Title: Evaluating strategies for pandemic response in Delhi using realistic social networks

Authors: Huadong Xia  
Kalyani Nagaraj  
Jiangzhuo Chen  
Madhav V. Marathe

Contact: Madhav V. Marathe  
Email: mmarathe@vbi.vt.edu  
Tel.: +1 540 231 8832  
Fax: +1 540 231 2891

Acknowledgements: We thank our external collaborators and members of the Network Dynamics and Simulation Science Laboratory (NDSSL) for their suggestions and comments. This work has been partially supported by DTRA Grant HDTRA1-11-1-0016, DTRA CNIMS Contract HDTRA1-11-D-0016-0001, NIH MIDAS Grant 2U01GM070694-09, NSF PetaApps Grant OCI-0904844, and NSF NetSE Grant CNS-1011769.
Abstract

We analyze targeted layered containment strategies to contain an influenza pandemic in the National Capital Territory of India (NCT-I, including New Delhi and its surrounding areas). A key contribution of our work is to synthesize a realistic individual-based social contact network for NCT-I using a wide variety of open source and commercial data. New techniques were developed to infer daily activities for individuals using aggregate data published in transportation science, combined with human development surveys and targeted local surveys. The resulting social contact network is the first such network constructed for any urban region of India. The time varying spatially explicit network has over 13 million people and more than 200 million people-people contacts. The network has several interesting similarities and differences as compared to similar networks for US cities.

As a second step, we use a high performance computing based modeling environment to study how an influenza-like illness (ILI) would spread over the NCT-I network. We also analyze well understood pharmaceutical and non-pharmaceutical containment strategies to control a pandemic outbreak. Our methodology builds on earlier work in this area. The results suggest: (i) pharmaceutical containment strategies typically are more effective than non-pharmaceutical for NCT-I residents; (ii) the epidemic dynamics of the region are strongly influenced by activity pattern and demographic structure of the local residents; (iii) a high resolution social contact network helps us make a better public health policy. To the best of our knowledge this is the first such study in the Indian sub-continent.

1 Introduction

Today’s densely populated urban regions enable rapid transmission of infectious diseases [1]. Additionally, urban contact networks in regions like India and China are rapidly growing. The National Capital Territory region of Delhi predicts a rise in population from 16.7 million in 2011 to 22.5 million in 2021 primarily due to the high rate of in-migration in Delhi [2]. In Beijing, the population rose from 12.9 million in 2000 to 18.8 million in 2010 [3]. The densely populated large urban regions provide a perfect fabric for the rapid spread of infectious diseases. Public health authorities have focused on developing effective interventions and policies to control the spread of diseases using pharmaceutical interventions and social distancing measures. Both strategies effectively reduce the connectivity within the social contact network or change the transmission probability between individuals.

Over thepast 10 years, we have developed a formal methodology for network computational epidemiology – development and use of computer models to understand the spatio-temporal diffusion of disease through populations using a synthetic yet realistic representation of the underlying social contact network [4]. The basic approach has now become widely accepted in the literature. It is based on the idea that a better understanding of social contact network characteristics can allow more insight into disease dynamics and intervention strategies for epidemic planning.

A methodology to synthesize social contact networks is already in place for US cities. Contact networks for US cities are generated by following a hierarchical composition of data-driven stochastic processes: (i) the baseline population is synthesized based on sociodemographic statistics and microsample data from the United States Census; (ii) mobility patterns from a nationwide household survey and land use data in the form of work, retail, recreational, and school and college locations are used to estimate region-specific contact networks. The structure of the resulting social networks calibrated to the above data, has been shown to influence the outcome of disease outbreaks in our simulated epidemic models [5, 4].
Since the synthetic network should provide a realistic representation of the contact network specific to that region, the process to generate the contact network utilizes region-specific data. The US synthetic population captures details of household structure by utilizing the 5% Public Use Micro Sample for each Public Use Microsample Area modeled. The US National Household Travel Survey (NHTS) [6] captures the interdependence of people’s activities, especially adults, in the same household across all surveyed households in the United States. Data with a similar level of detail is not available for Delhi (and will not be for many other regions), making it impossible to replicate the US network generation process for regions outside the US.

1.1 Summary of contributions

Here, building on our earlier work, we construct a synthetic social contact network for NCT-I. To overcome data limitations for Delhi, we developed several new methods – many of which are general and can be applied to synthesize networks for urban regions in other developing countries. To the best of our knowledge, this is the first such synthetic network developed for any urban region in Southeast Asia. Using a variety of data sources, demographic information for each person and location, and a minute-by-minute schedule of each person’s activities and the locations where these activities take place is generated by a combination of simulation and data fusion techniques. This yields a dynamic social contact network represented by a (vertex and edge) labeled bipartite graph \( G_{PL} \), where \( P \) is the set of people and \( L \) is the set of locations. If a person \( p \in P \) visits a location \( \ell \in L \), there is an edge \( (p, \ell, \text{label}) \in E(G_{PL}) \) between them, where \( \text{label} \) is a record of the type of activity of the visit and its start and end points. The synthetic social contact network is: (i) spatially explicit – home locations, work locations, business locations, educational institutions, government institutions and other places of interest are explicitly represented; (ii) time varying – individuals carry out daily activities based on a normative day and visit appropriate location in turn interacting with other individuals visiting the locations during the same time and (iii) labeled – both individuals and locations carry a range of attributes described in the subsequent sections. It is impossible to build such a network by simply collecting field data; the use of generative models to build such networks is a unique feature of this work.

We then use high-performance agent-based simulations to study the spread of influenza-like illness over the synthetic social contact network of NCT-I. We study the efficacy of various intervention strategies. This includes pharmaceutical and non-pharmaceutical interventions. We rank order these strategies based on their efficacy and discuss how these results compare with results reported for other cities in the world.

2 Related Work

Traditionally, mathematical and computational modeling of epidemics has focused on aggregate models using coupled rate equations [7]. In this approach, a population is divided into subgroups (compartments) according to an individual’s health state (e.g., susceptible, exposed, infected, and recovered) and demographics. The evolution of the infectious disease is characterized by ordinary differential equations. An important assumption in all aggregate differential equation-based models is homogeneous mixing. This limits use of these models for spatially sensitive processes.

In recent years, high-resolution individual-based computational models have been developed to support planning, control and response to epidemics. These models support networked epidemiol-
ogy – the study of epidemic processes over explicit social contact networks. Research in this area can be divided into three distinct subareas.

The first subarea aims to develop analytical techniques and computer simulations over classes of progressively sophisticated random graphs [8, 9]. These models relax the mean field assumption to some extent, but still use the inherent symmetries in random graphs to analytically compute important epidemic quantities of interest. The primary goal of these results is to obtain closed form analytical results.

The second subarea aims to develop individual based models using important statistics of a region. The two important statistics used are: (i) density and is usually obtained using LandScan data and (ii) basic census information that provides the demographic distribution of individuals within a population. A simple template is used to represent a community and these communities are joined hierarchically to obtain larger regions. See [10, 11, 12, 13, 14] for examples of this approach. These models can extend to obtain hybrid models. In a hybrid model, counties are represented as nodes and edges are added between counties to capture the movement of individuals – see [15, 16, 17] for a comparative study. Epidemic dynamics within a county are computed using an individual-based model. The dynamics over the network of counties are captured using coupled rate equations.

The final class of models use the most realistic representation of social contact networks; see [18, 19, 20]. In [20, 19, 21, 22] each individual in the United States is modeled with detailed demographic profiles and daily activities. Our synthetic social network for NCT-I is constructed using this class of models.

3 Network Generation: data and methodology

To study the intervention strategies for pandemic response, it is important we create a faithful people-people contact representation for the region. In this section, we describe how to construct a realistic social contact network for the city of Delhi. During the procedure, both data and methodology play a key role.

3.1 Data Collection

Precisely, Delhi refers to the National Capital Territory of India (NCT-I). It is the capital of India, including New Delhi and several adjacent urban areas. Delhi contains over 13 millions people and is one of the areas with highest population density in the world. The average population age is fairly young with a higher male to female ratio. Some population statistics can be found in table 1 in comparison with two other representative cities in the world.

<table>
<thead>
<tr>
<th>City</th>
<th>Population size</th>
<th>Average age</th>
<th>Average household size</th>
<th>sex ratio (M/F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>16,191,340</td>
<td>37.9</td>
<td>2.6</td>
<td>0.99</td>
</tr>
<tr>
<td>Delhi</td>
<td>13,850,507</td>
<td>25.6</td>
<td>5.14</td>
<td>1.22</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>16,228,759</td>
<td>32.9</td>
<td>3.0</td>
<td>0.97</td>
</tr>
</tbody>
</table>

Table 1: Demographic statistics of Delhi in comparison with other two cities

In constructing a contact network, multiple sources of data are required including demographics, activity pattern and land use information about the region. The data we collected to construct the
Delhi network is listed in Table 2.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>India Census 2001 [23]</td>
</tr>
<tr>
<td>Household Microdata: India Human Development Survey 2005 by the University of Maryland and the National Council of Applied Economic Research [24]</td>
<td>Micro samples for household structure. It describes each household sample: hh size, hh heads age, hh income, house types, animal care; and also for each individual in the hh: demographic details, religion, work, marital status, relationship to head, etc.</td>
</tr>
<tr>
<td>Locations</td>
<td>MapMyIndia Dataset [25]</td>
</tr>
<tr>
<td>Activity</td>
<td>Thane Travel Survey by UFL [26]</td>
</tr>
<tr>
<td>India residential area activity survey by NDSSL</td>
<td>The survey focused on approximately 40% of adults in India who do not travel to work. People's age, gender, and contact duration near their home are collected. The survey is conducted by our group of NDSSL.</td>
</tr>
</tbody>
</table>

Table 2: The demographics, location and activity data used in the construction of the Delhi network.

3.2 Network Construction Methodology

Our method follows the steps we do when constructing networks of another area [27]: synthesize a baseline population with detailed individual structure and same aggregate statistic properties of the real population; assign each individual a reasonable activity schedule; and create locations in the region where synthetic people can take their activities. Our methods are similar to what was done in [27], but to accommodate region-specific data in Table 2, we also design some novel methodologies. In the following, section 3.2.2 and 3.2.4 are new methods we created and are described in more detail; other steps can be found with more details in paper [27].

3.2.1 Synthetic population generation with the India Census 2001 and micro household sample data

Our objective in creating a synthetic population is to create all the individuals with disaggregate demographic features that fit aggregate distributions of demographic variables as a whole, while building a realistic household structure for all those individuals. For this purpose, the summary
statistics of interested demographic variables at a household level (India Census 2001 [23]) and a collection of household samples from a survey of Delhi [24] are required.

Assuming the surveyed household samples are representative, any household in the real population can be estimated with a carefully selected sample in terms of its household size and household members. We are able to replicate the samples to create all Delhi households. The family members in those synthetic households naturally compose the Delhi population.

During the procedure, sample selection is critical. From the census data we collected the distributions of available demographic variables on a household level, i.e., householder’s age and household size. We then select and replicate the samples based on the joint distributions of those variables. We choose these two variables because they characterize important household structure features. Many other variables regarding household structure in the micro sample data are related to these two variables to some extent. For example, the number of children in a family is correlated to the householder’s age and household size; the number of workers in a family is correlated with the household size and to some extent reflects the household income.

While the marginal distribution of the two variables (householder’s age and household size) are presented in census data, their joint probability is unknown. To estimate a reasonable joint probability, we apply the iterative proportional fitting (IPF) procedure to fulfill the task. IPF is an iterative algorithm for estimating cell probabilities in a contingency table such that the marginal totals remain fixed, the details can be found in [27]. To calculate the cell probabilities, we will assign an initial value for each marginal variable combination cell, and then iteratively fit the cell values through the IPF procedure until we get converged results. Since the selection of an initial cell value might affect the accuracy of the final results, we put in the estimation value from the micro household sample data before the first iteration.

3.2.2 Activity assignment using the 2001 Thane, India household travel survey statistics

Due to the unavailability of travel survey data in India at the time of this study, we devise a discrete-time simulation to assign detailed activity schedules to the synthetic population in DelhiNet-V2 using the 2001 Thane, India household travel survey statistics described by Nehra [26] and Banerjee [28].

The 2001 Thane household travel survey is a trip-based survey that collected travel data in the form of 24-hour trip diaries from 14,428 respondents from 3,505 households in the metropolitan region of Thane, a city in the western state of Maharashtra, India. Additionally, the survey collected sociodemographic information from respondents and their respective households. Literature on the Thane travel survey describes the sociodemographic profile of mobile adults (adults that recorded at least one trip) and people recording no trips, as well as travel data statistics in the form of empirical frequency distributions of trip start times and trip durations of the survey sample population. Statistics of personal and household trip rates are split by mode of transportation, household size and individual worker status. The literature also briefly describes trip frequency, activity characteristics, and time use characteristics of students younger than 16 years old, students older than 15 years old, and mobile adults. Detailed trip chaining analysis is also reported for commuters (adults reporting at least one work-based trip). All trips reported in the survey began at home and ended at home. Based on the Thane survey statistics reported in [26] and [28], the activity assignment process stated in Algorithm 1 generates a sequence of activities along with their start and end times for a normative 24-hour day for each synthetic person in the population. Each set of
activity assignments for a synthetic person are independent of the activity assignments to all other people in the synthetic population.

For each person in the baseline population, the algorithm first assigns an activity class to the synthetic individual depending on his/her demographics (age and gender, et al.). For adults, this is achieved by sampling from the commuter status and demographic distribution of adults in the survey population reported in [28]. The algorithm classifies synthetic adults as commuters (adults reporting at least one work related trip), noncommuters (mobile adults with no work related trips), or zero trip makers. We further assume all adult noncommuters below the age of 23 have school related activities and classify them as college attendees. Since the literature reports commuter status statistics only for adults, we make the following assumptions about individuals aged 17 years or less, henceforth referred to as kids. A kid under the age of 6 years is assigned the same sequence of activities as an adult from the same household having no work related or school related activities. Kids 6 to 10 years old are classified as primary school attendees, non school goers making at least one trip in a day, or zero trip makers. Similarly, kids 11 to 17 years old are classified as secondary school attendees, non school goers that make at least one trip in a day, or zero trip makers. The distribution of primary and secondary school attendees, non school going kids and kids with no trips in the synthetic population is set to match the net enrollment ratios of primary and secondary schools all over India from 2000 to 2007 [29] and the fraction of zero trip makers in the age range 6 through 17 years in the Thane sample. The activity class assignment process for both kids and adults is represented by function $f_1$ in step 1 of the algorithm.

In step 2, the activity class of the synthetic individual is then used to decide his/her activity sequence by sampling from an empirical frequency distribution of reported activity sequences in the Thane survey. The Thane survey describes each recorded trip by the origin and destination of the trip, namely, home, work, shop, school (or college), social/recreational, and all other location categories. These six location types along with ‘travel’ define the seven distinct activities that constitute an activity sequence. Individuals classified as zero trip makers are assigned a home activity for all 24 hours of the day. More than 99% of the students in the Thane survey report exactly two trips in a day [28]: home to school and school to home. As a result, we assign the activity sequence home – travel – school (or college, in the case of adults) – travel – home to all school or college attendees. The algorithm defines all non working adults and non school going individuals reporting at least one trip during the day and with no school or work related activities as noncommuters. Close to all noncommuting adults report exactly two trips in a day [28], of which approximately half reported the activity sequence: home – travel – shop – travel – home, a quarter reported the activity sequence: home – travel – social/recreational – travel – home, and the remainder reported the sequence: home – travel – other – travel – home. Since the literature provides no information on noncommuter kids in the survey, we assume that the above frequency distribution of activity sequences of noncommuting adults applies to noncommuter kids as well. Commuters report eight distinct activity sequences, of which 97.34% report only two trips in a day: home to work and work to home. The activity sequence assignment process for both kids and adults is represented by function $f_2$ in step 2 of the algorithm.

Finally, in step 3 of the algorithm, a detailed activity schedule with start and end times for each activity in the sequence is generated by sampling from reported empirical frequency distributions of trip start times and trip durations. For each activity in the activity sequence, the algorithm samples from the relevant trip start time and trip duration empirical distributions (represented by functions $g$ and $h$, respectively, in the algorithm) by conditioning on the time left till the day ends.
Since the literature does not report start time and the trip duration distributions for school or college related trips, we assign a fixed schedule to all primary school, secondary school and college attending individuals.

**Algorithm 1: Assign Activities**

**Input**: baseline synthetic population file with age and gender of each synthetic individual, input random seed $\xi$

**Output**: activity file with start and end times of each activity for each person in the synthetic population

**for each synthetic individual i do**

1. $[\xi, \text{actCLASS}_i] = f_1(\text{age}_i, \text{gender}_i, \xi)$; /* assign activity class */
2. $[\xi, \text{actSEQ}_i] = f_2(\text{actCLASS}_i, \xi)$; /* assign activity sequence */
3. for each activity $j$ in $\text{actSEQ}_i$ do

   /* generate detailed schedule */

   $[\xi, \text{startTime}_{i,j}] = g(\text{actSEQ}_i, \text{activity}_j, \text{endTime}_{i,j-1}, \xi)$

   $[\xi, \text{endTime}_{i,j}] = h(\text{actSEQ}_i, \text{activity}_j, \text{StartTime}_{i,j}, \xi)$

**Output**: base-synthetic-population-file-with-activity-schedule

3.2.3 Location creation, assignment and contact network estimation

Locations are where people conduct their activities. They decide how people are distributed in the geographical space of the city. The dataset of MapMyIndia [25] contains the land use statistics in the NCT of Delhi, including the coordinates for multiple types of real locations where people work, study, shop and have entertainment, respectively. We extracted those coordinates and assigned people to those locations for their daytime activities. Here, schools, colleges, shopping centers and other places are also work places. For example, schools are places students take classes, but they are also work places for teachers.

Home locations are another type of location for people’s home activities, which usually occur at night. We don’t have a complete data set for people’s real home coordinates. However, the city of Delhi is divided into 114 wards and we know the number of households in each ward, which helps us precisely distribute home addresses over the whole city.

Once we know for each location the subgroup of people who visit it, a people-location bipartite graph $G_{PL}$ is then yielded. Here $P$ is the set of people and $L$ is the set of locations. If a person $p \in P$ visits a location $\ell \in L$, there is an edge $(p, \ell, \text{label}) \in E(G_{PL})$ between them, where label is a record of the type of activity of the visit and its start and end points along the time axis.

With the presence of the labeled $G_{PL}$, we are able to further model people-people interactions and create a people-people contact network $G_P$ comprising them. Potential people-people contacts occur when two persons coexist in the same location at the same time. If a location is large, however, two persons won’t meet even they are there simultaneously. Therefore, we measure people-people interactions within a location via its sublocation structure. Sublocation is the division of people in a location wherein all people in the same sublocation are in contact with each other. Sublocation size is considered the largest possible sublocation within a given area. Apparently, the sublocation size is an important parameter characterizing the interactions of people within a location. We will discuss this further in section 5.
3.2.4 Contacts in residential area

The above methodology has been applied to generate several other cities in the world [27, 30]. However, as an unusual social-economic phenomenon in Delhi, about 40% of the population do not have a formal job and they stay around their residential area for the whole day. The data is verified from two independent sources, a nation-wide household survey conducted for India [24], and the travel survey we retrieved from [26]. Reference [26] claims 40% of people do not travel, excluding 32% of commuters, 12% of Non-commuters, and 16% of school kids.

Therefore, it is nontrivial to model the interactions among those people who stay home. We conducted a survey in Delhi and several other cities nearby, collecting data regarding those “at-home people” within a residential area. Since those people claim they do not travel, we assume they are in contact only with those people within their own community. Those contacts are generated randomly following certain patterns retrieved from the survey. Those new generated contacts form a contact network we call the residential network. We then incorporate the residential network into the Delhi network.

4 Analysis to the contact structure and optimal public health policy in Delhi

As introduced in the last section, we generate the Delhi network (including $G_{PL}$ and $G_P$) for the city of Delhi based on the high resolution data and novel methodologies. A high resolution social contact network reserves an effective contact structure in the population. It will provide insights for policy makers when studying the epidemic dynamics and evaluating effective public health policies. In the following, we conduct detailed analysis to the synthetic Delhi population and the Delhi network. In using such a network and powerful epidemic simulation platform EpiFast [22], we study various intervention strategies to contain the spread of disease in the city.

4.1 Demographics and Daily Activity pattern of the synthetic population

The individuals in the synthetic population are synthesized by aggregating members from those representative household samples based on the distributions of household level demographics. This is different from the coarse synthetic population, where individuals are built directly based on individual level demographic variables. We take the new method as an improvement because it incorporates more details and provides a realistic household structure. However, we hope the synthetic population is statistically similar to that of a real population in terms of individual level demographics. To verify, we plot Figure 2, comparing the synthetic population in 16 different age groups to the census data on an individual level. The observation with a Q-Q plot visualization in Figure 2a shows that the synthetic population in each age group is very close to the real statistics. The sex ratio of the synthetic population deviates a little bit for adults (refer to the deviation in counts of unisex groups in Figure 2b). Given that the number of micro samples are small, the deviation is not very large. The statistic deviation will diminish as we collect more representative household samples.

Figure 3 compares the statistics of the synthetic activities for all the people. We calculate for each hour in a typical day the number of people taking a specific type of activity, at-home, work, school, etc. The aggregated activities have a bias towards “at-home”. For anytime, there are more people staying at home than going out for other activities. This is due to the special economic
phenomenon in India where about 40% people do not have a job, as discussed earlier. Most people work or study during the day, and almost all people stays at home late night. If the survey basis is accurate, this unique cultural feature is quite different from an US city example.

Figure 1: The left side depicts the age-group counts for Delhi from the India Census 2001 [23]. The right side depicts the age-group counts of the synthetic population based on micro sample data and household level statistics. Visually we see the synthetic population conforms to the real statistics quite well. QQ-plots in Figure 2 give us a clearer visualization on this.

Figure 2: Q-Q plots of the age-group quantiles for the synthetic Delhi population. In Figure 2a, the age-group quantiles of the synthetic population conforms very well with the expected value based on the census data. If we count for unisex only, Figure 2b plots the comparison for male age-group counts, it deviates a little more but is still acceptable.

4.2 Graph Structural Properties of the Contact Networks

A detailed profiling of the network structure for $G_{PL}$ and $G_P$ is shown in Figure 4. In the bipartite people-location graph $G_{PL}$, there are 13.85 million people and 1.11 million locations. Its degree distribution is plotted in log-log scale in Figure 4a. A large part of the degree sequence follows a power law distribution, which conforms to findings in many other work [4]. For $G_P$, we plot the
distribution of node degrees, clustering coefficients and contact durations. Obviously, the degree
distribution in $G_P$ is totally different from that of $G_{PL}$. To get a better understanding of the
epidemic implication of those measurements, we compare these structural properties of the Delhi
network against those of the Los Angeles network in Table 3. Different from theoretic assumptions
such as power law degree distribution, the degree distribution of the Delhi network $G_P$ is peaked
around a degree of 20. The average degree is about 30; this is a relatively small number based on our
other study to USA cities [31]. Similarly, compared to the Los Angeles network, the average edge
weight (representing accumulated contact duration) is longer, and the average clustering coefficient
is significantly higher. Those structural features suggest residents in Delhi tend to stay with a few
fixed acquaintances for a long time instead of meeting many unfamiliar people for a short time.
Such a contact structure leads to implications regarding the pandemic spreading in the population.

<table>
<thead>
<tr>
<th>people-person network</th>
<th>No. of nodes</th>
<th>number of edges</th>
<th>Avg. degree</th>
<th>Avg. edge weight (minute)</th>
<th>Avg. CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>the Delhi network</td>
<td>206,787,386</td>
<td>13,850,507</td>
<td>29.86</td>
<td>363</td>
<td>0.546</td>
</tr>
<tr>
<td>the Los Angeles network [31]</td>
<td>459,273,880</td>
<td>16,228,759</td>
<td>56.60</td>
<td>141</td>
<td>0.389</td>
</tr>
</tbody>
</table>

Table 3: Average structure properties of several city-scale contact networks

Another interesting observation can be found when dividing contacts in the network into differ-
ent activity types. By plotting the duration distribution of different activity contacts, we find that
contacts from different activities may show totally different patterns. As shown in Figure 5, contact
duration in work places has a wide range and on average are much longer than those of shop and
other activities. However, we should be very careful with the implications of such a difference. It
does not necessarily suggest work contacts are more important than contacts in a shopping area,
we should at least look at the number of contacts in a typical location of those activities before
jumping to conclusions.

4.3 Epidemic Dynamics and Intervention Policies

Now we run epidemic simulations on the Delhi contact network to study epidemic dynamics and
the effects of public health policies in the Delhi population. We assume the disease simulated is
(a) Degree distribution of locations and people in the bipartite graph $G_{PL}$. The location degrees range from 1 to 8230 and show a power law like distribution.

(b) Degree distribution of people-people contacts in the network $G_P$. The average degree is 29.86.

(c) The clustering coefficients (CC) in $G_P$, its average CC is 0.546.

(d) Duration distribution of people-people contacts in the network $G_P$. The average contact duration is about 363 minutes.

Figure 4: Network structure profiling for the Delhi Network ($G_{PL}$ and $G_P$). Comparing $G_P$ to several other city-scale networks (Table 3), we can get a better sense of the relation between the structure and the epidemic dynamics.
H1N1, which occurred in a 2009 global outbreak and is still prevalent in India [32]. To address the variations in different estimates of $R_0$ of H1N1 in literatures [33, 34, 35], we choose a set of values: 1.35, 1.40, 1.45, and 1.60. We believe the range of these values covers most estimates for $R_0$ of H1N1 found in the literature.

### 4.3.1 Analysis of node vulnerability

Node vulnerability is measured as the probability a node is infected during an epidemic. We estimate it based on results of 10,000 random simulation runs. The distribution of node vulnerability when $R_0 = 1.35$ is shown in Figure 6. The distributions for other $R_0$ are very similar to that of $R_0 = 1.35$ (omitted here to save space, refer to appendix E for complete results), indicating the node vulnerability is more relevant to the network structure than to the disease property.

Figure 6: The vulnerability histogram for nodes in the Delhi network. This is based on 10,000 random epidemic simulations with $R_0 = 1.35$. In the figure, the fraction of low vulnerability nodes are more than the fraction of high vulnerability, where about 40% of people have a vulnerability lower than 0.35.
The vulnerability distribution of the people varies from 0 to 1, biased toward the left side. Quite a few people have a vulnerability close to 0. Compared to other populations (refer to section 6), such a distribution suggests a contact network resistant to disease spreading. And we believe it is highly related to the fact that a large portion of people do not travel a great deal in the city as shown in Figure 3.

4.3.2 Optimal intervention strategies during epidemic spreading

Using a high resolution contact network modeled for Delhi, we are able to achieve a better understanding for the epidemics and effectiveness of different intervention policies. We simulate four public health policies frequently applied in the real world, including pharmaceutical interventions (PI) and non-pharmaceutical interventions (NPI). PI includes Antiviral and Vaccination; NPI includes School Closure and Work Closure. The simulation results when $R_0 = 1.35$ are presented in Figure 7. The results when $R_0$ is 1.40, 1.45 and 1.60 are omitted because they are all very similar to what we show when $R_0 = 1.35$. The complete results can be found in appendix D.

Figure 7: Epidemics under various intervention strategies in the Delhi network when $R_0 = 1.35$, including a base case where no intervention is conducted. Here we use the tuple (attack-rate, peak, peak-day) to characterize epidemic dynamics. For Vaccination and Antiviral, we randomly choose 25% of the population to apply corresponding pharmaceutical treatments. School Closure and Work Closure are applied for 3 weeks when 0.1% of the nodes in the city get infected.

Vaccination has the strongest effect in containing the disease spread. All the other policies, including Antiviral, School Closure, and Work Closure, have lower effectiveness. Vaccination is significantly better than other policies and seems the best choice without considering other factors. Vaccines are not always available, however, especially at the early stage of an emerging disease epidemic. This was the case for the 2009 H1N1 pandemic. Even if vaccines are available, they may not be sufficient to provide mass vaccination. It is meaningful to consider the other three intervention policies.

School Closure and Antiviral have their pros and cons. Antiviral will help reduce the attack rate more than a School Closure, but a School Closure works better in reducing the maximum number of cases on any day (peak), and in delaying the occurrence of the peak. School Closure, however, is better in all three parameters (attack rate, peak population, and peak day) than Work Closure. By dissecting into the subpopulation structure and comparing their epidemic dynamics, we could gain insights on controlling the disease spreading. In Figure 8 we plot the epidemic curve,
which is the fraction of people infected on each day, for each of the four subpopulations (preschool, school age, adult, and senior). As observed from Figure 8, among all subpopulations, only school age has an epidemic worse than the population average (green curve in figure). Closing schools can avoid disease transmissions between students within schools, which explains the high effectiveness of School Closure.

Figure 8: Epidemic curves show subpopulation epidemics in the Delhi network when $R_0 = 1.35$. The Delhi population is partitioned to four groups based on age: preschool, school age, adult, and senior. Each red curve shows the fraction of people in that subpopulation infected on each day in the base case in Figure 7, where there is no intervention. The green curve shows the fraction of people in the whole Delhi population infected on each day in the same base case.

4.3.3 Targeted-Layered Containment (TLC)

In the last section, we simulate different intervention strategies individually to test their efficacy. In real life, however, a set of interventions are typically conducted simultaneously. Therefore we consider combinations of several interventions, including targeted interventions and general interventions. Such a combination of various interventions is called targeted-layered containment (TLC) [36].

We consider the TLC policy with a combination of the following interventions, and examine multiple levels of compliance with the interventions and infection rate thresholds for initiating interventions.

- **Targeted-Anti-Viral**: diagnosed individuals are applied anti-viral therapy (under some compliance).
- **Targeted-Stay-Home**: diagnosed individuals are suggested to stay-at-home (under some compliance).
- **School-Closure**: students’ school activities are removed (under some compliance and initiated with specific infection-rate threshold).
- **Work-Closure**: work places are closed (under some compliance and initiated with specific infection-rate threshold).

For the disease, we assume asymptotic rate is 20%, ascertain rate is 75%, therefore, the probability an infected individual is diagnosed is 60%. For each intervention in the TLC, we change
the compliance levels to 30%, 60% and 90%; and change the initiating threshold to be 0.01% and 0.1%. The results are shown in table 4; a more straightforward visualization is plotted in Figure 10.

Figure 9: Epidemics under various intervention strategies in the Delhi network when $R_0 = 1.35$, including a base case where no intervention is conducted. TLC is the combination of all other interventions, therefore obviously and reasonably, the most effective among all interventions.

It seems the time when the TLC is initiated is more important than people’s compliance to the interventions. Compared to the baseline case where no intervention is conducted, a TLC initiated when infected rate is 0.1% brings significantly lower attack rate while a TLC initiated when infected rate is 0.01% does not reduce the attack rate much. The reason is likely due to the short-duration of the TLC. All intervention lasts one to three weeks only, so an early TLC cannot completely suppress a disease that will eventually outbreak later on.

For the same initiating threshold value, the compliance value impacts the outbreak date significantly. A higher compliance value will lead to a much later outbreak. On the other hand, however, a higher compliance value seems to not be able to greatly reduce the attack rate and the peak value.

<table>
<thead>
<tr>
<th>compliance (%)</th>
<th>threshold(%)</th>
<th>average attack-rate(fraction)</th>
<th>average peak (fraction)</th>
<th>average peak-day</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA (baseline)</td>
<td>NA</td>
<td>0.434817</td>
<td>0.00924687</td>
<td>117.6</td>
</tr>
<tr>
<td>30</td>
<td>0.01</td>
<td>0.432875</td>
<td>0.00901022</td>
<td>136.976</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.416667</td>
<td>0.00714037</td>
<td>140.633</td>
</tr>
<tr>
<td>60</td>
<td>0.01</td>
<td>0.433024</td>
<td>0.00902515</td>
<td>156.933</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.417798</td>
<td>0.00713408</td>
<td>164.2</td>
</tr>
<tr>
<td>90</td>
<td>0.01</td>
<td>0.433257</td>
<td>0.00904592</td>
<td>168.03</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.419244</td>
<td>0.00728672</td>
<td>176.8</td>
</tr>
</tbody>
</table>

Table 4: results of Targeted-Layered Containment in Delhi when $R_0$ is 1.35.

It’s worth to note that, the effect of different initiating threshold values and compliance values shown above is not an ad-hoc phenomenon. We apply the exact same TLC strategies to the Los Angeles network we generated in another paper [31] and list the simulation results in table 5. Despite disparate demographic and contact structure of the two cities (table 3 and reference [31], the TLC effect is influenced by the threshold and the compliance in a very similar way.
Table 5: results of Targeted-Layered Containment in Los Angeles when $R_0$ is 1.35.

### Sensitivity Test to Our Synthetic Network Model of the Delhi network

Detailed and comprehensive data of a region is critical in constructing a high resolution network. However, not all data is available for us to prepare the Delhi network. We have to make assumptions in our model when necessary data is not retrieved yet or unlikely to be available. Two important assumptions in our model, made based on an educated guess, are the sublocation size and the location assignment algorithm. As introduced, people within a location are divided into connected subgroups in a network view. Let *sublocation size* be the largest subgroup size within a location; it reflects the internal structure of a location. We choose for each type of locations an empirical value for their sublocation size. Also in our model we apply a specific algorithm for assigning activity locations. For example, we assume that a person’s selection of his/her workplace or shopping center is based on the gravity model, i.e., further a place is from his home, less likely the place will be chosen. Therefore, for each activity of each person, we assign a distribution over all locations, with those close to the person’s home having larger probabilities, and we randomly choose one from the distribution to be the location of that activity.

Divergence between our synthetic network and the real network could occur due to such assumptions. To evaluate the influence of such choice to the quality of the constructed network, we conduct sensitivity tests to measure the divergence in terms of epidemic output.

We choose the same experimental settings as those in Section 4. Here we assume $R_0 = 1.35$. We point out, however, that the observations are similar for the sensitivity experiments with the $R_0$ value being 1.40, 1.45, or 1.60.

#### 5.1 Sensitivity to Sublocation Size

The sensitivity test results regarding various sublocation size are shown in Figure 11. The effect of varying sublocation size has a significant impact to either disease spreading or the intervention to the spreading. Second, changing the sublocation size of some specific types of locations may change the topological structure of the network, which may eventually change the effectiveness of intervention policies. For example, in the baseline network, School Closure is more effective in delaying disease spreading compared to Work Closure. For the network constructed after we increase sublocation size of work places (w+10 in figure), however, the effect of closing work places is as significant as that of closing schools. This means that the change of sublocation size has a fundamental impact to the structure of the synthetic representation of the real contact network, which produce non-negligible impact to the control of disease spreading in the population. Therefore, choosing the
right sublocation size is essential in our network modeling.

Figure 11: Epidemics and policy efficacy in the Delhi network with various sublocation sizes (R0=1.35). Six types of locations are modeled in the Delhi network: home(h), work-place(w), school(s), college(c), shops(sh) and other(o). In base case, we define the sublocation sizes for those locations based on empirical data. We increase the sublocation sizes for some locations in control groups. For example, (s+10, c+10) represents increasing sublocation sizes of schools and colleges by 10 and keeping sublocation sizes for the other types of locations; (all+10) means increasing sublocation sizes for all location types (except home) by 10.

5.2 Sensitivity to Location Switches

To test the second assumption, we switch locations for two randomly chosen people with the same type of activities. By verifying the robustness of the results under location switching, we can understand what the epidemic dynamics could be for another possible location assignment algorithm. From the simulation results, shown in Figure 12, we can hardly tell the difference between all those location switching operations. Obviously, the location assignment algorithm doesn’t change the effective contact structure under the context of our model. We conclude in terms of epidemics, people’s interaction pattern in a local place is more important than the location distribution in the city globally, given that the population density is unchanged.

6 Comparison study between the coarse network and the refined network

In our previous paper [31], we generated a network for Delhi based on very limited data with a generic methodology. In this paper, we construct a network with more detailed information and new methodology. We call the former “the coarse network” and the latter “the refined network” in this section. The data sets used in generating the two Delhi networks are listed in Table 6. The coarse network does not have micro household samples so the accuracy of the household structure of its synthetic population is not guaranteed. It does not have the exact location information of Delhi so LandScan data is used, which contains only the population density distribution over the region. More importantly, no activity survey data was available for India when we created the coarse network, so the US NHTS results were used as a substitution. In summary, the refined network is much more realistic in the sense of data input. By intuition, more detailed data will make the constructed network more realistic, and allow more precise analysis and prediction for
the epidemics in the Delhi population. To this end, we conduct a brief comparison study regarding the contact structure and epidemic dynamics of Delhi based on the two networks.

<table>
<thead>
<tr>
<th>Demographics</th>
<th>the coarse network</th>
<th>the refined network</th>
</tr>
</thead>
<tbody>
<tr>
<td>School/College Statistics [38, 39]</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity</td>
<td>based on US Travel Survey [6]</td>
<td>2001 Thane Household Travel Survey [26]; India residential area activity survey by NDSSL</td>
</tr>
</tbody>
</table>

Table 6: Data used in construction of the coarse network and the refined network.

The network structure profiling for the two networks are listed in Table 7. On average, the refined network, the coarse network has a much higher degree (76.99 v.s. 29.86) and lower edge weight (contact duration). We deduce that such difference makes diseases spread easier in the coarse network\(^1\). Similarly, regarding the clustering coefficient distribution, the refined network has a higher average clustering coefficient, also helping hinder disease spreading. The key source of the differences is probably because the coarse network borrowed the US mobility survey data. We see the coarse network looks more similar to the Los Angeles network than to the refined network, despite the two Delhi networks are built for the same population.

\(^1\)Let’s consider two simplified cases. Case 1, a seed node \(u\) has two contacts with durations \(d_1\) and \(d_2\). Case 2, a seed node \(u\) has one contact with duration \((d_1 + d_2)\). The expected number of secondary infections in case 1 is \((1 - (1 - \tau)^{d_1}) + (1 - (1 - \tau)^{d_2})\); that in case 2 is \(1 - (1 - \tau)^{d_1+d_2}\), where \(\tau\) is the probability of disease transmission per unit of contact time. The expected number is almost the same in two cases, except that case 1 is larger by a second-order difference: \((1 - (1 - \tau)^{d_1}) * (1 - (1 - \tau)^{d_2})\).
### Table 7: Average structure properties of several city-scale contact networks

<table>
<thead>
<tr>
<th>Network Description</th>
<th>No. of nodes</th>
<th>Number of edges</th>
<th>Avg. degree</th>
<th>Avg. edge weight (minute)</th>
<th>Avg. CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>people-people network</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>the coarse network of Delhi [30]</td>
<td>526,234,615</td>
<td>13,850,507</td>
<td>75.99</td>
<td>162</td>
<td>0.482</td>
</tr>
<tr>
<td>the refined network of Delhi</td>
<td>206,787,386</td>
<td>13,850,507</td>
<td>29.86</td>
<td>363</td>
<td>0.546</td>
</tr>
<tr>
<td>the Los Angeles network [31]</td>
<td>459,273,880</td>
<td>16,228,759</td>
<td>56.60</td>
<td>141</td>
<td>0.389</td>
</tr>
</tbody>
</table>

#### 6.0.1 Epidemic Dynamics and Intervention Policies

We run exactly the same simulations with the coarse network as we do for the refined network in section 4. We listed the results with $R_0 = 1.35$ in Figure 13 and Figure 14 to compare against with the refined network. Similarly, the results with $R_0$ value of 1.40, 1.45 and 1.60 are omitted; but they are also very similar in the case of $R_0 = 1.35$ here. Please refer to appendix D for complete results.

![Figure 13: The vulnerability histogram in the coarse network and the refined network with H1N1 when $R_0 = 1.35$. The histogram shows very different structural properties for the two networks.](image)

The vulnerability distribution in Figure 13 shows a very clear difference between the two networks. Distribution of the refined network is generally flat, but the coarse network changes up and down violently. Compared to the refined network, the coarse network contains more high vulnerability nodes (those above 0.8) and less low vulnerability nodes (those less than 0.4). This difference is consistent with their different activity schedules. Nodes in the coarse network have a busier schedule, so they are exposed to more people and become more vulnerable. On the other hand, the refined network contains 40% at-home people and they should account for the large low vulnerability people.

The coarse network and the refined network differ a lot in epidemic dynamics despite the $R_0$ calibration, as shown in Figure 14, where we use the tuple (attack-rate, peak, peak-day) to characterize epidemic dynamics. For either the base case without intervention, or the cases under various intervention policy, the coarse network has a much higher attack rate, higher peak and earlier outbreak dates than the refined network, conforming to our earlier observation.

Both networks indicate that “Vaccination” is the most effective intervention strategy in all candidates. Non-pharmaceutical interventions such as school closure and work closure have similar effects across the two networks. Nevertheless, we can see the disparate network precision may lead us to draw different conclusions in selecting a right policy in some scenarios. For example, the
effects of two policies, Antiviral and School Closure are very different in the coarse network and the refined network. School Closure seems a better solution to delay the disease outbreak than Antiviral based on the network of the coarse network. However, the conclusion is quite different if we are going to choose between the two strategies based on the refined network.

7 Conclusion

Social contact network plays an important role in understanding disease spreading in the population. A realistic social contact network can also help to make a more reasonable public health policy. In this paper, we introduce a novel methodology to generate high resolution city-scale social contact networks by integrating data from multiple sources. We exemplify how to create the network for the city of Delhi.

Modeling the people-people contacts typically involves region specific data regarding the demographics, the land use and the activity patterns. The quality of the data is critical in creating a realistic network. In a previous paper, we have generated the first version of Delhi Network with very coarse data. We compare it to the high resolution network constructed in this paper in terms of their construction methodology, network structure and corresponding epidemic dynamics. The experiments show that, the prediction to the epidemic dynamics and the selection of intervention policy towards the same population might be quite different if we use networks based on different level of details. For the coarse network and the refined network, the difference is large regarding the contact structure. The epidemic dynamics varies a lot too despite that we calibrate the $R_0$ to have same value in the two networks.

Lastly, we try to attack a fundamental question in social network synthesis, what’s the requisite level of detailed data for generating social contact networks? we conduct a serial of sensitivity test to explore the problem. We figure that details are important but not all factors in the framework matters. For example, the location distribution over the geographical space seems a trivial factor within our framework however the sublocation structure inside a location is a significant factor.
References


A  additional tables for section 4.3.3 (TLC)

B  Comparison study for the Delhi network under the models with various sublocation size
Table 8: results of Targeted-Layered Containment when R0 is 1.40.

<table>
<thead>
<tr>
<th>R0</th>
<th>compliance</th>
<th>threshold</th>
<th>AR</th>
<th>peak</th>
<th>peak-day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.40</td>
<td>0.3</td>
<td>0.0001</td>
<td>0.459784</td>
<td>0.0103476</td>
<td>128.867</td>
</tr>
<tr>
<td>1.40</td>
<td>0.3</td>
<td>0.001</td>
<td>0.44541</td>
<td>0.00837585</td>
<td>132.3</td>
</tr>
<tr>
<td>1.40</td>
<td>0.6</td>
<td>0.0001</td>
<td>0.459968</td>
<td>0.0103727</td>
<td>148.067</td>
</tr>
<tr>
<td>1.40</td>
<td>0.6</td>
<td>0.001</td>
<td>0.446039</td>
<td>0.00849277</td>
<td>154.033</td>
</tr>
<tr>
<td>1.40</td>
<td>0.9</td>
<td>0.0001</td>
<td>0.461573</td>
<td>0.0105962</td>
<td>111</td>
</tr>
<tr>
<td>1.40</td>
<td>0.9</td>
<td>0.001</td>
<td>0.447562</td>
<td>0.00856637</td>
<td>165.73</td>
</tr>
</tbody>
</table>

Table 9: results of Targeted-Layered Containment when R0 is 1.45.

<table>
<thead>
<tr>
<th>R0</th>
<th>compliance</th>
<th>threshold</th>
<th>AR</th>
<th>peak</th>
<th>peak-day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.45</td>
<td>0.3</td>
<td>0.0001</td>
<td>0.484626</td>
<td>0.0117089</td>
<td>123</td>
</tr>
<tr>
<td>1.45</td>
<td>0.3</td>
<td>0.001</td>
<td>0.469992</td>
<td>0.00960018</td>
<td>125.333</td>
</tr>
<tr>
<td>1.45</td>
<td>0.6</td>
<td>0.0001</td>
<td>0.484776</td>
<td>0.0117296</td>
<td>140.667</td>
</tr>
<tr>
<td>1.45</td>
<td>0.6</td>
<td>0.001</td>
<td>0.472081</td>
<td>0.00967723</td>
<td>144.867</td>
</tr>
<tr>
<td>1.45</td>
<td>0.9</td>
<td>0.0001</td>
<td>0.486354</td>
<td>0.0119619</td>
<td>105.6</td>
</tr>
<tr>
<td>1.45</td>
<td>0.9</td>
<td>0.001</td>
<td>0.473562</td>
<td>0.0098646</td>
<td>156.3</td>
</tr>
</tbody>
</table>

Table 10: results of Targeted-Layered Containment when R0 is 1.60.

<table>
<thead>
<tr>
<th>R0</th>
<th>compliance</th>
<th>threshold</th>
<th>AR</th>
<th>peak</th>
<th>peak-day</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.60</td>
<td>0.3</td>
<td>0.0001</td>
<td>0.548933</td>
<td>0.015805</td>
<td>107.267</td>
</tr>
<tr>
<td>1.60</td>
<td>0.3</td>
<td>0.001</td>
<td>0.535696</td>
<td>0.0133405</td>
<td>109.033</td>
</tr>
<tr>
<td>1.60</td>
<td>0.6</td>
<td>0.0001</td>
<td>0.549292</td>
<td>0.0158388</td>
<td>122.833</td>
</tr>
<tr>
<td>1.60</td>
<td>0.6</td>
<td>0.001</td>
<td>0.538847</td>
<td>0.0135529</td>
<td>126.133</td>
</tr>
<tr>
<td>1.60</td>
<td>0.9</td>
<td>0.0001</td>
<td>0.549467</td>
<td>0.0158853</td>
<td>131.4</td>
</tr>
<tr>
<td>1.60</td>
<td>0.9</td>
<td>0.001</td>
<td>0.540429</td>
<td>0.0138356</td>
<td>135.26</td>
</tr>
<tr>
<td>1.60</td>
<td>0.9</td>
<td>0.01</td>
<td>0.55064</td>
<td>0.016119</td>
<td>93.3</td>
</tr>
</tbody>
</table>
Figure 15: Epidemics and policy efficacy in the Delhi network with various sublocation sizes (R0=1.35).

Figure 16: Epidemics and policy efficacy in the Delhi network with various sublocation sizes (R0=1.40).

C Sensitivity to Location Switches
Figure 17: Epidemics and policy efficacy in the Delhi network with various sublocation sizes (R0=1.45).

Figure 18: Epidemics and policy efficacy in the Delhi network with various sublocation sizes (R0=1.60).
Figure 19: impact of location switches to the Delhi network (R0=1.35).

Figure 20: impact of location switches to the Delhi network under different public health policies (R0=1.35).

D Comparison study between the coarse network and the refined network
Figure 21: impact of location switches to the Delhi network under different public health policies (R0=1.40).

Figure 22: impact of location switches to the Delhi network under different public health policies (R0=1.45).
Efficacy of Various Interventions (R0=1.60)

Figure 23: impact of location switches to the Delhi network under different public health policies (R0=1.60).

Epidemics and policy efficacy in the coarse network and the refined network (R0=1.35).

Figure 24: Epidemics and policy efficacy in the coarse network and the refined network (R0=1.35).

E Vulnerability Distribution for the coarse network and the refined network
Figure 25: Epidemics and policy efficacy in the coarse network and the refined network (R0=1.40).

Figure 26: Epidemics and policy efficacy in the coarse network and the refined network (R0=1.45).
Figure 27: Epidemics and policy efficacy in the coarse network and the refined network (R0=1.60).

Figure 28: Vulnerability Distribution for the coarse network and the refined network.