and Research under my supervision, during the academic year 2020-2021.

Prof. M.S. Santhanam

2.6.2021

Committee:

Prof. M.S. Santhanam

Dr. G.J. Sreejith

“ To those – who believe in the beauty of their dreams,
for the future belongs to them”

Who said that every wish
Would be heard and answered
When wished on the morning star?
Somebody thought of that
And someone believed it
Look what it’s done so far
What’s so amazing
That keeps us star gazing
And what do we think we might see
Someday we’ll find it
The Rainbow Connection
The lovers, the dreamers and me!

Declaration

I hereby declare that the matter embodied in the report entitled 'Infectious Disease Spread through Indian Transportation Network ', are the results of the work carried out by me at the Department of Physics, Indian Institute of Science Education and Research, Pune, under the supervision of Prof. M.S. Santhanam and the same has not been submitted elsewhere for any other degree.

01/06/2021

Onkar Sadekar

Acknowledgments

Throughout this thesis, I have received enormous help and support from my mentors, peers, family, and friends.

I would like to thank my supervisor, Professor M.S. Santhanam, without whose support this thesis would not be in its current form. His belief in me through the tough times of pandemic made the journey enjoyable. I would like to sincerely thank him for allowing me to work on this topic and helping me learn the essential things to be a researcher.

I would also like to thank Dr. G.J. Sreejith and Dr Sachin Jain for being an integral part of this project. The discussions during the meetings were lively and made me understand the art of asking the right questions and optimal ways to approach them. I would also like to thank Dr. Bijay Kumar Agarwalla and Dr. Urna Basu for enabling me with the necessary skills and always encouraging me. I would like to acknowledge the INSPIRE-grant from DST for the past 5 years.

This project would definitely be incomplete if not for my friend and colleague Mansi Budamagunta. She collected most of the raw data for this data-intensive project. She was always available to discuss academic as well as non-academic things and criticize me whenever necessary.

I would like to thank my dearest friends Mohan, Akhila, Adarsh, and Suman for always entertaining my stupid acts and patiently listening to my random blabberings throughout all these years at IISER. I would also like to thank Mitali, Rounak, Shambhavi, Shriya, Sagnik, Aagam, Harsha, and Sahiti for being an integral part of my life at IISER. I would also like to thank my juniors Shraddha, Aniket, Gautam, and Ramya for the incredible memories throughout these years.

I would like to thank Aai and Baba for making all the sacrifices over the years for me and always believing in me to do my best academically. Their unwavering support and care will always continue to be a prominent part of my life.

Finally, I would like to thank two extraordinary people in my life – Aanjaneya and Dada, for being present for every significant decision in my life in the past few years and always directing me to the correct path by all possible means! No words are enough to express my gratitude towards them, and I would forever be in their debt for making me what I am today.

Abstract

We study infectious disease spread through the Indian transportation network in this thesis. We use a hazard index to quantify the risk faced by 446 Indian cities for an epidemic starting from any city. This hazard index, also called as *effective distance* was first introduced by Helbing and Brockmann to explain the global spread of infectious diseases. Even though there have been a lot of India-specific studies to examine and predict the spread of infection, to our knowledge, none of them consider long-distance travel through multiple modes of transportation as the primary source of infection. We estimate the traffic for three modes of transport – air, rail, and road to construct the transportation network for India. We use the Susceptible-Infected-Recovered (SIR) metapopulation model to simulate the dynamical system and quantify the associated risk by the arrival time of the infection to the city. We show that the *effective distance* is an objectively better hazard index than geographical distance and that it works the best for higher values of SIR infection rate parameters and lower threshold of infected cases to define arrival time. We also illustrate that *effective distance* can be modified to cover the case of multiple outbreak locations. Before comparing with the real-life data of Covid-19 cases, we give evidence for removing critical links using the link salience treatment to curb the spread of the disease. Finally, we show that the SIR metapopulation model has some static and dynamical properties similar to the Fisher-KPP class of equations through numerical simulations. Our study opens up multiple new avenues to build a full-scale working model for India with better mobility and traffic data and study diffusion-like processes on heterogeneous networks.

Contents

Chapter 1

Introduction

As of July 5, 2021, more than 180 million people have been infected by the Covid-19 virus globally, while around 4 million have been deceased [1, 2]. In other words, 2% of the world population is infected by the virus strain starting from a few handfuls of cases in Wuhan in December 2019. The world is coming closer, but that has brought its own set of disadvantages. In the last few decades, long-distance travel has been a boon for the business, and collective growth of countries [3]. However, as shown by many incidents during the same time, it has also eased the spread of deadly infectious diseases to remote corners of the world in a short amount of time [4, 5, 6, 7, 8]. The most challenging problem for policymakers and governments worldwide is predicting the pandemic as soon as any new cases are found.

The ubiquitous Susceptible-Infected-Recovered (SIR) compartmental model helps predict the growth of infection in a *well-mixed* population – a system where each part has an equal probability of interacting with any other part [9, 10]. However, the assumption of a well-mixed population becomes invalid once we consider the real-life situation where the population is distributed across cities. The metapopulation concept was introduced in epidemiology to account for the movement dynamics based on the mobility data [11, 12]. Metapopulation considers the population divided into two or more levels of scale, where the interaction within a scale is different from the rest. Even though this approach allows one to include another level of complexity to the model, the pattern of spreading of any infectious disease is observed to be non-trivial when we consider the metapopulation to be cities connected in the form of a network [13, 14]. One of the long-standing goals in epidemiology has been to predict the pattern in which the infections spread. Since the resources are limited, forecasting how the patterns emerge is a practical concern that can save many valuable lives [15, 16, 17, 18].

Brockmann and Helbing first introduced the concept of '*effective distance* (D_{eff})' in [19]. They found that D_{eff} – a measure defined based on the probability of an agent traveling from one city to another works exceptionally well for predicting the hazard at a global level. There is a multitude of studies at the global as well as national level predicting the spread of Covid-19 to countries and states [20, 21, 22, 23]. Even though there have been many India-specific studies in the past few months for predicting the epidemic risk, to our knowledge, no study considered the Indian transportation network to the scale we did [24, 25, 26]. The disease majorly spreads between different countries through air traffic. However, for a country like India, there are multiple modes of transport for short and long-distance transport, which carry significant passenger loads each day. The difference in terms of actual speed and distribution of traffic across various modes makes the situation in India drastically different from a global level spread. Globally, India is the second-most populous country with more than 1.3 billion residents [27]. The diversity in socio-economic conditions in India is rarely seen in other parts of the world. Any infectious disease is perilous in this environment because many people live in suboptimal conditions with poor sanitation and ventilation. Predicting the spread of disease in such conditions combined with the heterogeneity in transport makes it an arduous, yet critical task [28, 29, 30].

Around 10 million people travel every day in India by various means. This number is higher than the population of many countries in the world. Unfortunately, there is no systematic record of the number of people traveling between two cities in the public domain. There are a handful of studies about railway and air transport [31, 32, 33] but given the uncertainty of schedules, and the constant addition of newer routes, it is hard to imagine the sustainability of such data over a long period. Thus, we are left with no other option but to collect, predict, or build the data ourselves using all the tools available. In order to be able to predict the risk of infection, we must have the passenger traffic data.

There have been few studies centered around India, but to our knowledge, none of them accounted for the long-distance transport for a large set of cities [34, 35, 36, 37]. The problem with including just the big cities is that the small cities can contribute significantly to the spread even though they might not contribute to the number of infected cases. Preventive measures like lockdown are economically harsh and can have undesirable effects if not appropriately executed [38, 39, 40, 41]. Proper planning based on the prediction of disease spread is still missing for policymakers. We will look at the case of predicting the spread of infectious diseases in India through a transportation network.

The plan for this thesis is as follows: In Chapter (2), we start by discussing the basics of the epidemiological compartmental models and the metapopulation model. After motivating the

equation of transport on a network, we combine the SIR and the transport dynamics model under some assumptions to write down the equations for the SIR metapopulation model. After proposing the model, we look into the details of the data collection process for the three modes of transport – air, rail, and road. The data for air transport was available freely, while only the train schedules were available for the rail transport. Virtually, no information was available for the road transport. We propose two algorithms for estimating rail and road traffic data, making the optimum use of the available data – train schedule, geographical coordinates, and population of the cities. Thus, by the end of Chapter (2), we will have all the necessary tools fully assembled to analyze the SIR metapopulation compartmental model.

We will begin Chapter (3) by showing the phase transition in the metapopulation model, a signature of the SIR well-mixed population model. We will then define the *Time of Arrival* in order to quantify the hazard associated with each city. We consider risk based on two different approaches, a fraction of the population is infected, or an absolute number of people are infected. Before finalizing the definition of D_{eff} , we will look into the effectiveness of D_{eff} using various definitions. We will then show that D_{eff} , defined using the transition jump probability $\boldsymbol{P}$-matrix, shows a much better correlation with the *Time of Arrival* defined using the absolute threshold than any other pairs of definitions. We will illustrate the robustness of this pair by averaging the best fit over outbreak locations and then looking at its trend for various parameters. We will show that the model works best when the SIR infection parameters are high and the absolute threshold is low.

After looking at more theoretically motivated aspects of the model, we will consider a few practical extensions, such as multiple outbreak locations and the effect of removing links from the network. We will show that a modified definition of D_{eff} works for two outbreak locations and can be extended to multiple outbreak locations. We will also verify the method of link salience first introduced by Brockmann *et al.* to find out the critical links in the network [42]. We find that there are no critical links in the network common to all nodes and hence, the dynamics of infection spread highly depends on the outbreak location. Finally, we will compare our results with the real-life data for Covid-19 and give possible reasons for the mismatch between the two. Amongst multiple things, the main reason for the mismatch remains to be the quality of the transport and the real-life data of Covid-19 cases.

In Chapter (4), we dive into the theoretical side of D_{eff} . We will show that the SIR metapopulation model can be transformed into an SI metapopulation model under certain limits. The SI model is very similar in form to the Fisher-KPP family of equations, which admit wave-like solutions [43]. In order to prove the correspondence between these two systems (which would give ample evidence for the effectiveness of D_{eff} for any network), we will compare numerical simu-

lations to see any similarity between the two systems. In order to make the correspondence even stronger, we will compare just the diffusive terms of the two models and show that the mean and the standard deviation of the infected fraction indeed show very similar trends. Thus, we show a weak and indirect connection between diffusion on a line and diffusion on a network, which would explain the linear relationship between *Time of Arrival* and D_{eff} . In Chapter (5), we will summarize all the key findings from our thesis and give future work directions.

The main focus of this thesis is to study the problem of infection spreading in India and construct a hazard index that would quantify the risk of a city for any pandemic in the future using the known tools from physics, epidemiology, network, and data science. Given the complexity, the scale of the problem, and a striking lack of raw data, the progress we have made certainly brings us closer to a full-scale working model, and we hope that the data we have collected will be helpful for future epidemiological as well as non-epidemiological models.

Chapter 2

SIR Metapopulation Model and Indian Traffic Data

2.1 Introduction

Susceptible-Infected-Recovered (SIR) model is a well-studied epidemiological model that successfully explains the gross features of spreading infectious diseases in many real-life cases [9]. Multiple extensions of this model have been proposed and successfully applied to specific cases of disease transmission. We will look at one such extension in Section (2.2) by separating the spatial scales of intracity and inter–city dynamics. We will understand and rationalize the details and the underlying assumptions of this extended SIR model (henceforth called as SIR metapopulation model). In Section (2.3), we will look at the methods and techniques used to collect data for our model. Finally, we summarize the notable points about the SIR metapopulation model and the traffic data in Section (2.4).

2.2 SIR Metapopulation Model

There has been much interest in understanding the spreading processes often observed in nature. In particular, epidemiology – a discipline dealing with questions like 'how, why, and what next' for the spread of diseases has developed mature tools to understand and analyze infections [44].

In the following few Subsections, we will look at the details of the SIR model before delving into the concept of the metapopulation. Towards the end, we will explain how these two completely

different dynamics can be integrated into a single equation (given by Eq. (2.4)), which will serve as the backbone for most of the analysis in this thesis.

2.2.1 Compartmental Models

One of the most successful approaches in epidemiology is the family of compartmental models [10]. The underlying assumption for these models is that the population is divided into various compartments based on the individual's state, and all individuals move from one compartment to another based on various environmental and health-related factors. However, given a large number of individuals and many uncontrollable and often unknown parameters in the system, it is hard to precisely predict which individual would be the next one to jump from one compartment to the other. To deal with this complexity, we replace the innumerable degrees of freedom in the system with randomness. We assume an effective jump rate for all individuals to move from one compartment to another and observe how well it agrees with reality. We may not know which individual will be next to make the transition. However, we can predict the size of the fraction of the population that will transition in a unit period of time. Thus, to compensate for the incomplete information, we settle for a probabilistic view of the system. These models cannot predict which individual will get infected next but can tell the evolution of the total number of people in all compartments with time.

2.2.2 SIR Model

Let us now look at the details of one of the most successful models of the family, namely the SIR-Model, where $S(t)$, $I(t)$, and $R(t)$ denote the susceptible, infected, and recovered (or removed) population respectively at time t. We assume that the population is well-mixed. A well-mixed population denotes the case where every individual interacts with every other individual with equal probability. Thus, it is easy to see that the probability of a susceptible individual getting infected is directly proportional to the number of infected cases in the population. We can also see that the number of individuals going from susceptible to infected and then from infected to recovered has to be proportional to susceptible and infected people, respectively. Putting all this together, we get the rate equation for the SIR model as follows:

$$\frac{\partial S(t)}{\partial t} = -\alpha \frac{S(t)I(t)}{N},$$

$$\frac{\partial I(t)}{\partial t} = +\alpha \frac{S(t)I(t)}{N} - \beta I(t),$$
$$\frac{\partial R(t)}{\partial t} = +\beta I(t). \tag{2.1}$$

In Eq. (2.1), α denotes the rate of infection per capita, while β is the recovery rate. We note that $S(t) + I(t) + R(t) = N$, where N denotes the total population, which remains constant with time.

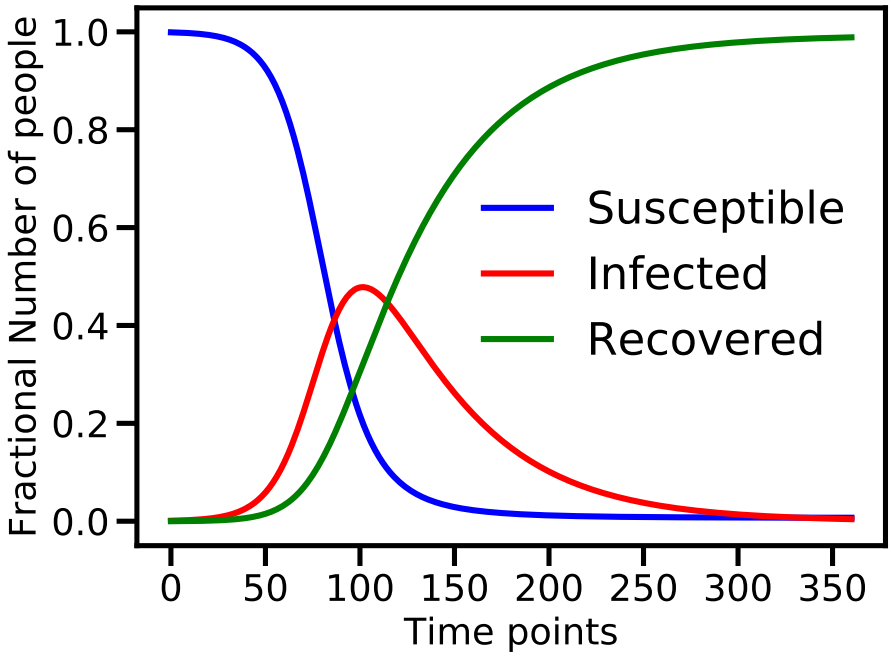


Figure 2.1: SIR Dynamics: Time evolution of the three compartments for some initial condition and parameter values given by $\alpha = 2.5$ and $\beta = 0.5$. We note that almost all people in the population are infected by the end of the evolution even with a small number of initial infected cases.

Figure (2.1) shows the time evolution of the fractional number of people in each compartment. It is important to note that the curve for the infected fraction is not symmetric about the peak. We see that there is an exponential growth of infected fraction in the initial period. Furthermore, after crossing the inflection point, the growth rate starts decreasing before finally reaching zero at the peak. We can understand this by looking at the second equation in Eq. (2.1). At very short times, $S(t) \approx N$, effectively making, $\partial I(t)/\partial t \approx (\alpha - \beta)I(t)$. Since, $\alpha > \beta$, we see exponential growth at very short times before the assumption breaks down.

The above approximation has another important significance. Notice that the rate of change of

infected cases can be greater than or lesser than zero, based on the sign of $(\alpha - \beta)$. If $\alpha > \beta$, there is an exponential growth of infected cases, while if $\alpha < \beta$, there is an exponential decay. Both these cases have a very different outcome, as there is an epidemic in one case, and in other cases, the infection never takes off. In literature, this is often written in terms of $R_0 = \alpha/\beta$, where R_0 is called the *reproduction number*. Intuitively, R_0 denotes the number of people each infected person infects further on an average. We can easily see that if a person infects more than one person on average, the infection will take off, and it will die down if each person infects less than one person on average. From a statistical physics point of view, $R_0 = 1$ denotes the phase transition point for the system. The observable here is the number of recovered people after a very long time. The existence of a threshold point for the SIR model has important and practical consequences for mitigating the spread of the pandemic [45]. We will come back to the phase transition aspect later.

2.2.3 SIR Model with Metapopulation

One inherent drawback of the SIR model is the assumption of a 'well-mixed' population. Spatial heterogeneity makes the assumption of equal probability of infection unrealistic. Considering a country of 1.3 billion people distributed in thousands of cities and towns as a well-mixed population is a gross over-simplification, and we can certainly add a layer of complexity to this simple setup. Introducing metapopulation helps us achieve this goal. A *metapopulation* is defined as a group of interacting species which are spatially separated. This division of scales allows one to explore many complex systems by allowing different interactions between and within the groups. [11]

As the first deviation from a simple SIR model, we consider the SIR model on metapopulation. We consider each city as a node of a network, and individuals can travel between the cities. Within a city, we consider that the population is well-mixed. This type of approach has shown to be very useful in predicting the risks to the cities in the literature [12, 15]. We will now rationalize the travel between two cities based on various real-life data and algorithms. Note that the next few steps are heavily derived from the *Science* paper by Brockmann and Helbing [19], which was one of the primary motivators for this problem.

We will look at the movement kinetics separately before adding in the SIR dynamics. We consider a directed network of cities connected by links. The link strength denotes the number of people moving along an edge in a particular direction. The rate of number of individuals moving from city n would be proportional to the number of individuals in city n (*i.e.* N_n). We are interested in finding $\mathscr{P}(m,t+\Delta t|n,t)$ – conditional probability that an individual in city n at time t is in city

m at time $t+\Delta t$. The conditional probability can be written in terms of jump rate as $\mathscr{P}(m,t+\Delta t|n,t) = W_i^j \Delta t$, where $\boldsymbol{W}$ is the transition rate matrix specifying the jump rate from city n to city m. Combining these ideas, we can write the movement dynamic equation for some city n as,

$$\frac{\partial N_n(t)}{\partial t} = \sum_m \Big[W_m^n N_m(t) - W_n^m N_n(t) \Big], \quad n,\, m = 1,\, 2,\, ...,\, M, \tag{2.2}$$

where M is the total number of cities in the network. The left-hand side denotes the rate of change of population for city n, while the right-hand side denotes the terms which increase or decrease the population for city n respectively. It is prudent to mention here that summing Eq. (2.2) over n, will give 0 on both sides, as all people starting from some particular city have to end up somewhere. The total population in the network is conserved as there are no source or drain terms. However, the population of city n can change with time. If we imagine a distribution of N_ns as some starting condition, it is easy to see that eventually, the population of each city would stabilize to some value based on the $\boldsymbol{W}$-matrix. The conditions for 'steady-state' existence depend on the form of $\boldsymbol{W}$-matrix, but we assume here that those conditions are met. Thus for the rest of the thesis, unless stated otherwise, we consider the population of each city to be constant with time.

In order to calculate the terms of the $\boldsymbol{W}$-matrix, we rely on real-life data. The underlying idea is to collect the data about the average number of people going from one city to another in unit time and define the transition rate. We also need to rely on the census data to get the population of the cities N_n's. We will look at the details of data collection in Section (2.3), but we will assume that we have all the required data for now.

We now define the traffic matrix $\boldsymbol{F}$, such that F_n^m denotes the number of people going from city n to city m in a unit period of time. Thus, $F_n = \sum_m F_n^m$ denotes the total number of people leaving city n in a unit period of time. Now, the rate at which an individual from city n goes to city m, can be written as, F_n^m/N_n. Substituting $W_n^m = F_n^m/N_n$ in Eq. (2.2),

$$\begin{aligned} \frac{\partial N_n(t)}{\partial t} &= \sum_m \Big[\frac{F_m^n}{N_m(t)} N_m(t) - \frac{F_n^m}{N_n(t)} N_n(t) \Big], \\ &= \sum_m [F_m^n - F_n^m] = [F^n - F_n], \qquad n,\, m = 1,\, 2,\, ...,\, M. \end{aligned} \tag{2.3}$$

In order to know if $N_n(t)$ is constant with respect to time, we either have to rely on the actual data or make some assumption about the structure of $\boldsymbol{F}$-matrix. As we will see in Section (2.3),

$F_m^n \neq F_n^m$. Thus, the movement kinetics is not an equilibrium process (as *detailed balance* is not satisfied). However, the system can be in a steady state. Before we proceed to explain why we expect/require the system to be in a steady-state, it is necessary to note that there is no birth-death process associated with the movement kinetics equation. The total number of people in the network remains constant. One of the main aims of the thesis is to build a model which can predict the spread of any fast-spreading infection in India so that mitigation strategies can be deployed. In such a scenario, we argue that the time scale in which the infection typically spreads is much smaller than the time scales in which the city population increases due to migration or enhanced birth rates [27]. Thus, we can assume that $N(t)$ is constant with time. We can readily see this by looking at the last equation in Eq. (2.3). F^n and F_n, denote the influx and outflux respectively for city n. If outflux is equal to the influx, then the right-hand side is equal to zero and thus $N_n(t) = N_n$.

In order to integrate this movement kinetics into the SIR-model, we make another simplification that the probability of getting infected (or recovered) during the transit is zero. Thus, if a person is susceptible (or infected) when leaving city n, he/she would remain susceptible (or infected) when he/she reaches city m. We can look at this in two ways; the typical time scale associated with getting infected/recovered is much larger than the typical time scale of traveling between any two cities. Hence, on an average most people don't change their state midway through the travel, or at least not in a way we can observe. Another way to look at this is by re-scaling the SIR-model parameters (α and β). If there is some probability of susceptible people getting infected on a particular link, we can say that α has city-dependence or route-dependence. We can think of it as if the en-route infections are incorporated in the variations of α for different cities and routes. However, that is an added complexity and for now we will ignore it. Thus, under the assumption of exhaustive nature of movement and SIR dynamics, we can write down the SIR-metapopulation model equations as,

$$\begin{aligned}
\frac{\partial S_n(t)}{\partial t} &= -\alpha \frac{S_n(t) I_n(t)}{N_n} + \sum_m \Big[\frac{F_m^n}{N_m} S_m(t) - \frac{F_n^m}{N_n} S_n(t) \Big], \\
\frac{\partial I_n(t)}{\partial t} &= +\alpha \frac{S_n(t) I_n(t)}{N_n} - \beta I_n(t) + \sum_m \Big[\frac{F_m^n}{N_m} I_m(t) - \frac{F_n^m}{N_n} I_n(t) \Big], \\
\frac{\partial R_n(t)}{\partial t} &= +\beta I_n(t) + \sum_m \Big[\frac{F_m^n}{N_m} R_m(t) - \frac{F_n^m}{N_n} R_n(t) \Big].
\end{aligned} \tag{2.4}$$

Here, n, $m = 1, 2, ..., M$ and all the parameters are as previously defined. We note that $S_n(t) + I_n(t) + R_n(t) = N_n$, which is constant in time. We can trivially check this by just summing all

the three equations. We will solely focus on the set of Eqs. (2.4) for our analysis in this thesis. To summarize, individuals change their compartment in a particular city following SIR dynamics, and then they travel between different cities following movement dynamics. In Eq. (2.4), these two dynamics can be separated, and each has its associated time scales. The difference between these time scales has an important impact on how we analyze the model results.

In order to proceed ahead from Eq. (2.4), we need to have some information about the $\boldsymbol{F}$-matrix. In the next Section, we will look at the ways and assumptions adopted to construct the $\boldsymbol{F}$-matrix.

2.3 Data Collection

The primary modes of transportation in India are air, rail, and road. We estimate that around 10 million people travel every day in India [Table (2.1)]. However, the major challenge encountered in this project was the unavailability of systematic datasets in the desired format for various modes of transport in India. We collected multiple datasets starting from 6 cities and going all the way up to 446 cities in increasing levels of complexity and difficulty in obtaining them. We skip the details about the unused data and focus on the final dataset for 446 cities for all three modes of transport.

There are two aspects to the data we are collecting. First is the topological aspect that concerns the bare network properties such as degree distribution and clustering. The other aspect is the transportation aspect derived from the properties of the agents traveling on them. Multiple properties such as centrality measures, shortest distances, and edge/node distribution can be defined in this aspect. However, the two properties that we will be looking at are the distribution of *local mobility* and *traffic symmetry*.

The total number of people traveling out of a city is some fraction of the city's population. As we will see in the following few Subsections, this fraction is almost the same across all cities if the data is available or our assumptions in the algorithm for constructing the data constrain the fraction to be the same. Despite this, for the final combined dataset, the fraction is not identical across the cities. We call this fraction – local mobility or simply mobility. Local mobility is not equal to global mobility.

Another important property is the difference between the traffic on edge in either direction. Or, in other words, the symmetricity of the traffic. Even in this case, few of the datasets are symmetric while others are not. As a consequence, the final dataset is not symmetric. As mentioned earlier, this breaks the detailed balance, and the system is not in equilibrium for the movement kinetics. We will now look into the details of the data.

2.3.1 Census

We rely on the official census data from the Government of India collected in 2011 to get the population of various cities [27]. First, we made a list of all the cities/towns with a population higher than 0.1 million. After that, we compared the names occurring in either the airway dataset or the railway dataset. Once we had a curated list of all cities having a population of more than 0.1 million, connected either by air or railway, we used that to generate the road data. Thus, in the end, we had a list of 446 Indian cities having a population greater than 0.1 million as per the 2011 census and connected either by rail, air, or road transport. We will now look at the individual datasets before looking at the combined dataset properties.

2.3.2 Air Transport

The passenger data used for airlines was estimated using the monthly air traffic statistics published on a private website [46]. This data was collected for the time duration between January-2018 to December-2018. It directly gave the number of passengers traveling between 85 Indian cities by airplane. We get the total passengers to be around 0.75 million per day as given in table (2.1), which is comparable to the total statistics given by the Directorate General of Civil Aviation, Govt. of India (DGCA) website [47]. The estimated passenger number for the duration between March-2018 to April-2019 was approximately 0.78 million. The slight difference can be attributed to the different time periods during which the data was collected and traffic through small airports.

2.3.3 Railway Transport

The raw data for railways was in the format of $\sim$ 8000 forward and backward routes with more than 3000 stations [48, 49]. As mentioned earlier, we included only those Indian cities with more than 0.1 million population as per the 2011 census. Thus, a few smaller sub-stations in the vicinity of major metropolitan cities were also treated as separate cities. Since we could not get direct traffic data, we had to rely on two assumptions to get the traffic data from train routes and populations,

1. The train is always full.
2. The number of people traveling between any two cities (not necessarily consecutive) by a single train is somehow proportional to their respective populations.

We have the information about the route, population, weekly frequency, and the type of train on each of the routes [49]. We use the knowledge about the train type to fix the capacity of each

train. We will now delve into the details of the algorithm.

Let us consider a particular route (Γ) with k stations labeled by indices 1, 2, ..., k. The population of the cities is N_1, N_2, ..., N_k respectively. Assumption 1 tells us that the train starts with full capacity from city 1, which we denote by C. According to assumption number 2, these 'C' people will alight at remaining stations proportional to the population. Going back to our first assumption, the number of people getting on the train at any city j ($j \geq 2$) would be the same as the number of people getting down at that station.

Now, the number of people getting on the train at city 2, will again get down at cities j ($j \geq 3$), proportional to their population. The number of people getting down at city 3 would be dependent on number of people who got on the train in city 1 and 2. And so on and so forth. Thus, we will continue this process until city $k-1$, and finally 'C' number of people will get down at city k. We now write the number of people getting on the train at station i, *i.e.* F_i in a more mathematical way as follows,

$$
\begin{aligned}
F_1 &= C, \\
F_2 &= F_1 \, \frac{N_2}{\sum_{j=2}^{k} N_j}, \\
F_3 &= F_1 \, \frac{N_3}{\sum_{j=2}^{k} N_j} + F_2 \, \frac{N_3}{\sum_{j=3}^{k} N_j}, \\
F_4 &= F_1 \, \frac{N_4}{\sum_{j=2}^{k} N_j} + F_2 \, \frac{N_4}{\sum_{j=3}^{k} N_j} + F_3 \, \frac{N_4}{\sum_{j=4}^{k} N_j}, \\
&\vdots \\
F_a &= \sum_{i=1}^{a-1} \left[F_i \, \frac{N_a}{\sum_{j=i+1}^{k} N_j} \right].
\end{aligned}
\tag{2.5}
$$

We note that these equations have a recursive form. Luckily, we can simplify each of them to get a simple form for F_a, which denotes the total number of people leaving city a. We get,

$$
\begin{aligned}
F_a &= F_1 \, \frac{N_a}{\sum_{j=a}^{k} N_j} = C \, \frac{N_a}{\sum_{j=a}^{k} N_j}, && \text{for,} \quad 2 \leq a \leq k-1, \\
F_a &= F_1 = C, && \text{for,} \quad a = 1.
\end{aligned}
\tag{2.6}
$$

We can quickly verify this for F_3 as follows:

$$
\begin{aligned}
F_3 &= F_1 \frac{N_3}{\sum_{j=2}^{k} N_j} + F_2 \frac{N_3}{\sum_{j=3}^{k} N_j}, \\
F_3 &= F_1 \frac{N_3}{\sum_{j=2}^{k} N_j} + F_1 \frac{N_2}{\sum_{j=2}^{k} N_j} \frac{N_3}{\sum_{j=3}^{k} N_j}, \\
F_3 &= F_1 \frac{N_3}{\sum_{j=2}^{k} N_j} \left[1 + \frac{N_2}{\sum_{j=3}^{k} N_j} \right], \\
F_3 &= F_1 \frac{N_3}{\sum_{j=2}^{k} N_j} \left[\frac{\sum_{j=2}^{k} N_j}{\sum_{j=3}^{k} N_j} \right], \\
F_3 &= F_1 \frac{N_3}{\sum_{j=3}^{k} N_j}.
\end{aligned} \tag{2.7}
$$

Finally, the number of people going from city a to city b for a particular route (Γ) would be,

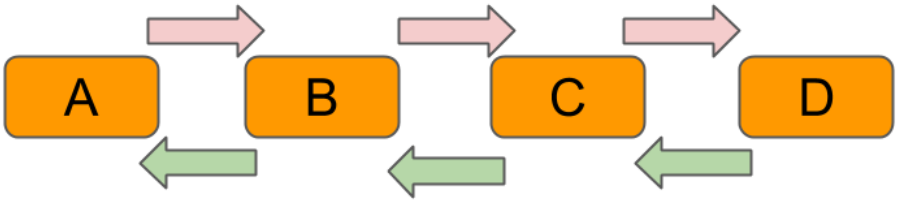
$$
\begin{aligned}
F_a^b(\Gamma) &= \left[C \frac{N_a}{\sum_{j=a}^{k} N_j} \right] \left[\frac{N_b}{\sum_{j=a+1}^{k} N_j} \right], && \text{for,} \quad 2 \leq a < b \leq k, \\
F_a^b(\Gamma) &= \left[C \frac{N_b}{\sum_{j=a+1}^{k} N_j} \right], && \text{for,} \quad 1 = a < b \leq k.
\end{aligned} \tag{2.8}
$$

Thus, we have obtained the expression for the number of people traveling from city a to city b for a given route (Γ). There is only one free parameter in the above expression $F_1 = C$, which we fix by knowing the type of train and its capacity. We can run the same algorithm for all the routes in our raw dataset to get the full $\boldsymbol{F}$-matrix for the railway as the mode of transport. It is pertinent to note that the second assumption is not precisely followed for all pairs of cities. The exception in the mobility of the starting city comes from fixing the train capacity, hence the outflux not being proportional to the population. The expression includes city populations of remaining cities in the route and hence, is different depending on which way the train is going. Thus, the asymmetry is introduced in the traffic matrix, and there is a spread of local mobility values even when we just one route.

We get the total passenger flux to be around 8.8 million after running the above algorithm on our dataset of 435 cities. As per the Indian Railways facts and figures, [50], the non-suburban passengers' traffic was 3.55 billion in 2016-17, which gives daily non-suburban traffic of around

9.7 million passengers per day. This difference can be explained by the fact that we have not included all the smaller stations in our dataset, and many times, the trains run with more than their designated capacities. We illustrate this algorithm more visually and straightforwardly, using an imaginary toy route and population in Figure (2.2).

Let us consider a train route going through 4 cities A, B, C, and D, having equal population N. We will assume that the capacity of the train is 120. Figure (2.2) shows the final results after running the algorithm for the forward and backward routes. As we can readily see, the $\boldsymbol{F}$-matrix is not symmetric even if we consider the simplest case of just one route through cities of equal population. It is also worth noting that the number of people traveling in/out of a city is not proportional to the city population. The assymetricity of $\boldsymbol{F}$-matrix and unequal local mobility become more prominent when we consider all cities and all the routes.



Capacity=120

	A	B	C	D
A	0	40	40	40
B	60	0	20	20
C	20	20	0	60
D	40	40	40	0

Figure 2.2: Model Train Algorithm: We consider two routes here: A-B-C-D and D-C-B-A. For simplicity, let us assume that the population of all four cities is equal and the train's capacity is 120. We run the algorithm using Eq. (2.8) and each entry in the table T_{ij} specifies the number of people traveling from i to j. The color in the table corresponds to the route in the upper part of the figure.

2.3.4 Road Transport

In this Subsection, we will look at the algorithm used to generate road traffic. To our knowledge, there is no systematic source of road traffic data between major cities in India. The National Highway Authority of India (NHAI) collects toll booth data for all major national highways [51]. However, as per their records, they do not account for all the road traffic in India. Moreover, frequently, the state highway records are not well maintained. All of this inspired us to build our algorithm to generate road traffic. We note that we have a list of 446 unique cities connected by railway or airway. We consider only these cities to include in our dataset for roadway traffic. First, we will look at the underlying assumptions that we have used to generate the road data.

As a most straightforward case, we assume that the final $\boldsymbol{F}$-matrix is symmetric. Note that this condition was not satisfied for the railway data. We can move away from this assumption; however, the algorithm to make sure that the population of each city remains constant becomes more complex.

The next assumption we make is that most people use the road to travel for short-distance and long-distance travel is usually undertaken through railways or by airplane. Unfortunately, we do not have any data to support this. Furthermore, we make the final assumption that the number of people traveling by road is proportional to the city's population. We will later see how these specific assumptions help us fix the free parameters in our algorithms under the given set of constraints.

The algorithm consists of two steps. First, we create the adjacency matrix and fix the connections between the cities. Once we have the adjacency matrix, we run another algorithm (similar to the train algorithm) to generate the $\boldsymbol{F}$-matrix. First, we get the latitudes and longitudes of all the cities in our dataset using the *geopy* library in python. Once we have the coordinates, we draw an imaginary circle with some specific radius around each city. We then create the adjacency matrix by checking which cities fall into each imaginary circle by measuring distance on a sphere.

Once we have the adjacency matrix, we sort it according to the population. We start with the city with the lowest population. We get the number of people traveling from this city by fixing a value for desired mobility and then distributing the traffic proportionally to the connected cities' population. We then make symmetric entry in the $\boldsymbol{F}$-matrix. Next, we move on to the city with the second-lowest population. We get the number of people traveling by multiplying the population with the desired mobility. We subtract those people who are already accounted for through the symmetric entry in the first step and distribute the rest proportionally. We repeat this process until we reach the end. If the number of people accounted for symmetric entry is higher than the desired mobility, we increase the mobility for that particular city by some amount.

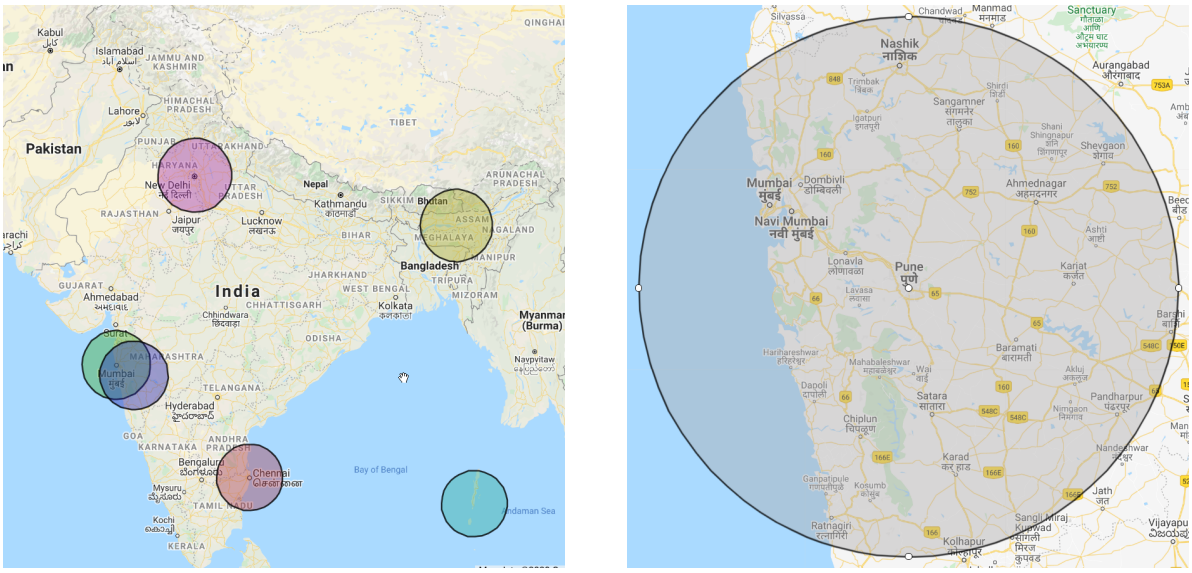


Figure 2.3: The relative size of 200 Kms circle for scale on map of India. On left-hand side, we draw the circles around Mumbai, Pune, Delhi, Chennai, Guwahati, and Port-Blair. The right-hand side figure shows the circle for Pune. We can identify few of the important cities in Maharashtra which are connected to Pune by road from our algorithm.

In order to understand this algorithm more clearly, we give a toy example to illustrate our case in Figure (2.4). We consider a network of 5 cities, such that few links are missing. The population of the five cities is in proportion as follows: A : B : C : D : E :: 10 : 25 : 50 : 75 : 100. Let us suppose that the population of city A is 0.1 million, and desired mobility is 0.0175. The algorithm proceeds as follows:

- First, we get the adjacency matrix whose indices are sorted according to the population. We start with the least populous city, i.e., A, and distribute the desired number of people into all possible connections (blue colored) in the first row.
- We make symmetric entries in the first column (blue), subtract that number from desired traffic for city B and distribute the rest into possible connections (red) in row 2. We make a similar symmetric entry in red in column 2.
- For city C, we proceed as before, subtract, distribute, and make symmetric entries.
- We continue the algorithm until the table is filled.

Coming back to our original original network 446 cities, the parameters that we used in our algorithm are as follows:

- desired mobility $\gamma = 0.015$.

	A	B	C	D	E	Final	Desired
A	0	250	500	0	1000	1750	1750
B	250	0	0	1768	2357	4375	4375
C	500	0	0	3536	4714	8750	8750
D	0	1768	3536	0	7821	13125	13125
E	1000	2357	4714	7821	0	15892	17500

Figure 2.4: Road Algorithm: The $\boldsymbol{F}$-matrix for the toy model is symmetric and the final traffic, matches exactly with desired traffic for most cities. We also note that for the smaller cities, the mobility condition will almost always be satisfied.

- radius of circle = 200 Kms.
- The increase in traffic in case the symmetric entries already sum up more than the desired traffic was ($0.5\times$ desired traffic).

We now mention a few of the main results of the $\boldsymbol{F}$-matrix for road transport.

1. We increased the mobility of only 6 (out of 446) cities to account for overflow of traffic due to symmetric entries.
2. The cities have 20 connections on an average.
3. 92% (or ≈ 410) cities have a local mobility of 0.015. The average (global) mobility is 0.0115 and the standard deviation of the same is 0.0021.

We get the daily road traffic to be around 2.5 million. The only source to compare this with is the NHAI toll booth data. Using some assumptions about the Passenger Car Unit and the number of toll booths a typical vehicle passes through while using national highways, we estimate that the National Highways carry somewhere around 2 to 3 million passengers every day [51]. Compared to that, our estimate for local transport seems reasonable.

2.3.5 Combined dataset

We thus have three datasets corresponding to the three modes of transport. We note that each of the datasets gives average traffic per day between a pair of cities. We add all three datasets together to get a combined dataset. The final dataset is neither symmetric nor is the local mobility the same across cities.

The time scales associated with the three modes of transport are very different. Airway and road typically take around 2 to 6 hours, while railways may take up to 24 or more hours to travel from one city to another in our dataset. As we will see later, our approach for predicting the risk depends on the number of people traveling, but it does not touch upon the timescale difference if there are multiple modes of transport. For now, we assume that the time scales for the spread of disease are much higher than the time scales associated with the average traveling time in the network. This assumption might not always be valid, as there might be some highly infectious disease that spreads everywhere in a very brief amount of time. Our approach will not work in such cases, but the time to respond with a mitigation strategy would be far too less. Under these conditions, we consider it reasonable to add the three datasets without modifying the dynamical equations.

In the future, we can look at the effects of virtual competition between the physical speed of travel (dictated by the mode of transport) and the artificial speed of infection (dictated by the number of people). However, in this thesis, we consider that the number of people dominates spreading while completely suppressing the effect of physical speed of various modes of transport.

2.4 Summary

As we come to the end of the Chapter, we summarize the essential points for us to proceed. We started by describing the SIR metapopulation model by touching upon the various aspects of the SIR model and the movement dynamics equation. Equation (2.4) highlights the separation of dynamics of the two processes and will form the basis for all our future computations. This equation lead us to discuss the methods for collecting traffic data. We discussed the algorithms, methods, assumptions, and sources for collecting the traffic data for three modes of transport – air, rail, and road. We summarize the key results of the three datasets in Table (2.1). With all the necessary pieces collected, we are ready to look at the results and analyze them.

Property	Airway	Railway	Roadway	Combined
Number of Nodes	85	435	446	446
Number of Edges	1182	41594	9128	46448
Average Degree	13	95	20	104
Symmetricity of Data	Yes	No	Yes	No
Locality of Mobility	Same	Different	Same	Different
Number of passengers	7.5×10^5	8.8×10^6	2.5×10^6	1.2×10^7
Fraction of total	0.06	0.73	0.21	1.0

Table 2.1: Properties of different datasets. Airway contributes very less to the final dataset and may become important if we consider the speed of transportation. As roadway is locally dominant mode, most of the meaningful long-distance connections would arise from railways.

Chapter 3

Main Results and Practical Aspects

3.1 Introduction

Now that we have collected all the necessary tools and motivated the dynamical equations for the system's evolution, we are ready to look at the results. First, we will define some observables for our model in Section (3.2) and show that D_{eff} correlates linearly with *Time of Arrival* when the latter is defined using an absolute threshold. We will also look at the robustness of this relationship when we vary the parameters of the system. After that, we will look into a few mitigation aspects in Section (3.3), mainly motivated by practical requirements. We summarize the major results in Section (3.4)

3.2 Robustness of Linear Relationship

3.2.1 Phase Transition

It is a well-known result that the SIR model for a well-mixed population shows a phase transition [10] at $R_0 = \alpha/\beta = 1$. We saw an intuitive explanation for this in Section (2.2.2). The mitigation strategies rely on pushing the reproduction number (R_0) to fall below 1. However, once we transition from a well-mixed population to a metapopulation, it is not obvious that the phase transition will still be observed. The local infection may show a phase transition, but there is no reason to believe that even the global infection will show a phase transition.

In order to check this, we first define the observable. We start with a small fraction of infected

in one of the cities. As we can see from Eq. (2.4), the final state of any individual can either be susceptible or recovered. No individual can stay infected for a very long amount of time. Thus, if we let the simulation run for a sufficiently long time, the whole population can be divided into two exhaustive groups – susceptible and recovered. We denote the number of people in the recovered fraction at the end of simulation as R_∞, meaning recovered people after infinite time. R_∞ will tell us the epidemic's severity, as all those who are recovered were infected at some point in time.

Thus, if the epidemic took off, we will see that R_∞ will have a substantial value, while if the epidemic did not take off, R_∞ will be a tiny fraction of the total population. We keep the value of β fixed and vary α so that R_0 covers a range of magnitude. We plot our result in Fig. (3.1).

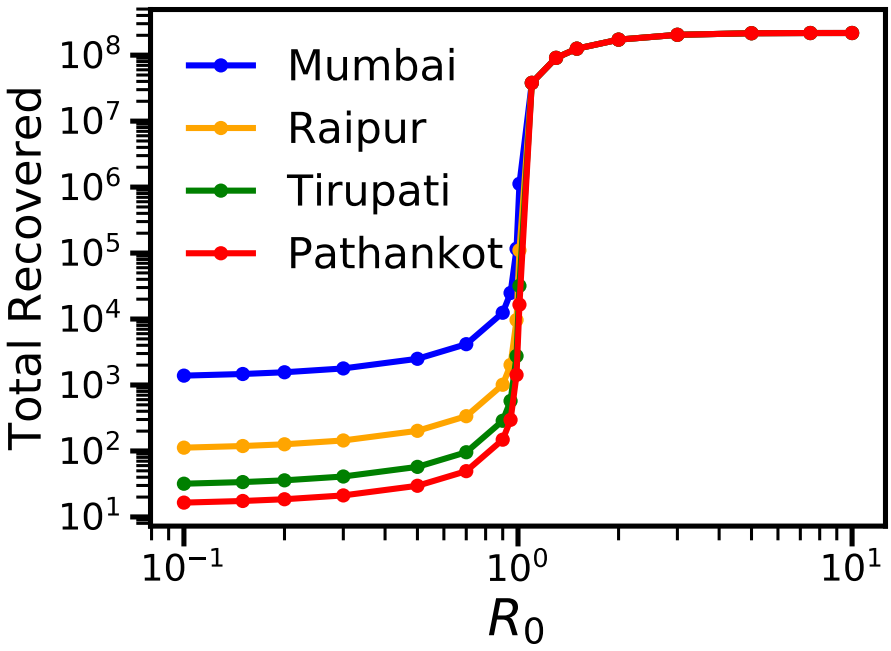


Figure 3.1: Phase transition for SIR metapopulation model. The x-axis is the reproduction number $R_0 = \frac{\alpha}{\beta}$, while the y-axis denotes the total number of the recovered people at the end of pandemic. Note that plot is in log-log scale. We can see a sharp transition for R_∞ around $R_0 = 1$. The four curves correspond to four outbreak locations decided on the basis of population and geographical location. The initial infected fraction is 0.0001 of the respective city's population. $\beta = 0.5$ here.

We see that there is a phase transition for the SIR metapopulation model independent of the outbreak location. As long as $R_0 \leq 1$, the total recovered individuals are not very different from the initial condition. However, as soon as $R_0 > 1$, the total recovered fraction goes up exponentially, and for substantial values of R_0, the dependence on the initial condition washes away completely. Since we are interested in predicting the risk associated with the epidemic, we have to consider

that the epidemic spreads. Hence, we work in the $R_0 > 1$ regime for the rest of the thesis unless stated otherwise.

3.2.2 *Time of Arrival*

As far as we know, the analytical solution to Eq. (2.4) does not exist. Hence, we solve the set of equations numerically using Runge-Kutta 4 (or RK4 for short) method. Starting from some initial conditions, we let the system's state evolve following the dynamical equations given by Eq. (2.4). Thus, we have the fractional number of people's trajectories in each compartment for all cities for all times. Since the main aim of this thesis is to predict the risk, we are mainly interested in a timescale where we can predict and adopt strategies to curb the spread. Thus, we are mainly interested when the number of cases is meager but enough to predict the risk.

As we had mentioned earlier, the SIR metapopulation model has two distinct dynamics – Intracity SIR spread and inter–city traffic spread. When the time scale of the intracity SIR spread is much lower than the time scale of inter–city traffic, we can assume that once an infection reaches a city, it quickly grows into a local outbreak. Thus, we are only interested in a timescale when the number of cases crosses a certain low threshold. The validity of this assumption increases as we increase the infection rate α, as the intracity cases increase much faster than infection spreading to neighboring cities.

Thus, to quantify the risk associated with a city, we need to fix a threshold for infected cases and see how long it takes for the city to reach that threshold. We call this the '*Time of Arrival*.' Formally, we define the *Time of Arrival* for city n as the first instance since $t = 0$, when the number of infected cases in a city crosses a certain threshold θ_n, for any outbreak location city m. City m could be the same as city n.

Now, we can define the threshold in multiple ways. Here, we look at two primary ways of defining threshold, which tell us about very different scenarios.

1. **Fractional Threshold:** The threshold θ_n for city n is dependent on city population N_n, such that $\theta_n/N_n =$ constant. Thus, we are looking at a case when a fixed fraction of each city gets infected. This scenario is not intuitive as it treats 100 cases in a city with a population of 1 million, the same as 1000 cases in a city with a population of 10 million. The latter would create more fear in the minds of the public than the first case. However, if we assume that all the resources scale with city population, then the fractional threshold is a good way of predicting the risk. The disadvantage is that the risk for smaller cities can be exaggerated. Since the traffic data is averaged over, the fluctuations are overlooked. The fluctuations might

have a drastic effect on the small cities.

2. **Absolute Threshold:** This leads us to think of another way to define the threshold θ independent of the city's population. We consider a fixed number of infected cases as the threshold for a city to be called infected. We see that this threshold is inherently skewed towards cities with high populations. The trajectory of fractional infected cases in a city follows almost the same trajectory across all cities. Thus, if we put an absolute threshold, the infected fraction crosses that threshold much earlier for bigger cities. However, this is also a more publicly accepted approach. The risk is usually perceived in terms of absolute numbers.

3.2.3 D_{eff}

In the last Section, we defined *Time of Arrival* to facilitate assigning risk to each city. Once we know the initial condition, we can numerically solve the differential equation and get the trajectory of infected cases in all cities. By choosing either of the thresholds, we can get the *Time of Arrival* for each city. Arranging the *Time of Arrival* in ascending order will give us the risk of each city in decreasing order. Technically, this is the list we started to seek. However, we cannot produce this list unless we know the initial condition precisely, which might not always be possible. We need a way to predict the risk without actually knowing the initial condition.

The concept of '*effective distance*' was first introduced in [19]. It is a well-known fact that the spread of infectious diseases does not follow any particular pattern with respect to the spatial topology [18]. The risk of a city is not necessarily high (or low) just because it is geographically closer (or far away) from the outbreak location. Various network topology parameters like degree, centrality measures, and other neighborhood indices were tried to predict the risk, but with no success [15]. Helbing and Brockmann came up with a non-trivial yet intuitive concept to define a measure of distance on the network in [19] which could successfully predict the risk of contagion spread to an accuracy that was never achieved before. We will now look into the details of this idea.

D_{eff} is a probabilistically motivated way to define distances on a network. Since individuals travel and carry the infection from one city to another, the risk is higher if more people travel. However, we are interested in only the relative number of people traveling, as the final goal is to get a relative measure of hazard associated with each city. As mentioned earlier in the movement dynamics equation, the individuals jump from one city to another at some rate. We consider two probabilities associated with this process.

1. Jump Probability γ_n: As mentioned in sec. (2.3), not everyone in the city travels on a daily basis. Hence, we can define the jump probability as $\gamma_n = F_n/N_n$, as the probability that an individual from city n travels outside the city in unit amount of time.
2. Travel Probability P_n^m: Given that an individual is traveling outside the city, we can define an associated probability for traveling along a particular edge $P_n^m = F_n^m/F_n$.

Note that $\gamma_n P_n^m = (F_n\, F_n^m)/(N_n\, F_n) = F_n^m/N_n$, which gives us the probability that an individual from city n will travel to city m in unit time. We can define effective distance for adjacent cities in the following two ways,

$$\begin{aligned} d_{\text{eff}}^1 &= 1 - \log(P_n^m), \\ d_{\text{eff}}^2 &= 1 - \log(\gamma_n P_n^m). \end{aligned} \tag{3.1}$$

Now, $0 \leq \gamma_n P_n^m, P_n^m \leq 1$. When we take the negative log of both these quantities, $0 \leq -\log(\gamma_n P_n^m),\ -\log(P_n^m) < \infty$. Thus, $1 \leq d_{\text{eff}}^1, d_{\text{eff}}^2 < \infty$. Here we have dropped the indices m and n in $d_{\text{eff}}^1, d_{\text{eff}}^2$ for brevity.

If two cities are not adjacent, $d_{\text{eff}}^{1,2}$ would be ∞. However, we know that as long as the two cities are connected, the infection will spread to them in a finite time. In order to overcome this shortcoming, we define the effective distance in a more general way. We define a set $\{\Gamma\}$ which consists of all possible paths from city n to m. We define $\lambda(\Gamma)$ for any two cities n and m (not necessarily adjacent) as the sum of d_n^m's for all the pairs of cities along a path. Thus, for a given pair of cities, there would be multiple distances of varying lengths. We define the effective distance $D_{\text{eff}}^{1,2}$ as the minimum out of all possible distances,

$$D_{\text{eff}}^{1,2} = \min_{\{\Gamma\}} \lambda(\Gamma). \tag{3.2}$$

We propose that the effective distance proposed in Eq. (3.2) is a good indicator to predict the risk associated with each city for a given outbreak location. We can define $\boldsymbol{D}$-matrix, as the distance matrix, which gives the *effective distance* between any two cities. We manually put the diagonal entries equal to zero. Note that the $\boldsymbol{D}$-matrix is not symmetric.

We now have three measures of distance for the network, namely, Geographical distance (D_{geo}), D_{eff}^1, and D_{eff}^2. We also have two definitions of *Time of Arrival* for the infection defined

in Sec. (3.2.2). In order to see which one fits the best, we linearly regress the *distance* with the *time* and see which one shows the best fit. We quantify the fit by calculating the regression coefficient R^2, also known as the coefficient of determination. $0 \leq R^2 \leq 1$. If $R^2 = 0$, the two datasets are completely uncorrelated, while if $R^2 = 1$, the two datasets are exactly correlated. Thus, a higher R^2 value would indicate a better fit.

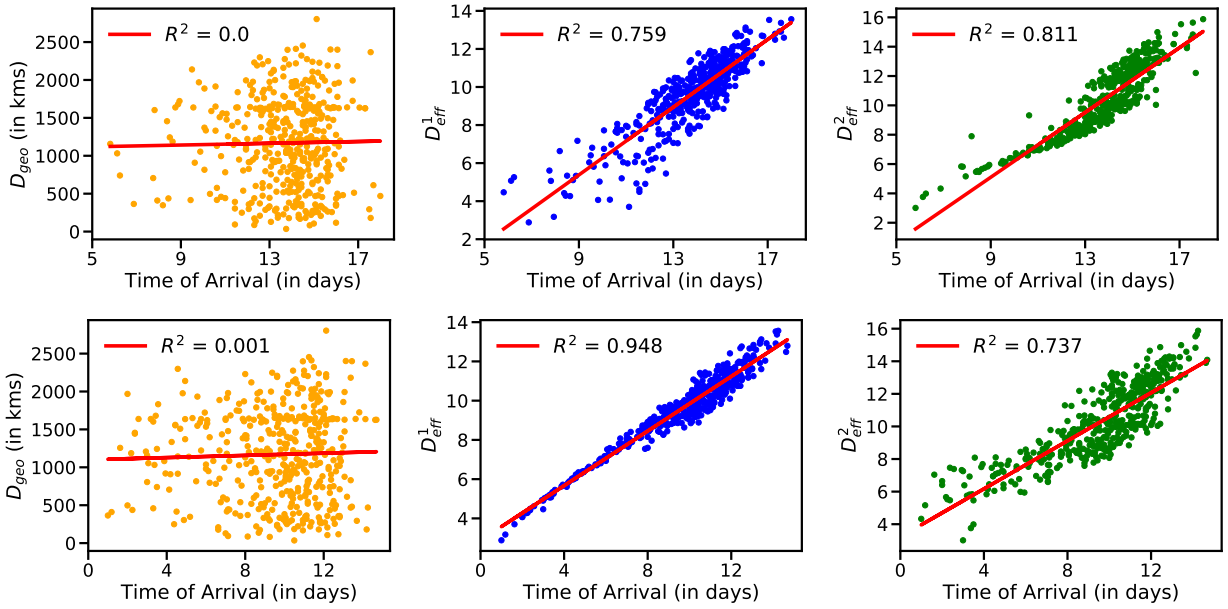


Figure 3.2: Comparison of scatter plots: The upper row denotes fractional threshold (0.0005), while the lower row denotes absolute threshold (10). The first column is with respect to geographical distance, while the second and third column are with respect to D^1_{eff}, and D^2_{eff} respectively. $\alpha = 1.5, \beta = 1.0$. The outbreak location is Mumbai here. The red line denotes the best fit line for the scatter plot, while the legend denotes the value of R^2 for each linear regression.

As we can see in Figure (3.2), D^1_{eff} correlates the best with the *Time of Arrival* when the latter is defined using an absolute threshold. However, we see the difference in the fit for only one outbreak location. To be more certain, we need to see the best fit for more outbreak locations.

In order to settle this, we rely on averaging the value of R^2 for various outbreak locations. We fix the values of α, β, θ, and the outbreak location. We then use different definitions of D_{eff} to see which one correlates the best with different definitions of *Time of Arrival* . We then average the value of R^2 over all outbreak locations. We look at the trend with respect to R_0 in Figure (3.3).

In figure. (3.3), we see that D^1_{eff} is an objectively better index for *Time of Arrival* defined using absolute threshold, as the average R^2 (always above 0.9) is better than any other pair of definition over a range of R_0. We also note that all indices work better at higher values of R_0. Henceforth,

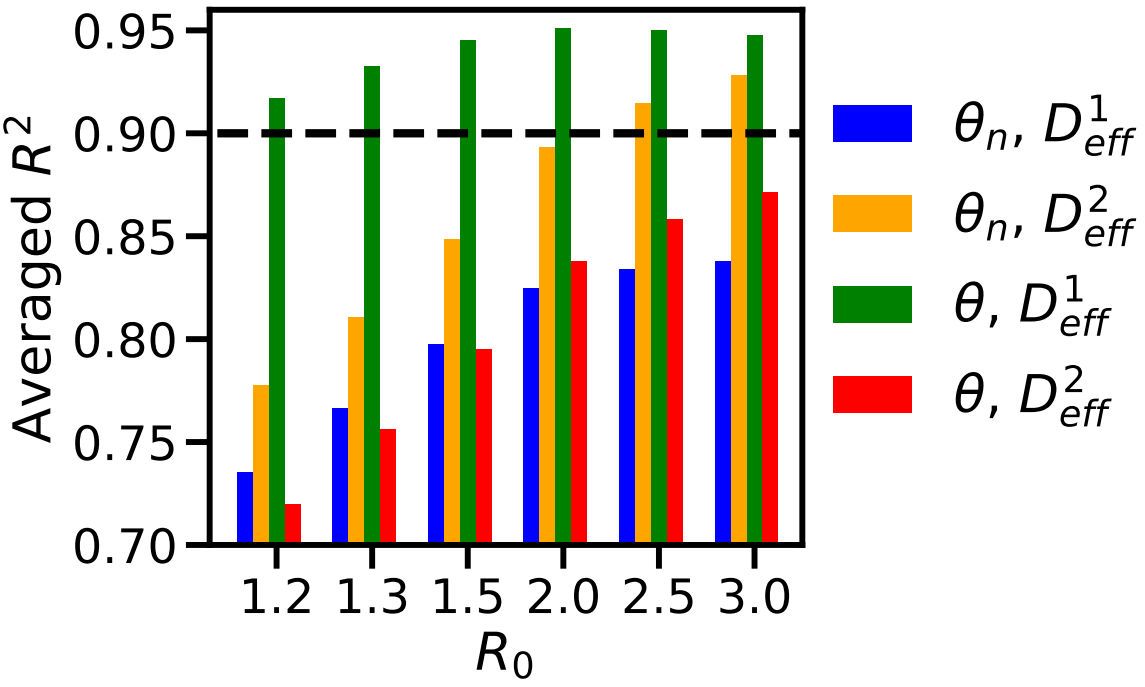


Figure 3.3: Averaged R^2 as a function of $R_0 = \alpha/\beta$. The four cases correspond to two definitions of *Time of Arrival* and D_{eff} each. θ_n corresponds to fractional threshold, while θ corresponds to absolute threshold. The parameter values are $\theta_n = 0.0005$, $\theta = 10$, and $\beta = 1.0$.

we drop the superscript '1' and just use D_{eff} to denote the *effective distance*. We plot few more scatter plots in Figure (3.4) for various outbreak locations of different sizes and locations to show the effectiveness of D_{eff} as the hazard index.

3.2.4 Robustness of D_{eff}

In the last Section, we looked at various ways to quantify the hazard index and settled on D_{eff} defined using the $\boldsymbol{P}$-matrix. However, we only verified the linear relationship between *Time of Arrival* and D_{eff} for selected outbreak locations. Usually, it is hard to predict the outbreak location for the next epidemic, and hence, we would like the hazard index to work for all possible outbreak locations and all parameter values, not just the ones we used in the previous Section.

In order to check the robustness of D_{eff}, we vary the parameters and see if the linear relationship still holds good. The parameters whose values are not known beforehand are α and β. Additionally, the outbreak location and threshold θ can also be regarded as unknown parameters.

We define the robustness by averaging R^2 (Coefficient of determination) over all possible out-

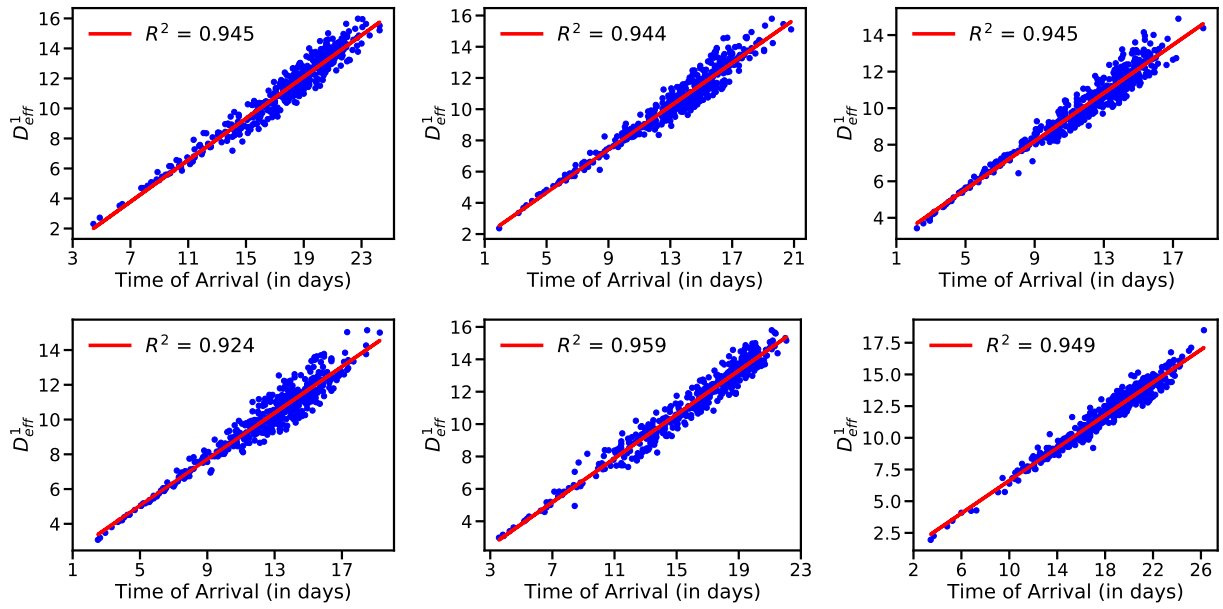


Figure 3.4: Scatter plots between Time of Arrival and D_{eff} for 6 outbreak locations. The Outbreak locations from left-hand side top corner in clockwise order are Ahmadnagar, Bhopal, Chennai, Pathankot, Mangalore, and Jaipur. $\alpha = 1.5, \beta = 1.0, \theta = 10$. We can see that there is an excellant match between the hazard index (D_{eff}) and the hazard (*Time of Arrival*).

break locations. For a fixed set of values of α, β, and θ, we consider the best fit coefficient for a given outbreak location. We repeat this exercise for all cities as outbreak locations. Thus, we have 446 values of R^2 for a given fixed set of parameters. We then calculate the mean and standard deviation of this set. We do this for multiple sets of parameter values and see the mean and standard deviation trends concerning any parameter.

α and $\beta \to$

We vary α and β by keeping the threshold $\theta = 10$. We plot the result in Figure (3.5). We note that for smaller values of α and β, the linear relationship is not as strong as for the higher values. The dependence of R^2 on α and β seems to be non-existent for higher values of α and β. One possible explanation for this phenomenon could be that the timescales of intracity spread for smaller values of infection rates are comparable to the inter–city spread. In such a case, the infection can take multiple routes for spreading, which are not solely determined by the $\boldsymbol{P}$-matrix. Thus, when we increase α and β, the time scales get disjoint, and the spread is mostly dictated by the $\boldsymbol{P}$-matrix, which makes the linear relationship stronger.

We have not plotted the fluctuations in the linear relationship (standard deviation of R^2 over the outbreak locations) in Figure (3.5), but we have verified them to be decreasing with higher SIR

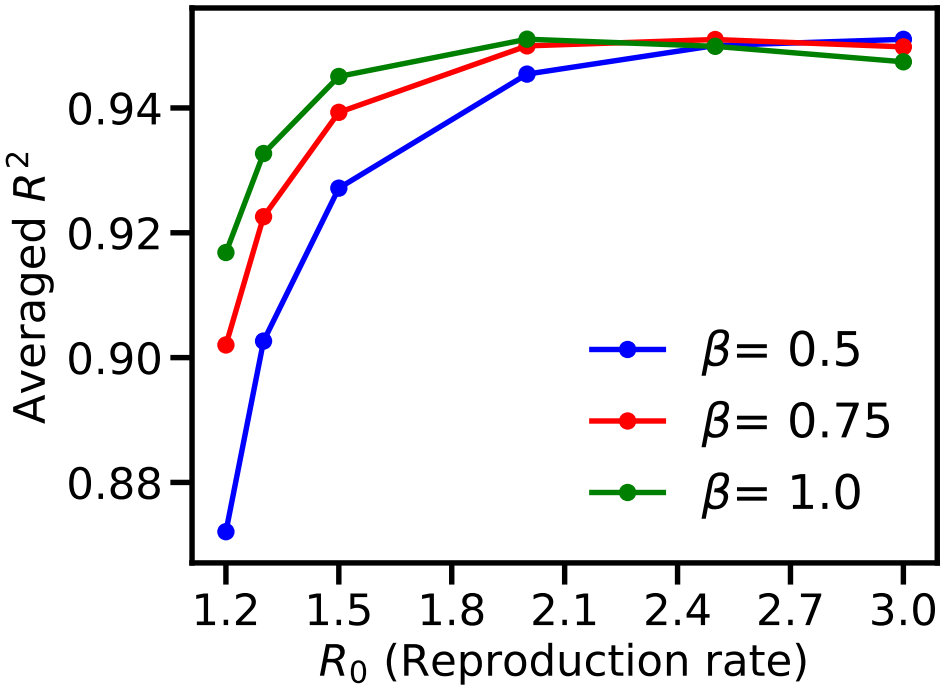


Figure 3.5: Averaged R^2 over all outbreak locations as a function of α and β. $\theta = 10$ here. Note that $R_0 = \alpha/\beta$.

model infection parameters. Thus, the linear relationship is not only weaker for smaller α and β, but it is also weakly dependent on what the outbreak location is. However, there is no particular trend for the strength of the relationship for any natural properties of the system.

Threshold $\theta \rightarrow$

Another critical parameter is the threshold that we use to define the *Time of Arrival*. θ does not appear anywhere in the dynamics and is an artificial as well as a subjective cutoff. However, in order to label any city as infected, we need to define some threshold. We follow the same procedure for averaging R^2 over all possible outbreak locations for various values of θ.

Figure (3.6) shows some very interesting results. First, we note that when R_0 is significantly high, the threshold does not matter. One possible explanation for this observation is that the threshold unexpectedly introduces another timescale in the model. Thus, in the parameter regime that we have chosen, the additional time scale interferes with the existing time scales (determined by SIR rates and $\boldsymbol{P}$-matrix). Thus, for smaller values of R_0 and high threshold values, the time it takes for the infection to spread within a city is of the order of time it takes to spread to its neighbors. Thus, D_{eff} does not dictate the spread effectively. However, for higher R_0, these time scales are different,

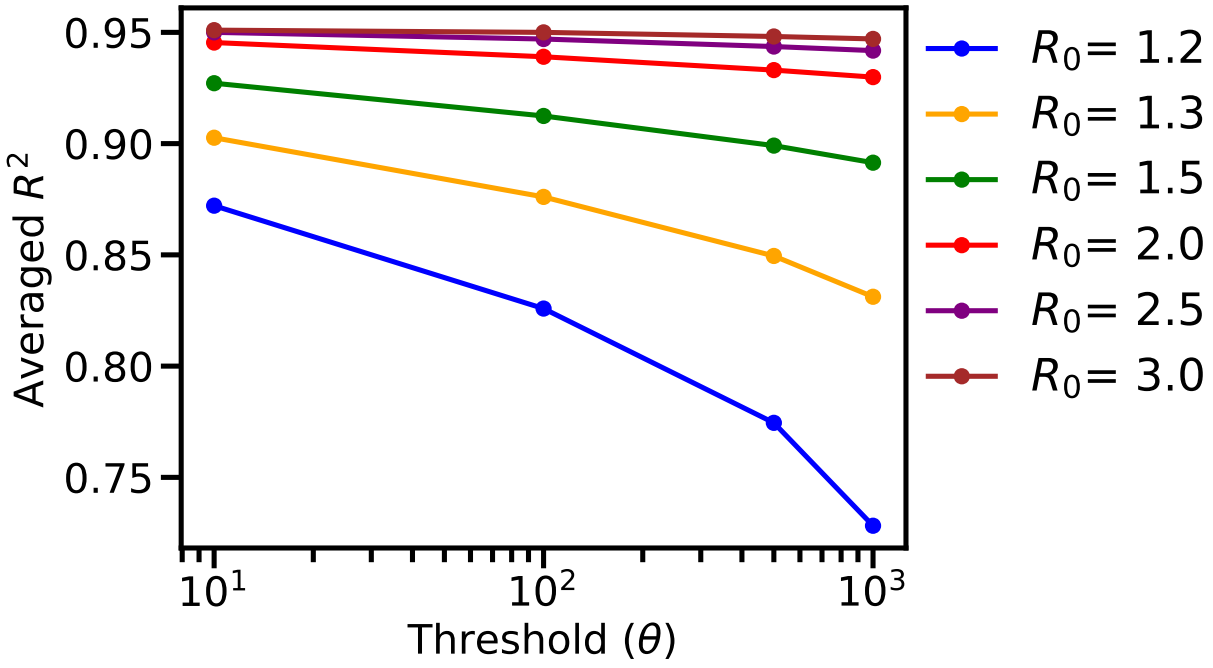


Figure 3.6: Trend of averaged R^2 over all outbreak locations with respect to θ. We keep $\beta = 0.5$. Note that the x-axis is in log scale.

and D_{eff} continues to predict arrival time effectively.

3.2.5 Effective Velocity

So far, we have seen that D_{eff} correlates linearly with the *Time of Arrival* . This allows us to define effective velocity V_{eff} as,

$$V_{\text{eff}} = \frac{D_{\text{eff}}}{\textit{Time of Arrival}}. \tag{3.3}$$

Once we know V_{eff} , we can easily predict the *Time of Arrival* – a quantity of practical concern. However, note that V_{eff} depends on many parameters, like α, β, θ, $\boldsymbol{P}$-matrix, and the outbreak location, all of which might not be known beforehand. Instead, we can look at the average velocity for the network and see if we can predict *Time of Arrival* with incomplete information.

Following a similar approach, we can calculate the V_{eff} for each outbreak by keeping the rest of the parameters fixed. We then take the average of all these velocities and look at its trend for other

parameters. We plot these results in Figure (3.7).

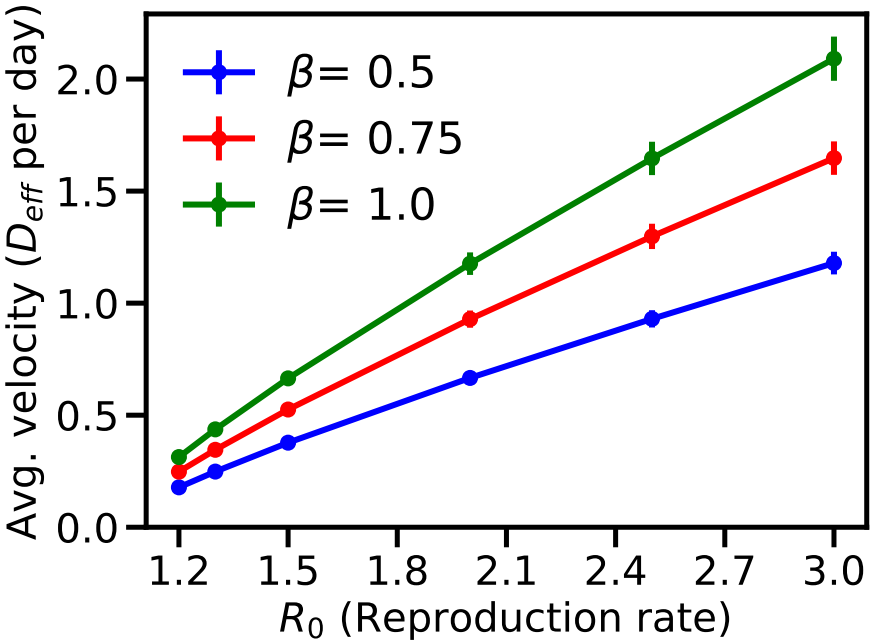


Figure 3.7: Average Velocity with respect to α and β. Threshold $\theta = 10$ here. The vertical bars denote the fluctuation in V_{eff} for a given set of parameters.

The fact that the fluctuations in V_{eff} are negligible compared to the actual V_{eff} has significant implications. It implies that once we know the SIR model rate parameters, V_{eff} is known independent of the outbreak location to a good confidence level. Thus, we can predict the *Time of Arrival* to some extent at the very beginning of the pandemic. We also note that the average velocity increase linearly with α. Thus, even for an unknown SIR infection rate, we can estimate the bounds of V_{eff} beforehand. Similarly, V_{eff} decreases linearly with β (not plotted here). As we will see later, V_{eff} allows us to compare the SIR metapopulation model to other known spreading processes.

3.3 Practical Aspects and Comparison with Real-Life Data

3.3.1 Multiple Outbreak Location

We have considered that the infection starts from one city and then spreads to all other cities. However, as we have seen in the past few pandemics, most of them originated outside India. Thus, there were multiple possible entry points for the infection to enter the country. Since D_{eff}

is outbreak location-specific, it is not clear how to extend this approach to a case when there is more than one outbreak location [52]. Another likely scenario is that the infection arrives in a city from overseas. However, it is not detected until it has spread to few more cities, and the memory of origin is lost. Even in this scenario, it becomes vital to be able to predict the spread. In this Section, we will consider the results of the simplest case of two outbreak locations. We will consider the most straightforward way of extending the definition of D_{eff} and check if it works by running the simulation.

Two outbreak locations can be considered as two competing spreading processes. We saw in Section (3.2.5) that V_{eff} depends on the outbreak location. We also hypothesize that for lower threshold values, the spreading process is dependent on the first arrival of either process rather than the additive effect of the two processes. The first arrival means that whichever infection reaches first dominates the spread in that particular city. The two spreading patterns do not pair up to infect a city. We can see that the effective spreading pattern would depend on V_{eff} of the individual cities.

Since, we hypothesized that the two spreading processes act like independent and disjoint processes, we define a modified D_{eff} as, $D_{\text{eff}}^{\text{mod}} = min\left(D_{\text{eff}}^{OL\,1}, D_{\text{eff}}^{OL\,2}\right)$, where OL stands for outbreak location. For each city that is not an outbreak location, we calculate D_{eff} from either of the outbreak locations and assign whichever is the lowest. We then linearly regress $D_{\text{eff}}^{\text{mod}}$ against *Time of Arrival* and see if there are any interesting features that emerge out of it.

We rely on $D_{\text{eff}}^{\text{avg}}$ to select the cities as outbreak locations. We define $D_{\text{eff}}^{\text{out}}$ (respectively, $D_{\text{eff}}^{\text{in}}$) as the average of D_{eff} from (respectively, to) all other cities from a chosen city. These $D_{\text{eff}}^{\text{avg}}$'s tell us how far the network is from the given city and how far is the given city from the network respectively. We can use $D_{\text{eff}}^{\text{avg}}$ as a proxy for V_{eff} to decide which cities to consider together as outbreak locations.

Figure (3.8) reveals many exciting features. Firstly, we see an excellent symmetry in the overall color of the plot. The plots along the left diagonal (almost identical $D_{\text{eff}}^{\text{out}}$) are well mixed and show equal domination by both outbreak locations. We also notice that there are two dominant spread patterns in some of the scatter plots. Even though the spread patterns are disjoint initially, the slope seems the same. A physically intuitive explanation for the difference in intercept is that the spread pattern is not exactly additive. We see that the city with lower $D_{\text{eff}}^{\text{out}}$ value hugely dominates the spreading pattern along the non-left diagonal elements, and the linear fit is not as good as the left-diagonal elements.

In conclusion, $D_{\text{eff}}^{\text{mod}}$ works well when both the outbreak locations have a similar value of $D_{\text{eff}}^{\text{out}}$. When the cities are unequally close to the network, the closer city dominates the spread. $D_{\text{eff}}^{\text{mod}}$

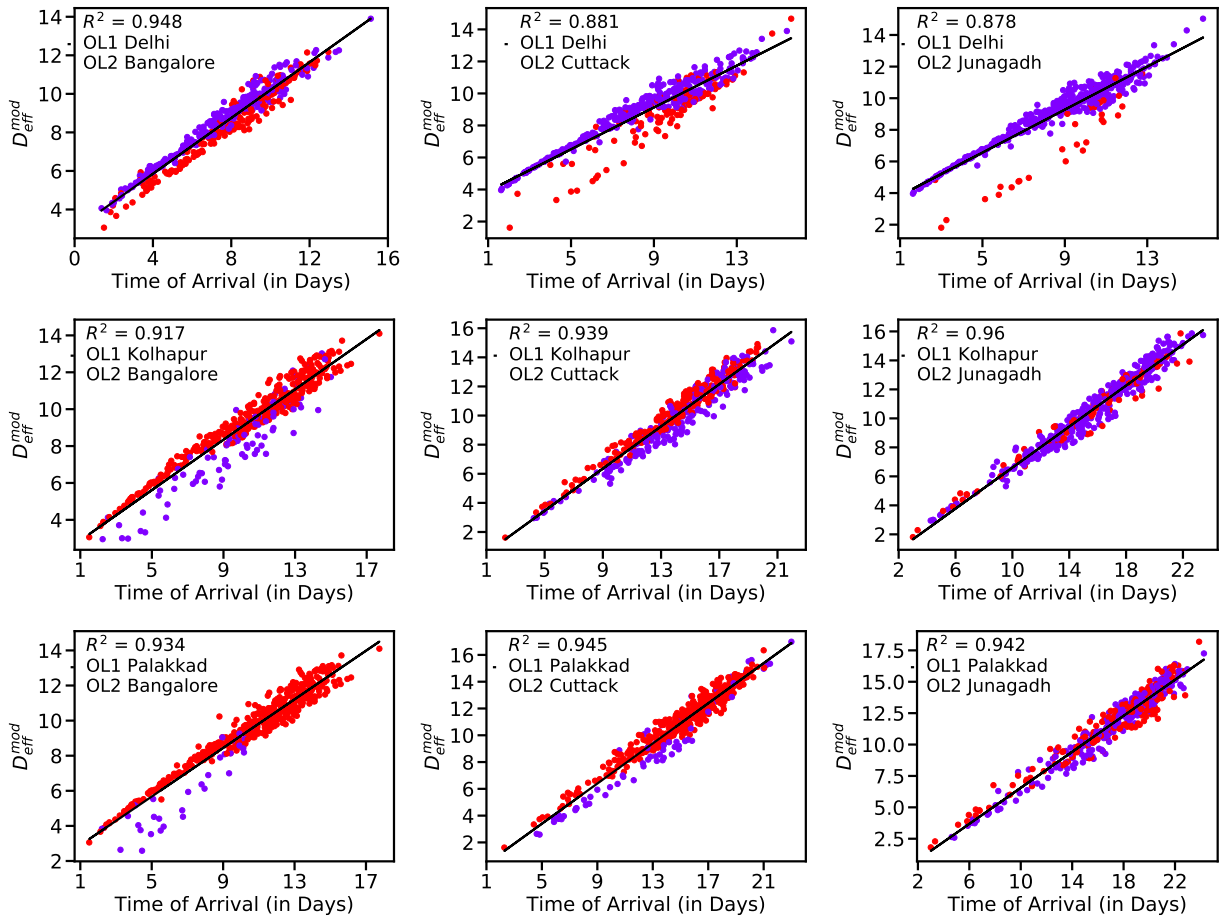


Figure 3.8: Two outbreak locations. Red color denotes outbreak location 2 (OL2) was closer to a given city, while purple color denotes OL1 was closer. Black line denotes the best linear fit with $D_{\text{eff}}^{\text{mod}}$ and *Time of Arrival*. The parameter values are $\alpha = 1.5$, $\beta = 1.0$, $\theta = 10$. Delhi and Bangalore are cities with lowest $D_{\text{eff}}^{\text{out}}$, Kolhapur and Cuttack are with medium $D_{\text{eff}}^{\text{out}}$, while Palakkad and Junagadh have very high values of $D_{\text{eff}}^{\text{out}}$.

does not work as well, but that is fine since D_{eff} can parameterize the spread pattern. We also note that we can extend this approach to cases with more than two outbreak locations. The curious case of non-zero intercept can be analyzed further; however, that is beyond the scope of this thesis.

3.3.2 Link removal

After analyzing the models' properties, we can tweak some aspects of the model to get desired results. Even though predicting hazard is a meaningful exercise in itself, suggesting counter-strategies to curb the spread is far more critical. This Section will look at one such strategy that

may help us delay the spread of infection.

Since our model consists of many parameters, we can change them in any desired fashion to get favorable results. However, we are more interested in practically relevant options. First off, we consider that α, β, and θ are fixed. Even though there are clinical ways to reduce infection rate α and increase recovery rate β, we assume here that those ways have already been implemented, and the SIR rates can not be changed further. The next parameter we consider is γ_n which tells us the fraction of people in city n who travel to other cities daily. We can uniformly scale down γ_n for all cities. The overall mobility in the network is reduced, and the time scale of the infection spread is increased. Nevertheless, it is evident that this approach is very cost ineffective, and we would like to change minimum parameters to achieve maximum change in the *Time of Arrival*.

Another way to decrease V_{eff} is to keep the traffic in the network constant but redistribute it across the edges present. We saw that (results not plotted here) even though the *Time of Arrival* for the at most risk cities increases, there are new cities with increased risk than before, thus keeping the average *Time of Arrival* almost equal. Hence, we need to find a suitable trade-off to change the traffic matrix by not changing too many parameters but still achieving maximum impact. One extreme way reduces traffic on all edges, while the other extreme way redistributes the traffic on all edges. It is easy to see that to reduce the average *Time of Arrival* we have to decrease the overall traffic. The only question that remains to be answered is – 'on what basis?'.

To summarize, we are looking for ways to change the $\boldsymbol{F}$-matrix in a meaningful manner to reduce the speed of infection (or V_{eff}). We can not close down all links in the network as such conditions are incredibly harsh on most people. Additionally, since the spread happens over an extended time, the restrictions are unnecessary most of the time. In other words, we need to find out the essential links in the network, specific to the outbreak location or otherwise.

Before we go into the details of criteria to choose the important links, we need to be aware of two things – $\boldsymbol{F}$-matrix is not symmetric, and the population of each city is conserved at all times. Thus, when we select a particular edge to remove, we stop the traffic only along a particular direction. Hence, the influx is not equal to outflux now, and the population of the city will change with time. Since we mostly consider the threshold of θ as low, removing even select edges will affect the dynamics drastically. In order to go around this problem, we remove the edges along both directions. Now, since the $\boldsymbol{F}$-matrix is not symmetric, the population still changes. However, the time scale for significant population change is much larger now, given that $\boldsymbol{F}$-matrix is not completely asymmetric (the difference between the influx and the outflux for most cities is orders of magnitude lesser than the total incoming traffic).

We use two criteria to select the most critical edges in the network – maximum $\boldsymbol{F}$-matrix entries and maximum link-Salience matrix entries. We digress here to understand the Link Salience matrix $\boldsymbol{S}$.

Link Salience Matrix ($\boldsymbol{S}$)

It is hypothesized that the reason D_{eff} works so well is that it correctly identifies the shortest-path trees in the network. Since epidemic spreading is a 'fastest way wins' process, the shortest path algorithm works well to predict the risk. Even though there are multiple ways for an infection to go from one point to another, it is typically observed that only a few paths dominate [17, 19, 53]. Thus, we have a natural way to quantify the critical links based on D_{eff} .

The concept of Link Salience was first introduced in [42]. In order to find the link salience of a particular edge, we use the following algorithm.

1. Considering a chosen city as a source node, we identify the shortest paths to all other nodes. Thus, we construct the shortest path tree for one city.
2. We repeat the above exercise for all possible source nodes and count the number of times a particular edge occurs in the N (corresponding to N nodes) shortest-path trees, uniquely. Thus, the maximum number of times an edge could occur in the shortest path trees is N. We label the normalized count as Link Salience. In [42], it was observed that the distribution of $\boldsymbol{S}$-matrix (Link Salience matrix) was bi-modal, peaked around two values – 0 and 1 for a wide variety of networks, implying that there are a few edges that help the spread, while most of the other edges do not take part in the spreading process at all.

We calculated the $\boldsymbol{S}$-matrix defined using the above method, but we got very different results than expected. The distribution of $\boldsymbol{S}$-matrix entries was not bi-modal. Rather it was mostly uniform. The existence of uniform distribution pointed out that there was no particular set of edges which was important for the spreading process for all outbreak locations.

We now compare the results for two criteria. We identify the links based on maximum entries of $\boldsymbol{F}$-matrix or maximum entries of $\boldsymbol{S}$-matrix. As a baseline, we also plot the change in time of arrival when links are removed randomly. Since, we are changing the $\boldsymbol{F}$-matrix, we are also changing the $\boldsymbol{P}$-matrix, and hence D_{eff} and V_{eff} too. Thus, the only quantity that we can now compare is the *Time of Arrival* .

Figure (3.9) shows the results when the links are removed. The top figure shows the case where $\sim 2\%$ links were removed out of the total links, while the bottom panel illustrates the result when $\sim 8\%$ of the total links were removed. The red color denotes the case when no links were

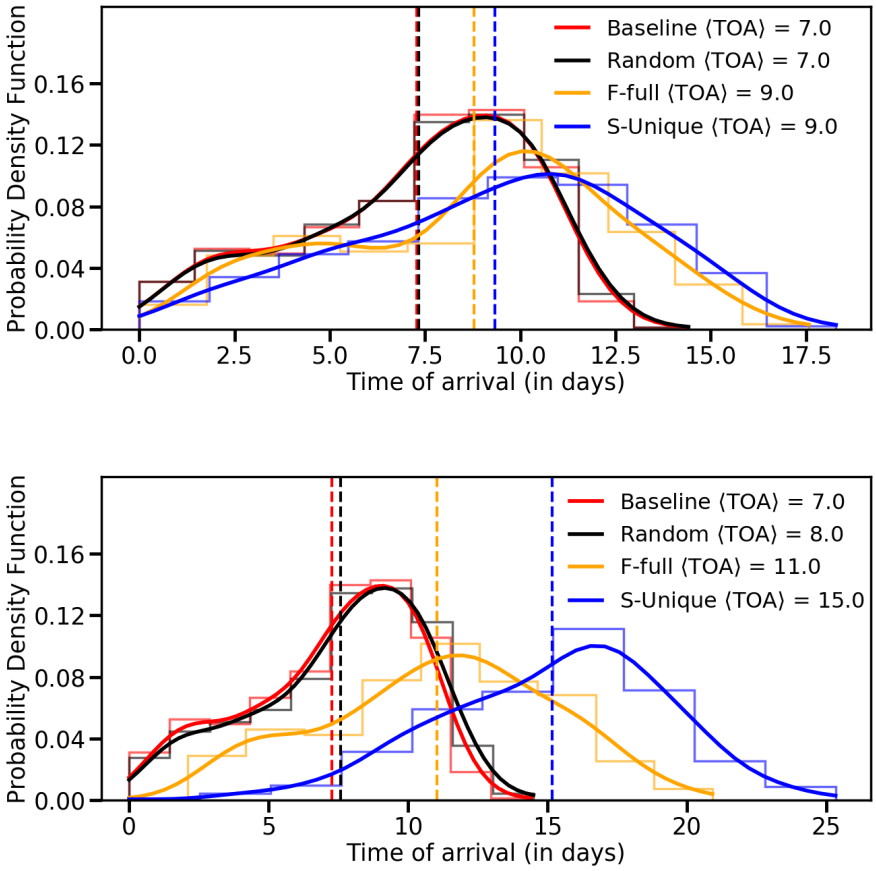


Figure 3.9: Probability density function for the time of arrival with Bangalore as the outbreak location. The top figure denotes the case with 1000 links removed, while the bottom figure denotes removing 3886 links. Various colour denotes the protocol used for removing the link. The dotted lines denote the average *Time of Arrival* after removing the links.

removed. Black color denotes when the links are randomly chosen and removed, yellow when they are selected based on $\boldsymbol{F}$-matrix, and blue when they are selected based on $\boldsymbol{S}$-matrix. The solid line denotes the Gaussian approximation for the PDF, while the dotted lines denote the average *Time of Arrival* for each case. The faded curves are the actual PDFs. $\alpha = 1.5,\ \beta = 1.0,\ \theta = 100$.

For the top figure, the traffic is reduced to 96%, 37%, 52% of the original traffic when the links are chosen randomly, on the basis of $\boldsymbol{F}$-matrix and $\boldsymbol{S}$-matrix respectively. While the same percentages for the bottom figure are 91%, 22%, 35% of the original overall traffic. These percentages are contradictory to our initial prediction about the non-existence of critical links. We see that even with just $\sim 8\%$ of the links removed in a specific way, the traffic is reduced to $\sim 30\%$ of its original

value. One possible explanation for this phenomenon is that even though there are no important links in terms of $\boldsymbol{S}$-matrix, the same is not the case with $\boldsymbol{F}$-matrix. Almost all links help contribute to the spread. However, the majority of the traffic is still carried by a minimal subset of all links. We can refine the results by considering more outbreak locations, but we skip that exercise for now.

We also note that deciding the links based on $\boldsymbol{S}$-matrix is a much better strategy than choosing based on $\boldsymbol{F}$-matrix. If we look at the bottom plot in Figure (3.9), we notice that the shift in PDF is non-linear. It seems that the overall width of the curve has also increased. Thus, removing links based on $\boldsymbol{S}$-matrix changes the mean *Time of Arrival* and reduces the number of cities with very high risk (very small *Time of Arrival*). The maximum *Time of Arrival* is almost increased by 60% ($15 \rightarrow 25$ days), while the average *Time of Arrival* is increased by 100% ($7 \rightarrow 15$ days). Thus, even though this approach is not entirely desirable, it still gives us a good enough criterion to curb the spread of the infection in the network.

3.3.3 Real-life data

As we come to the end of this Chapter, there is one final thing that we need to verify – the match with real-life data. India is one of the worst-hit countries globally in terms of Covid-19 cases [2]. This Section will compare the two waves of Covid-19 cases in India and see how well they agree with our dataset.

The first wave in India was pretty small (compared to the second wave) and spread out over a long period. India was classified as a 'cluster of cases' category, rather than in 'community transmission' – benchmark of epidemics by WHO [1]. The ongoing second wave is even more devastating and is spreading at a much faster speed. Nonetheless we compare the results for both in tables (3.1) and (3.2) respectively.

As we can see in the tables, there is a difference between the observed and predicted cases. Even though many cities appear in both lists, the order is not mismatched. There are many possible reasons for this, and we mention a few of them below.

Firstly, the majority of our dataset is estimated. Even though we have given justification for our algorithms, we have no way of verifying if the data is correct or not. Additionally, the census data is ten years old too, and hence, there might have been some critical changes in the demography that rendered our algorithm suboptimal. The reliability of the real-life dataset is also not very high. We had an incomplete dataset in terms of districts rather than cities. The heterogeneity between the SIR infection and recovery rates for various cities might also be another reason for D_{eff} not to work so accurately. Finally, there was a lockdown during most of the period when the infection

City	D_{eff}	City	Real TOA
Thane	2.88	Delhi	15
Pune	3.18	Chennai	19
Delhi	3.7	Ahmedabad	20
Surat	4.06	Thane	30
Ahmedabad	4.08	Pune	51
Pimpri Chinchwad	4.27	Hyderabad	58
Nashik	4.29	Bangalore	66
Vasai	4.42	Guwahati	73
Vasco Da Gama	4.47	Kolkata	83
Bangalore	4.49	Nashik	91
Hyderabad	4.62	Visakhapatnam	91
Kolkata	4.91	Kurnool	91

Table 3.1: Initially we look at the *Time of Arrival* of all cities and consider the city with the lowest *Time of Arrival* as the outbreak location. Then we look compare the results of D_{eff} from that city with the real *Time of Arrival* . We also note that the real-life data is in the form of daily new infections and we do some transformations to get the number of people infected at a given time. In this table, the starting date is 26$^{\text{th}}$ April, 2020. The threshold for real-life *Time of Arrival* is 5000 infected cases. Mumbai is considered as the outbreak location here.

City	D_{eff}	City	Real TOA
Mumbai	2.17	Mumbai	25
Mysore	3.06	Nagpur	25
Pimpri Chinchwad	3.23	Thane	28
Tumkur	3.67	Nashik	43
Delhi	3.85	Delhi	57
Chennai	3.86	Durg	58
Solapur	3.86	Raipur	61
Thane	4.08	Jalgaon	63
Salem	4.15	Chennai	63
Hyderabad	4.2	Hyderabad	64

Table 3.2: The procedure remains the same as Table (3.1). The starting date for this simulation is considered as 1$^{\text{st}}$ February, 2021. Since Pune and Bangalore were both affected around the same time, we consider both the cities as outbreak locations and use $D_{\text{eff}}^{\text{mod}}$ instead of D_{eff} . The threshold for real-life *Time of Arrival* is 7500 here.

spread and the mobility patterns were affected non-uniformly throughout the country.

3.4 Summary

We started our analysis by showcasing some crucial results for the SIR metapopulation model. After that, we defined the observables D_{eff} and *Time of Arrival* and verified the key linear relationship between the two. We then showcased the robustness of this relationship by varying the parameters of the model. After the theoretical considerations, we looked into more practical aspects of the model and suggested ways in which the spread could be slowed down. We finally compared it with real-life data. You can find more information about the project at [56, 57]. In the next Chapter, we will investigate the linear relationship between D_{eff} and *Time of Arrival* from a more academic perspective.

Chapter 4

The Effectiveness of D_{eff} and Fisher-KPP Equation

4.1 Introduction

In the last two Chapters, we have motivated and given evidence for the effectiveness of D_{eff} to predict the hazard associated with any infection spreading in India using the transportation network. We used the SIR metapopulation model to run the simulation and verify the results. Finally, we also looked at few practical extensions of this model by considering multiple outbreak locations and the effect of tweaking the traffic matrix. Even though D_{eff} works very well in most situations, we do not have a perfect understanding of why it works.

Since we know that the *Time of Arrival* increases linearly with D_{eff} , the kinetics of the spread is ballistic or wave-like. An infected person in one city has a probability of jumping to one of its many neighbors. Thus, if we look microscopically, the dynamics is stochastic, and hence, diffusive behavior leading to wave-like patterns makes more sense. We pursue this direction in this Chapter.

The plan for this Chapter is as follows: First, we will give an intuitive explanation for why D_{eff} works so well in Section (4.2). After that, we will show the similarity between the SI-model and the Fisher Model equations before comparing the results for both. As a natural next step, we will consider diffusion on a network with $\boldsymbol{P}$-matrix as the stochastic jump probability matrix for random walkers and see the connection with diffusion on a line in Section (4.3). Finally, we will summarize the key results from this Chapter in Section (4.4)

4.2 Fisher-KPP equation

The primary reason for D_{eff} to work so well is that the city gets infected with a tiny but earliest inflow of infected cases. The later infections do not matter. D_{eff} , on the other hand, selects the most probable path between the two nodes [17, 19]. Since the physical speed of travel is the same for any traveler between any two nodes, the most probable path is also the fastest path between any two nodes.

Let us think of a toy model which consists of a big central hub (labeled as 'C') and n (> 2) smaller periphery vertices (labeled as v_i) connected only to the central hub. Now, there are n edges in this network. For now, let us assume that each of these edges carry equal traffic both ways. Thus, the population of all $n+1$ nodes is conserved. However, the $\boldsymbol{P}$-matrix for the network is not symmetric. $P_C^{v_i} = \frac{1}{n}$, $P_{v_i}^C = 1$, and rest all entries are zero. Correspondingly the effective distances are $d(C \to v_i) = 1 + \log n$, $d(v_i \to C) = 1$, and $d(v_i \to v_j) = 2 + \log n$ for $i \neq j$. Thus, larger the n, farther are the smaller cities from the hub and also from each other. On the other hand, distance from periphery nodes to hub remains the same independent of n. We illustrate this with an example in Figure (4.1) for $n = 4$.

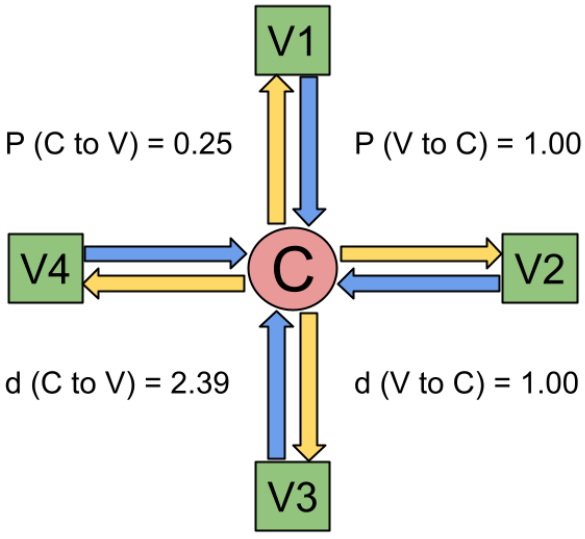


Figure 4.1: Toy Model for 5 nodes. Note that all periphery nodes V_i are equivalent to one another. C denotes the central hub. Here $d(V_i \to V_j) = 3.39$. The probabilities and the effective distances are as shown in the diagram.

Now, suppose the infection starts at the central hub. Assuming that the local mobility $= 1$ for all nodes, the infected person's probability of traveling to one of the periphery nodes is $1/n$. Thus,

the probability of an infected person reaching a node V_i decreases as more nodes are added. We can extend this reasoning to multiple initial infections and a higher threshold for calling a node infected. On the other hand, if the infection starts from a periphery node, it will always spread to the hub in a one-time step, and then it will take some more time to reach another periphery node. D_{eff} precisely predicts all these results. Even though the computations become hard, the idea of D_{eff} can be expanded to more complex network topologies with diverse traffic properties.

We will now try to gather more evidence for the effectiveness of D_{eff} . We start by summarizing the discussion in the *Science* paper [19]. The main hypothesis behind the notion of D_{eff} was that only the most probable path matters. Once the first infection arrives at some city, it quickly grows within the city, rendering the later arrivals irrelevant. Hence, the assumption of the timescale of local proliferation (dictated by α, β) should be much lower than the timescale of global spread (dictated by γ) was important.

However, our dataset differs from the dataset in [19] in many ways. Firstly, there is a significant spread in the mobility for various cities, and $\boldsymbol{F}$-matrix is not symmetric. These assumptions lead to a steady state for movement kinetics instead of an equilibrium state where detailed balance is followed. At this stage, we are unsure if both these properties have any significance on the results.

As a first approximation, we can assume that the recovery rate β is very low. This implies that all cities get infected by the time $R_n \approx 0$ for all n. Thus we can assume the model to be comprised of just two compartments $S_n + I_n \approx N_n$. This assumption effectively reduces the model to just $2n$ equations instead of original $3n$ equations, and conservation of city population ensures that we can specify the evolution in terms of just n equations of I_n. This model is often called as SI model in the literature. One major difference between SI and SIR model is the non-existence of a threshold for the former to become a pandemic. We can write the n equations for I_n, as follows,

$$\frac{\partial I_n}{\partial t} = \alpha I_n\left(1 - \frac{I_n}{N_n}\right) - \gamma_n I_n + \sum_m \frac{F_m^n}{N_m} I_m, \quad n,\, m = 1,\, 2,\, ...,\, M. \tag{4.1}$$

Making the transformation $I_n/N_n \to i_n$ and rearranging some terms we can write the above equation as,

$$\frac{\partial i_n}{\partial t} = \alpha i_n(1 - i_n) + \frac{1}{N_n}\sum_m \left[\gamma_m P_m^n N_m i_m - \gamma_n P_n^m N_n i_n\right], \quad n,\, m = 1,\, 2,\, ...,\, M. \tag{4.2}$$

Equation (4.2) looks very similar to the Fisher-KPP equation given by,

$$\frac{\partial u}{\partial t} = ru(1-u) + D\frac{\partial^2 u}{\partial x^2}, \tag{4.3}$$

where $u(x,t)$ denotes the concentration of a quantity at position x, at time t. D is the diffusion constant, and r is the growth parameter. The family of these equations is also called reaction-diffusion equations. It is known that the Fisher-KPP type of equations have a wave-like solution, even though the exact analytical solution for them have not been found yet [43].

We can discretize Eq. (4.3) in space as,

$$\frac{\partial u_n}{\partial t} = ru_n(1-u_n) + D\sum_m \left[\mathscr{P}^n_m u_m - \mathscr{P}^m_n u_n\right], \quad n,\, m = 1,\, 2,\, ...,\, M, \tag{4.4}$$

where $\mathscr{P}$ is a tri-diagonal matrix, the same as the symmetric random walk 'transition jump probability' matrix. We note that Eq. (4.2) and Eq. (4.4) are very similar in form to each other.

We will now try to show similarities between the two equations by adopting a numerical approach. Since we know that *Time of Arrival* scales linearly with D_{eff} for the SIR metapopulation model, we first check if the linear relationship still holds for the SI metapopulation model. Simultaneously, we will also verify the wave-like solution for the discrete Fisher-KPP equation by similarly defining the *Time of Arrival* and using the lattice spacing as a proxy for D_{eff} . We plot the scatter plots for both these models in Fig. (4.2).

We need to keep in mind several subtle points while comparing the SI model and Fisher-KPP equation in Figure (4.2). Even though the forms of equations are similar, there is heterogeneity in the SI model. Additionally, the SI model has more associated variables with the process.

First, we consider the Fisher-KPP equation. There are two parameters in this equation – r and D. However, we can divide throughout by D and rescale time in Eq. (4.4), so that the equation remains the same. That is equivalent to putting $D = 1$. If we ignore the diffusion terms in Eq. (4.4), $ru_n(1-u_n)$ acts like a local source term, which makes the concentration equal to one for any non-zero initial concentration. Since the diffusive term only diffuses but does not create anything new, it is natural that the concentration of all nodes would be unity after a very long time. Thus, any non-trivial initial condition for the Fisher-KPP equation on a finite line is immaterial for long-time dynamics. Thus, we can start with a concentration of 1 and look at its spread. Even though we have set $D = 1$, it is necessary that $r/D \sim 1$ or higher for the wave solution to hold. If the diffusive term dominates, *Time of Arrival* is not linearly correlated to 'distance from origin'. Thus, this

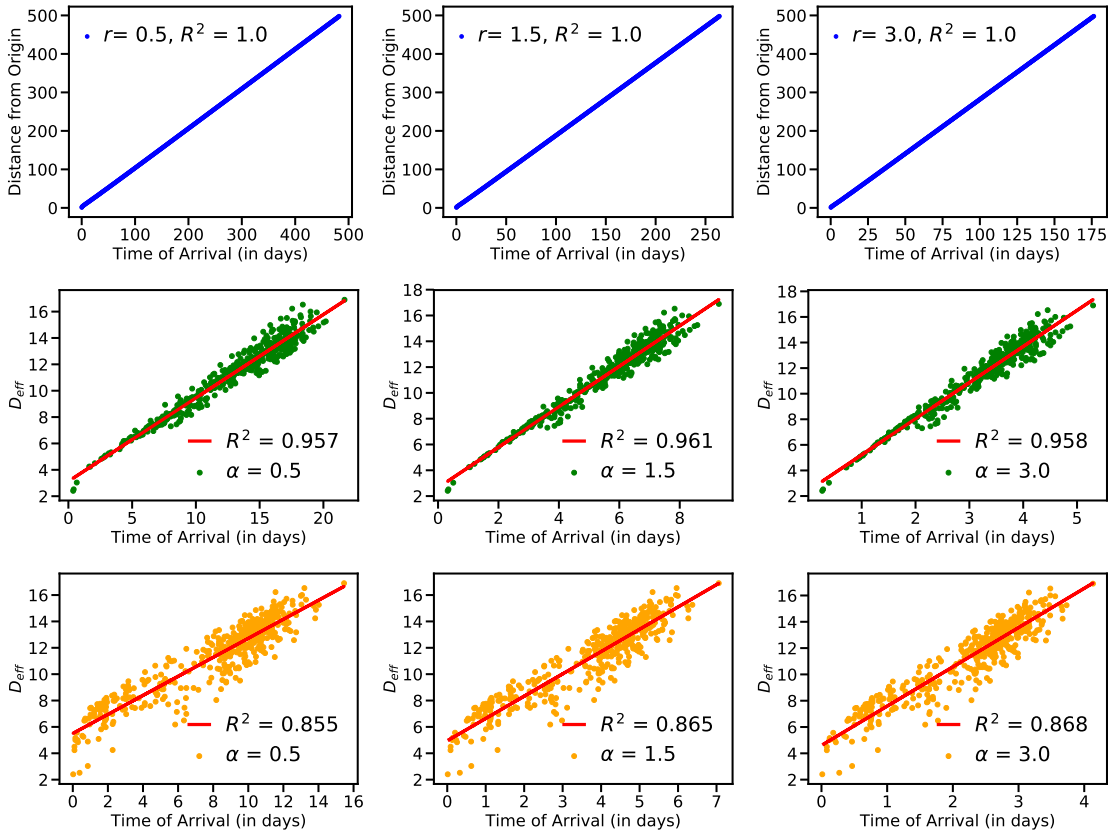


Figure 4.2: We plot the *Time of Arrival* against D_{eff} (for SI-model) and distance from origin (for Fisher-KPP equation). The first row shows Fisher-KPP equation scatter plot. The next two rows illustrate the plots for SI equation for absolute and fractional thresholds. The outbreak location for SI model is Tirupati. (Refer to text for more details)

relationship holds when r is of the order of or greater than D.

It is essential to mention here that the discussion about the evolution of concentration in the Fisher-KPP equation follows strictly for the infected cases in the SI model. In other words, the initial condition for the SI model is that the whole city is infected, which is certainly not equivalent to our discussion for the SIR model, but given that the steady-state conditions for both these models are different, it is reasonable to assign the above initial condition. We take a small city as the outbreak location for this exact reason. Fisher-KPP equation was in terms of fractional concentration, and hence, the corresponding way to compare with SI-model should be a fractional threshold. The results, however, do not seem to match as seen in the third row of Figure (4.2), where we have considered the fractional threshold. One alternative but a non-rigorous reason for this might be that there is no associated population with the vertices in the Fisher-KPP equation.

We can redefine D_{eff} ; however, we will skip that exercise for now and concentrate on the absolute threshold. If we compare the top two rows in Figure (4.2), we see that D_{eff} correlates very nicely with *Time of Arrival* . Thus, the velocity is well-defined for both these cases. In order to solidify the similarities between the two models, we look at the trend of V_{eff} for the respective source parameters, – α and r. Please note that this is not a very strict comparison, as we have arbitrarily set $D = 1$ in the Fisher-KPP equation, but we cannot do so for the SI model. So, to be more accurate, we are looking at the trend of V_{eff} for r/D.

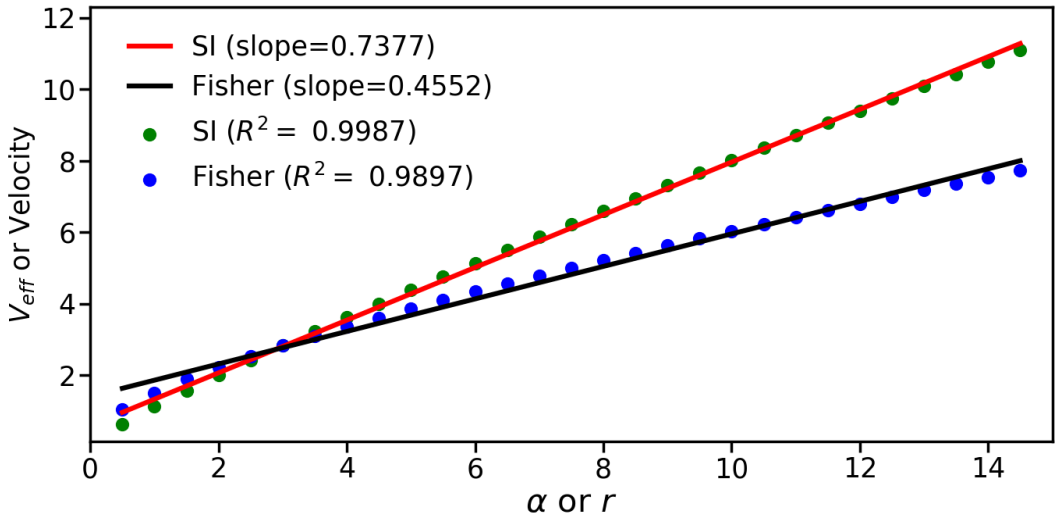


Figure 4.3: Comparison of velocity trend for SI-model and Fisher-KPP equation with respect to the infection/source rate (α or r). Tirupati is chosen as the outbreak location for the SI-model, while the fisher equation consists of 300 lattice points.

We see that V_{eff} for both systems increase linearly with α and r respectively. So, we have a partial agreement between r and α. Another reason might be that the slope for the SI model depends on the outbreak location. In that case, we can still see a match, but it is not universal relation then. The slope and the intercept are not equal or close by, thus eliminating the possibility of a direct connection between Fisher and SI equations. In the next Section, we will look into more microscopic properties of the spread pattern and check if there is any match with the existing models.

4.3 Diffusion on a Network

The SI equation presents a set of barriers towards connecting it with the Fisher equation. Firstly, the underlying natural topology of the network does not appear anywhere in the equations, unlike the Fisher equation, where the distance is defined using the distance between consecutive points. Secondly, D_{eff} is being used at two places – to define the jump rates between various nodes and define the space itself. This is unlike the Fisher equation (or any other diffusive process), where space is defined using natural topology, and the jump rates are usually defined independently. Thirdly, in the Fisher equation, the $\boldsymbol{P}$-matrix only connects the nearest neighbors, and no direct long-distance connections are present.

On the other hand, for the SI model, D_{eff} first flattens out a network into a 1D line for a specific outbreak location. However, there are still long-distance connections between various cities, which may drastically alter the system's evolution. Finally, the significant $\boldsymbol{P}$-matrix entries for any city are spread out over more than five cities, making it unreasonable to assume that a random walker on this network will only jump to the nearest neighbor in one-time step.

Owing to all the above reasons, we do not expect a direct microscopic match between the SI-model and Fisher-KPP equations properties. In order to further simplify the problem, we consider $\gamma_n = 1 \; \forall \, n$, so that the only difference between the two systems is the $\boldsymbol{P}$-matrix. We ignore *Time of Arrival* and D_{eff} and look at the properties of the infected fraction at each node in Figure (4.4).

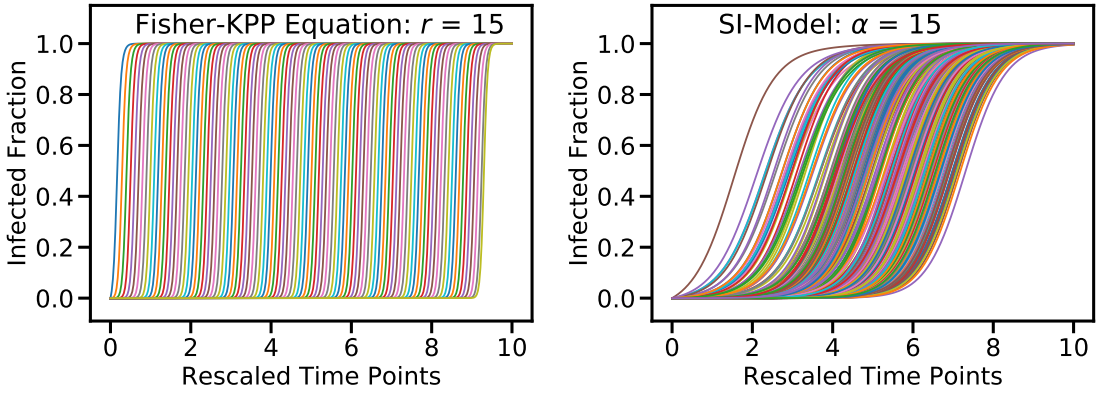


Figure 4.4: Evolution of infected fraction for 446 (100) cities for the SI-model (Fisher-KPP equation). $r = \alpha = 15$, and $D = 1$. The x-axis is rescaled so that we can compare the shapes properly. The figure on the left denotes Fisher-KPP equation, while the one on the right denotes SI-model. The outbreak city's time evolution is not plotted in this figure. Tirupati is the outbreak location for SI model.

Figure (4.4) shows the time evolution of infected fraction for both models. We consider the regime where r and α are high. We can see that the plots look similar in some sense. In order to further quantify this, we look at the evolution of mean and standard deviation of the infected fraction for all cities. Qualitatively, we can predict that $\mu(0) = 0$ and after long time $\mu(t) = 1$. Similarly, $\sigma = 0$ at the endpoints, and hence we expect a peak somewhere near the middle of the evolution. We plot both these quantities for both models in Figure (4.5).

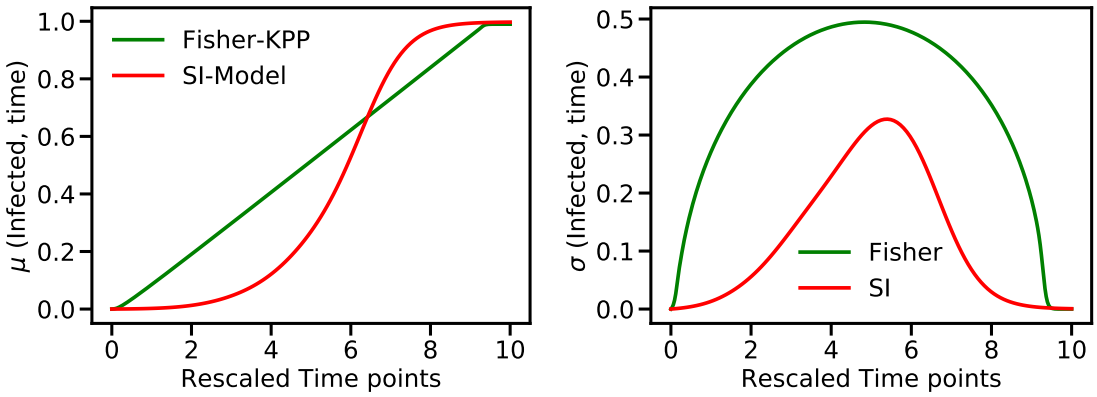


Figure 4.5: Mean and Standard deviation of the infected fraction across the cities as a function of time for both the models. All the parameters remain same as Figure (4.4).

Even though the infected fraction looks similar, the first two cumulants of the two models behave differently with time. We see that the mean increases linearly with time for the Fisher Equation. However, the same growth is non-linear for the SI model. Similarly, the standard deviation seems to follow a quadratic trajectory for the Fisher equation. However, for the SI model, it looks more like a Gaussian evolution.

As a final line of attack (for this thesis), we consider no source term in both equations such that the infection diffuses to all nodes, but the total concentration summed over all the nodes remains constant with time. Fisher-KPP equation reduces to ordinary diffusion equation on a line in the absence of the source term. The diffusion equation is a well-studied phenomenon whose properties are known with certainty [54]. Here, however, we compare diffusion on a line with diffusion on a network [55]. As before, we first look at the evolution of the infection on all nodes as a function of time.

Figure (4.6) shows the time evolution of the infection on various nodes as a function of time. We can see that the curves resemble some extent. Of course, the curve for diffusion on a line is very well structured, while the network diffusion curves have some heterogeneity. Even though the steady-state distribution seems different for the two models, it is primarily because of the different

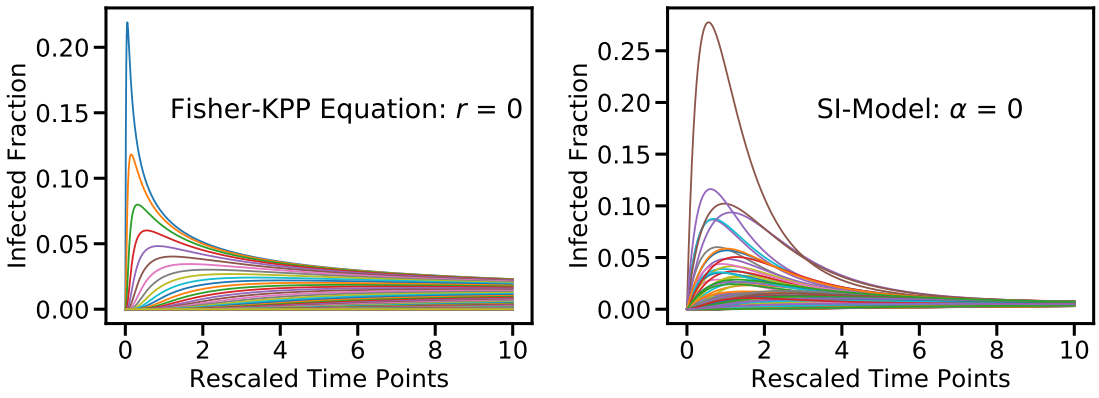


Figure 4.6: Diffusive behavior of the two models for the same parameters as Figure (4.4). Note that the diffusion on line is for 100 sites, while the network consists of 446 cities.

number of nodes in the two models. And finally, we compare the mean and standard deviation as before.

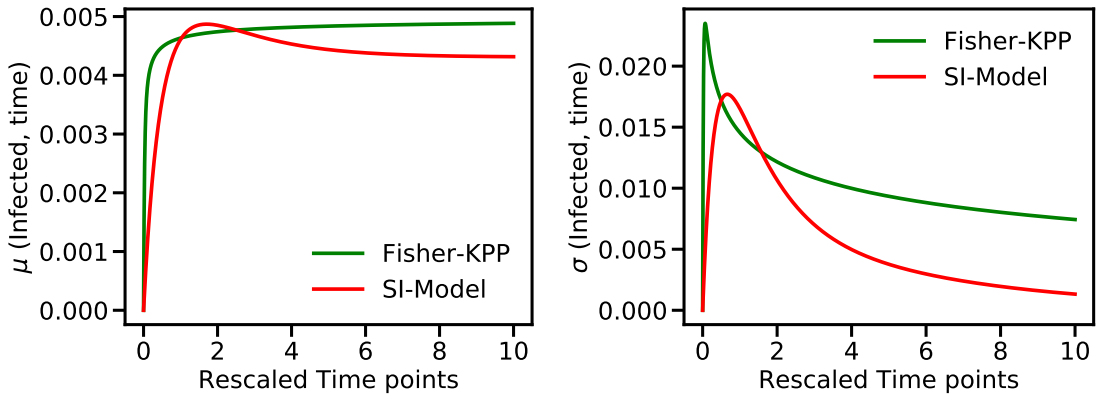


Figure 4.7: The first two cumulants of the infected fraction for the two models as a function of time. All the parameters remain same as Figure (4.6). The left panel denotes the mean, while the right panel denotes the standard deviation.

The mean (μ) and the standard deviation (σ) look much similar qualitatively for the two diffusion processes. When $r \neq 0$, the end values matched, but the evolution took very different routes for both the models. However, for the diffusion processes, *i.e.* $r = \alpha = 0$, the evolution looks very similar for both the models. The extremities depend on the static network properties, which are different for both models.

These results tell us that the diffusion on a line is similar to diffusion on a network, where the underlying distance is defined by the $\boldsymbol{P}$-matrix. Though we have found some evidence to show the

similarity between the two models, it is primarily numerical and specific to our network. We still do not understand the concept of D_{eff} very well, and the future work will mainly focus on this, as we will discuss in the next Section of the summary.

4.4 Summary

The main focus of this Chapter was to dig deeper into the concept of D_{eff} and the reasons behind its effectiveness from a more theoretical perspective. We started this Chapter by giving some non-rigorous but intuitive motivation for D_{eff} . Inspired by the fact that D_{eff} works better for higher SIR-model rate parameters, we took the extreme limit when there are no recoveries and displayed that the modified (also called as SI model) is similar in the structural form of the equation to the Fisher-KPP equation – a model which is known to have wave-like solutions. However, we saw that the macroscopic quantities such as V_{eff} do not agree very well with the two models – Fisher and SI.

Since there was no robust macroscopic match between the two models, we compared the two models at a more microscopic level by looking at their respective time evolutions of the infected fractions. Though there was some qualitative agreement, it was not enough to make definite statements. This approach forced us to go towards the final approach of comparing the pure diffusive processes on the two networks (line and the transport network). The match was the best in this case, as evident from the evolution of mean and standard deviation for the two processes.

We still do not have a complete understanding of the effectiveness of D_{eff} . We will look at a few possible ways of extending this study in the last Chapter to conclude this thesis.

Chapter 5

Conclusion and Future Directions

In an increasingly connected world, the risk of infectious diseases spreading has been prophesied, and the matching reality has been well-documented in the past few years. Even though much better computational models exist to predict the risk of infection spreading globally, a minuscule number of such models exist tailored for India. In this thesis, we studied the problem of understanding the spread of infectious diseases in India using some well-established epidemiology and network science tools.

After giving a motivation for studying the problem in Chapter (1), we looked into collecting and generating the traffic data methods using the available sources in Chapter (2). Our algorithms were based on a simple assumption that the number of people traveling from a city is proportional to the city population. Using this simple assumption, we built the dataset and simulated the SIR metapopulation model. One fundamental distinction of our dataset from the existing datasets was that it consisted of three modes of transport and considered geographically local and global traffic.

We introduced the concepts of *Time of Arrival* and D_{eff} in Chapter (3) and showed that they correlate linearly with each other. We then showed few main results concerning the robustness and effectiveness of our model. We also looked into more practical aspects of this model, which gave us some insights into the possible measures that could be taken to curb the spread of the disease. Finally, we also compared our simulation with the actual life data. Different infection rates in cities, incomplete information about the strain of virus propagating in different cities, the unreliable data of the infected cases, and the changes in the mobility pattern due to lockdown are a few out of multiple possible reasons for the mismatch between simulation and real-life data.

In Chapter (4), we looked into the reasons behind why D_{eff} works well in some settings while not so well in others. We looked into the limits of the SIR model, where we could consider it as an

SI model, which is very similar to the Fisher-KPP equation. We showed few interesting similarities between these two models for two cases – in the presence of the source terms and or the absence thereof. The main result was that D_{eff} selects the path which an ordinary diffusion on a line would undertake, with the line coordinates playing a part of D_{eff} .

There are multiple ways the current study can be extended. In this thesis, we considered a static traffic matrix. However, we know that there are traffic fluctuations for all periods ranging from a day to specific months. It would be interesting to run this model for dynamic traffic data (if available) to incorporate the fluctuations into the model. We can also increase the compartments of the model if the correct parameters and corresponding data are found. It would also be an exciting exercise to integrate fluctuations at a city level regarding infection and recovery rates. Finally, given that the traffic data exists, one can in-principle use it to study multiple transport-related problems on networks not restricted to epidemic spreading. The concept of D_{eff} also opens multiple avenues into studying diffusion on a network. We can explore the mechanism of why D_{eff} works so well by simulating the dynamics on synthetic networks.

We thus come to the end of this thesis. To summarize the whole thesis in a line, we can say that – 'we collected the static traffic data for Indian transportation network for three modes of travel and predicted the risk associated with Indian cities based on the concept of *effective distance*, before providing some evidence for this effectiveness using tools from statistical physics, epidemiology, network, and data Science.' Even though this thesis considers the simplest of models available, we hope that future studies build upon this and create better models.

Bibliography

[1] https://covid19.who.int/, https://coronavirus.jhu.edu/map.html

[2] https://www.covid19india.org/

[3] Coscia, M., Neffke, F.M., Hausmann, R., *Knowledge diffusion in the network of international business travel.* Nature Human Behaviour **4**, 10111020. (2020).

[4] https://www.who.int/ith/ITH_EN_2012_WEB_1.2.pdf, https://www.who.int/tdr/publications/documents/seb_topic3.pdf

[5] Belik, V., Geisel, T., Brockmann, D., *Natural human mobility patterns and spatial spread of infectious diseases.* Physical Review X **1**, 011001. (2011).

[6] Coltart, C.E., Behrens, R.H., *The new health threats of exotic and global travel.* British Journal of General Practice **62**, 512513. (2012).

[7] Helbing, D., *Globally networked risks and how to respond.* Nature **497**, 5159. (2013).

[8] Barbosa, H., Barthelemy, M., Ghoshal, G., James, C.R., Lenormand, M., Louail, T., Menezes, R., Ramasco, J.J., Simini, F., Tomasini, M., *Human mobility: Models and applications.* Physics Reports **734**, 174. (2018).

[9] Ronald, R., and Hilda, P., *An application of the theory of probabilities to the study of a priori pathometry.–Part III,* Proc. R. Soc. Lond. A **93**: 225240. (1917).

[10] Kermack, W. O. and McKendrick, A. G., *A contribution to the mathematical theory of epidemics,* Proc. R. Soc. Lond. A **115**:700721. (1927).

[11] Colizza, V., Pastor-Satorras, R., and Vespignani, *A. Reaction – diffusion processes and metapopulation models in heterogeneous networks.* Nature Phys. **3**, 276282 (2007).

[12] Gong, Y., Song, Y., Jiang, G., *Epidemic spreading in metapopulation networks with heterogeneous infection rates,* Physica A: Statistical Mechanics and its Applications, **416**, 208-218. (2014).

[13] Colizza, V., and Vespignani, A., *Epidemic modeling in metapopulation systems with heterogeneous coupling pattern: Theory and simulations*. Journal of Theoretical Biology, **251**, 3, 450-467, (2008).

[14] Citron, D. T.,Guerra, C. A., Dolgert, A. J., Wu, S. L. , Henry, J. M., Sánchez C., Héctor M., and Smith, D. L., *Comparing metapopulation dynamics of infectious diseases under different models of human movement,* Proc. of the Nat. Acad. of Sci., **118**, 18. (2021).

[15] Pastor-Satorras, R., and Vespignani, A., *Epidemic Spreading in Scale-Free Networks,* Phys. Rev. Lett. **86**, 3200. (2001).

[16] Gautreau, A., Barrat, A., Barthelemy, M., *Global disease spread: statistics and estimation of arrival times.* Journal of theoretical biology **251**, 509522. (2008).

[17] Iannelli, F., Koher, A., Brockmann, D., Hvel, P., Sokolov, I.M., *Effective distances for epidemics spreading on complex networks.* Physical Review E **95**, 012313. (2017).

[18] Taylor, D., Klimm, F., Harrington, H.A., Kramr, M., Mischaikow, K., Porter, M.A., Mucha, P.J., *Topological data analysis of contagion maps for examining spreading processes on networks.* Nature communications **6**, 111. (2015).

[19] Brockmann, D., Helbing, D., *The Hidden Geometry of Complex, Network-Driven Contagion Phenomena.* Science **342**, 13371342. (2013).

[20] Althouse, B.M., Wenger, E.A., Miller, J.C., Scarpino, S.V., Allard, A., Hbert-Dufresne, L., Hu, H., *Stochasticity and heterogeneity in the transmission dynamics of SARS-CoV-2.* `arXiv:2005.13689`. (2020).

[21] Arenas, A., Cota, W., Gmez-Gardees, J., Gmez, S., Granell, C., Matamalas, J.T., Soriano-Paos, D., Steinegger, B., *Modeling the Spatiotemporal Epidemic Spreading of COVID-19 and the Impact of Mobility and Social Distancing Interventions.* Physical Review X **10**, 041055. (2020).

[22] Feng, L., Zhao, Q., Zhou, C., *Epidemic in networked population with recurrent mobility pattern.* Chaos, Solitons & Fractals **139**, 110016. (2020).

[23] Garcia-Gasulla, D., Napagao, S.A., Li, I., Maruyama, H., Kanezashi, H., Perez-Arnal, R., Miyoshi, K., Ishii, E., Suzuki, K., Shiba, S., *Global Data Science Project for COVID-19 Summary Report.* `arXiv:2006.05573`. (2020).

[24] Bedi, P., Gole, P., Gupta, N., Jindal, V., *Projections for COVID-19 spread in India and its worst affected five states using the Modified SEIRD and LSTM models.* `arXiv:2009.06457`. (2020).

[25] Jha, V., *Forecasting the transmission of Covid-19 in India using a data driven SEIRD model.* `arXiv:2006.04464`. (2020).

[26] Khajanchi, S., Sarkar, K., *Forecasting the daily and cumulative number of cases for the COVID-19 pandemic in India.* Chaos: An Interdisciplinary Journal of Nonlinear Science **30**, 071101. (2020).

[27] https://censusindia.gov.in/2011-common/censusdata2011.html

[28] Khajanchi, S., Sarkar, K., Mondal, J., Perc, M., *Dynamics of the COVID-19 pandemic in India.* arXiv:2005.06286. (2020).

[29] Mishra, R., Gupta, H.P., Dutta, T., *Analysis, Modeling, and Representation of COVID-19 Spread: A Case Study on India.* arXiv:2008.13116. (2020).

[30] Sarkar, K., Khajanchi, S., Nieto, J.J., *Modeling and forecasting the COVID-19 pandemic in India.* Chaos, Solitons & Fractals **139**, 110049. (2020).

[31] Sen, P., Dasgupta, S., Chatterjee, A., Sreeram, P.A., Mukherjee, G., Manna, S.S., *Small-world properties of the Indian railway network.* Physical Review E **67**, 036106. (2003).

[32] Gopalakrishnan, R., Rangaraj, N., *Capacity management on long-distance passenger trains of Indian Railways.* Interfaces **40**, 291302. (2010).

[33] Rajput, N.K., Badola, P., Arora, H., Grover, B.A., *Complex Network Analysis of Indian Railway Zones.* arXiv:2004.04146. (2020).

[34] Gopal, R., Chandrasekar, V. K. , and Lakshmanan, M. *Dynamical modelling and analysis of COVID-19 in India.* Current Science, **120**, 8. (2021).

[35] Pujari, B.S., Shekatkar, S., *Multi-city modeling of epidemics using spatial networks: Application to 2019-nCov (COVID-19) coronavirus in India.* medRxiv. https://doi.org/10.1101/2020.03.13.20035386 (2020).

[36] Gupta, S., Shah, S., Chaturvedi, S., Thakkar, P., Solanki, P., Dibyachintan, S., Roy, S., Sushma, M.B., Godbole, A., Jaseem, N., *et al. An India-specific compartmental model for Covid-19: projections and intervention strategies by incorporating geographical, Infrastructural and response heterogeneity.* arXiv:2007.14392. (2020).

[37] Das, A., Dhar, A., Goyal, S., Kundu, A., Pandey, S., *COVID-19: Analytic results for a modified SEIR model and comparison of different intervention strategies.* Chaos, Solitons & Fractals 110595. (2021).

[38] Ray, D., Salvatore, M., Bhattacharyya, R., Wang, L., Du, J., Mohammed, S., Purkayastha, S., Halder, A., Rix, A., Barker, D., *Predictions, role of interventions and effects of a historic national lockdown in Indias response to the COVID-19 pandemic: data science call to arms.* Harvard data science review 2020. (2020).

[39] Sharma, A., Arya, S., Kumari, S., Chatterjee, A., *Effect of lockdown interventions to control the COVID-19 epidemic in India.* arXiv:2009.03168. (2020).

[40] Tiwari, V., Bisht, N., Deyal, N., *Mathematical modelling based study and prediction of COVID-19 epidemic dissemination under the impact of lockdown in India.* Frontiers in Physics **8**, 443. (2020).

[41] Venkateswaran, J., Damani, O., *Effectiveness of testing, tracing, social distancing and hygiene in tackling covid-19 in india: A system dynamics model.* arXiv:2004.08859. (2020).

[42] Grady, D., Thiemann, C., Brockmann, D., *Robust classification of salient links in complex networks.* Nature commun. **3**, 864. (2012).

[43] Ablowitz, M.J., Zeppetella, A., *Explicit solutions of Fishers equation for a special wave speed.* Bulletin of Mathematical Biology **41**, 835840. (1979).

[44] Trapman, P., *On analytical approaches to epidemics on networks.* Theoretical Population Biology, **71**, 2, 160 - 173. (2007).

[45] Driessche, P. *Reproduction numbers of infectious disease models.* Infect. Dis. Model. **2**(3):288-303. (2017).

[46] http://www.knowindia.net/aviation.html.

[47] https://www.civilaviation.gov.in/sites/default/files/MoCA_Annual_Report_2018_19.pdf, Page No. 49

[48] https://indianrailways.info/all_trains/, https://www.makemytrip.com/railways/list-of-indian-railway-stations.html

[49] https://enquiry.indianrail.gov.in/mntes/

[50] https://indianrailways.gov.in/railwayboard/uploads/directorate/stat_econ/IRSP_2016-17/Facts_Figure/Fact_Figures%20English%202016-17.pdf, Page no. 6.

[51] https://tis.nhai.gov.in/

[52] Hasegawa, T., Nemoto, K., *Outbreaks in susceptible-infected-removed epidemics with multiple seeds.* Physical Review E **93**, 032324. (2016).

[53] Toli, D., Kleineberg, K.-K., Antulov-Fantulin, N., *Simulating SIR processes on networks using weighted shortest paths.* Scientific reports **8**, 110. (2018).

[54] Kampen, N. G. V., *Stochastic Processes in Physics and Chemistry,* (Third Edition), Elsevier. (2007).

[55] Masuda, N., Porter, M., and Lambiotte, R. *Random walks and diffusion on networks,* Physics Reports, **716717**,1-58. (2017).

[56] For more information, visit: https://www.iiserpune.ac.in/~hazardmap/

[57] Sadekar O, Budamagunta M, Sreejith G.J., Jain S, and Santhanam M.S., *An infectious diseases hazard map for India based on mobility and transportation networks.* arXiv:2105.15123. (2021).