Evaluating strategies for pandemic response in Delhi using realistic social networks

Huadong Xia, Kalyani Nagaraj, Jiangzhuo Chen and Madhav Marathe
Network Dynamics and Simulation Science Laboratory at VBI,
Virginia Tech, Blacksburg, VA 24061, USA
Email: {xhd,kalyanin}@vt.edu,{chenj,mmarathe}@vbi.vt.edu

***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 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 suggests that: (i) pharmaceutical containment strategies typically are more effective than the non-pharmaceutical ones 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 better public health policy. To the best of our knowledge this is the first such study in the Indian sub-continent.

I. Introduction

Today's densely populated urban regions enable rapid transmission of infectious diseases [40]. Additionally, urban contact networks in regions like India and China are rapidly growing. The National Capital Territory region of Delhi is predicted to see 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 [32]. In Beijing, the population has risen from 12.9 million in 2000 to 18.8 million in 2010 [31]. The densely populated large urban regions provide a perfect fabric for rapid spread of infectious diseases. Public health authorities have focused on developing effective interventions and policies to control the spread of diseases using both pharmaceutical interventions as well as social distancing measures. Both the strategies effectively reduce the connectivity within the social contact network or change the transmission probability between individuals

Over the last 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 [37]. The basic approach has now become widely accepted in the literature. It is based on the idea that a better understanding of the characteristics of the social contact network can give better insights into disease dynamics and intervention strategies for epidemic planning.

A methodology to synthesize social contact networks for the US cities is already in place. 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, which are calibrated to the above data, has been shown to influence the outcome of disease outbreaks in our simulated epidemic models [16], [37].

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) [33] captures the interdependence of people's activities, especially adults, in the same household across all surveyed households in the United States. Data of similar level of detail is not available for Delhi (and will not be for many other regions as well), making it impossible to replicate the US network generation process for regions outside the US.

A. Summary of contributions

Building on our earlier work, we construct a synthetic social contact network for NCT-I. To overcome the limitations as regards to available data for Delhi, we developed several new methods – several of these methods 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 South 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 $(p, \ell, label) \in E(G_{PL})$ between them, where $label$ is a record of the type of activity of the visit and its start and end points. 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) is 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) is 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 disease over the synthetic social contact network of NCT-I. We study the efficacy of various intervention strategies. This includes both pharmaceutical interventions as well as 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.

II. Related Work

Traditionally, mathematical and computational modeling of epidemics has focused on aggregate models using coupled rate equations [2]. 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 epidemiology – 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 [4], [29]. 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], [17]–[19], [34] for examples of this approach. These models can be extended to obtain hybrid models as well. In a hybrid model, counties are represented as nodes and edges are added between counties to capture the movement of individuals – see [1], [12], [27] for a comparative study. Epidemic dynamics within a county are computed using an individual-based model. The dynamics over network of counties are captured using coupled rate equations.

The final class of models use the most realistic representation of social contact networks; see [6], [15], [28]. In [6], [8], [9], [15] 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.

III. Network Generation: data and methodology

To study the intervention strategies for pandemic response, it is important that 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.

A. Data Collection

Precisely, Delhi refers to the National Capital Territory (NCT) of India. It is the capital of India, including New Delhi and several adjacent urban areas. It contains over 13 million people and is one of the areas with highest population density in the world. The population is young with more males than females. Some statistics for the population can be found in table I in comparison with the two other representative cities in the world.

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 II.

B. Network Construction Methodology

Our method follows the steps we use when constructing networks of another area [5]: 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 [5], but to accommodate region-specific data in table II, we also design some novel methodologies. In the following, section III-B2 and III-B4 are new methods and are described in more detail, other steps can be found with more details in paper [5].

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, and meanwhile build a realistic household structure for all those individuals. For this purpose, both summary statistics of interested demographic variables in household level (India Census 2001 [21]) and a collection of household samples from a survey to Delhi [14] are required.

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

TABLE I: Demographic statistics of Delhi in comparison with other two cities

	Data Set	Description
Demographics	India Census 2001 [21]	Statistics for demographic variables such as age, gender, income, etc. in individual level and household level for Delhi residents.
	Household Microdata: India Human Development Survey 2005 by the University of Maryland and the National Council of Applied Economic Research [14]	Micro samples for household structure. It depicts each household sample: size, income, householder age, house types; and also for each individual in the hh: demographic details, religion, work, marital status, relationship to the householder, etc.
Locations	MapMyIndia Dataset [24]	It includes the following information for Delhi: (1) Ward-wise statistics for population and households; (2) Coordinates for locations such as residential areas, schools, shopping centers, hotels etc. (3) Infrastructures such as roads, railway stations, land use etc. (4) Boundary for each city, town and ward.
Activity	Thane Travel Survey by USF [36]	The dataset collect travel survey for residents in Thane, an Indian city similar to Delhi. Activity templates are extracted from travel statistics, and assigned to the synthetic population with a decision tree.
	00-07 school attendance statistics from the UNESCO Institute of Statistics (UIS) [25]	It is used to decide the fraction of kids as students.
	India residential area activity survey by Network Dynamics and Simulation Science Laboratory (NDSSL) at Virginia Tech	The survey focused on approximately 40% of population in Delhi without daily travels. We collected people's age, gender, and contact statistics near their home.

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

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, the sample selection is critical. From the census data we collected the distributions of available demographic variables in 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 [5]. 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 initial cell value might affect the accuracy of final results, we put in the estimation value from the micro household sample data before the first iteration.

2) Activity assignment using the 2001 Thane, India household travel survey statistics: Due to the unavailability of travel survey data in Delhi at the time of this study, we devise a discrete-time simulation to assign detailed activity schedules to the Delhi synthetic population using the 2001 Thane household travel survey statistics described by Nehra [36] and Banerjee [3]. Thane is a city in the western state of Maharashtra, India. A quick comparison based on census data [21] reveals the high similarity between the two India cities regarding demographic structure and religious/cultural habits, therefore we believe Thane is a reasonable proxy for Delhi.

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. Additionally, the survey collected sociodemographic information from respondents and their respective households. Literature on the Thane travel survey describes travel data statistics in the form of empirical frequency distributions of trip start times and trip durations of adults in the surveyed 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, students older than 15, and 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 [36] and [3], 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, gender, etc.). For adults, this is achieved by sampling from the commuter status and demographic distribution of adults in the survey population reported in [3]. 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 to 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 that make 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. Those assumptions are made based on observations from the real world for reasonability. 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 [25] 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 [3]: 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 [3], 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 ξ

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. $[\xi, actCLASS_i] = f_1(age_i, gender_i, \xi)$;

`/* assign activity class */`

2. $[\xi, actSEQ_i] = f_2(actCLASS_i, \xi)$;

`/* assign activity sequence */`

3. **for** *each activity* j *in* $actSEQ_i$ **do**

`/* generate detailed schedule */`

$[\xi, startTime_{i,j}] = g(actSEQ_i.activity_j, endTime_{i,j-1}, \xi)$

$[\xi, endTime_{i,j}] = h(actSEQ_i.activity_j, startTime_{i,j}, \xi)$

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

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 [24] 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. The locations of those households within each ward are missing, and we choose

to distribute them along streets etc. for two reasons. First, residential area is typically close to some streets/lanes in the real world. Second, the algorithm helps us avoid putting home locations to a place or infrastructure geographically unsuitable to live, such as lakes, rivers or railways.

Once people are assigned to locations, we will further model their interactions within the locations via sublocations. 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 V.

4) Contacts in residential area: The above methodology has been applied to generate several other cities in the world [5], [11]. 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 [14], and the travel survey we retrieved from [36]. Reference [36] claims that 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.

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

As introduced in the last section, we generate the Delhi network for the city of Delhi based on the high resolution data and novel methodologies. A high resolution social contact network reserve effective contact structure in the population. It will provide insights for policy makers in studying the epidemic dynamics and evaluating effective public health policies. In the following, we conduct a detailed analysis of the synthetic Delhi population and the Delhi network. In using such a network and the powerful epidemic simulation platform EpiFast [8], we study various intervention strategies to contain the spread of disease in the city.

A. 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

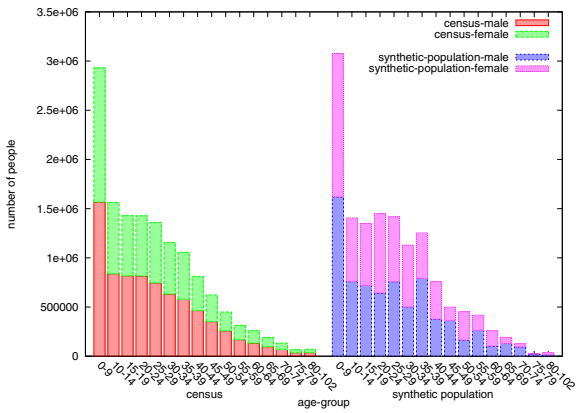


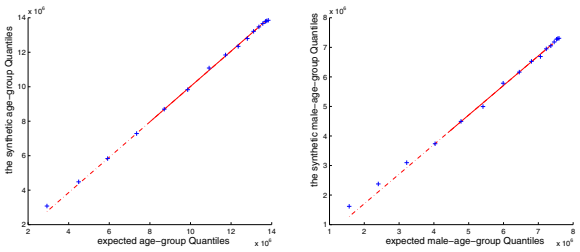
Fig. 1: On the left side is the age-group counts for Delhi from India Census 2001 [21]. On the right side is the age-group counts of the synthetic population based on micro sample data and household level statistics. Visually we see that the synthetic population conforms to the real statistics quite well. Q-Q plots in Figure 2 give us a clearer visualization on this.

in terms of individual level demographics. To verify, we plot Figure 1 and 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 slightly for adults (refer to the deviation in counts of unisex groups in Figure 2b). The results suggest that both our model and the implementation are reasonable. 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% of people do not have a job, as discussed earlier. Most people work or study during the day, and almost all people stay at home late at night. If the survey basis is accurate, this unique cultural feature is quite different from a US city example.

B. Graph Structural Properties of the Contact Networks

A detailed profiling to the network structure is shown in Figure 4. We plot the distribution of node degrees, clustering coefficients and contact durations. 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 III. Different from theoretic assumptions such as power law degree distribution, the degree distribution of the Delhi network is peaked around a degree of 20. The average degree is about 30, which is a relative small number based on our other study of US cities [11]. Similarly, compared to those of the Los Angeles network, the average edge weight (representing accumulated contact duration) is longer, the average clustering coefficient is



(a) the general age-group quantiles. (b) the male age-group quantiles.

Fig. 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.

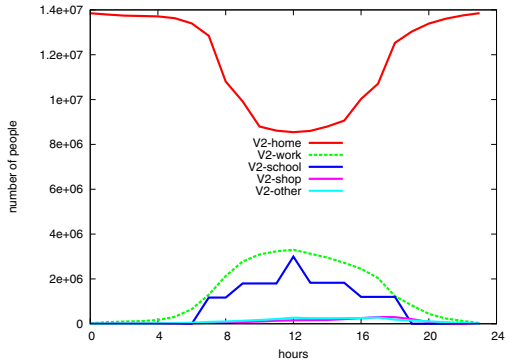


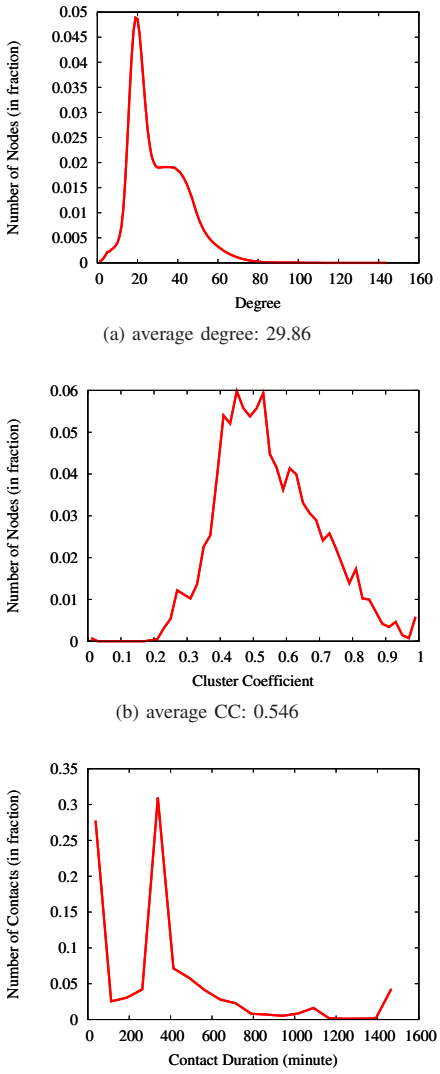
Fig. 3: Activity statistics of the synthetic population. For each type of activity, we calculate for each hour the number of people that conduct that kind of activity. During the late night(hour 0 and 24), almost all people stay at home.

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 contact structure has implications to the pandemic spreading in the population.

C. 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 to be simulated is H1N1 which occurred in a 2009 global outbreak and is still prevalent in India [20]. To address the variations in different estimates of R_0 of H1N1 in literatures [22], [26], [42], 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.

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 5. The distributions for other R_0 are very similar to that of $R_0 = 1.35$ (omitted here to save space, refer to reference [41] for complete results), indicating that the node vulnerability is more relevant to the network structure than to the disease property. This implies the following observations from the vulnerability distribution are applicable to a multitude of diseases regardless of their R_0.



(a) average degree: 29.86

(b) average CC: 0.546

(c) average contact duration: 363 minutes

Fig. 4: Network structure profiling for the Delhi Network. By comparing it with several other city-scale networks (Table III), we can get a better sense of the relation between the structure and the epidemic dynamics.

The vulnerability distribution of the people varies from 0 to 1, biased towards the left side. Quite a few people have a vulnerability close to 0. Compared to other populations (refer to section VI), 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.

2) Optimal intervention strategies during epidemic spreading: Using a high resolution contact network modeled for Delhi, we are able to get a good understanding for the epidemics and effectiveness of different intervention policies. We simulate four public health policies frequently applied in

Network	No. of nodes	number of edges	Avg. degree	Avg. edge weight (minute)	Avg. CC
the Delhi network	206,787,386	13,850,507	29.86	363	0.546
the Los Angeles network [11]	459,273,880	16,228,759	56.60	141	0.389

TABLE III: Average structure properties of several city-scale contact networks

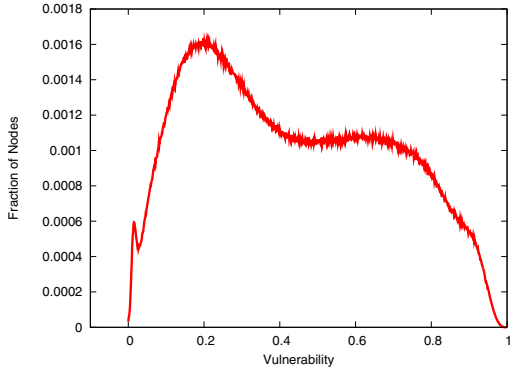


Fig. 5: The vulnerability histogram for nodes in the Delhi network, calculated based on epidemic simulations with $R_0 = 1.35$. In the figure, the fraction of low vulnerability nodes are more than the fraction of high vulnerability, where about 40% of people have a vulnerability lower than 0.35.

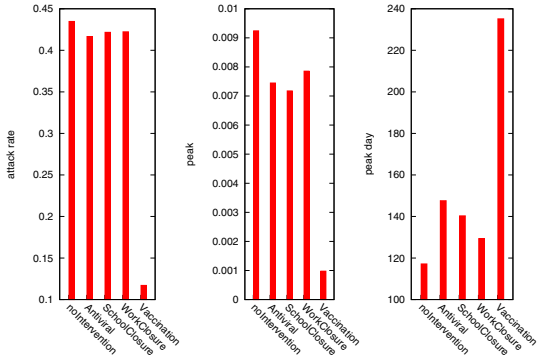


Fig. 6: Epidemics under various intervention strategies in the Delhi network when $R_0 = 1.35$, including a base case where no intervention is conducted. Here we use the tuple (attack-rate, peak, peak-day) to characterize epidemic dynamics. For Vaccination and Antiviral, we randomly choose 25% of the population to apply corresponding pharmaceutical treats. School Closure and Work Closure are applied for 3 weeks when 0.1% of the nodes in the city get infected.

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 6. 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 reference [41]. The insensitivity to R_0 here and in the following chapters suggests that the ordering of the policy effects remain the same regardless of the R_0 value of a disease.

Vaccination has the strongest effect in containing the disease spread. All the other policies, including Antiviral,

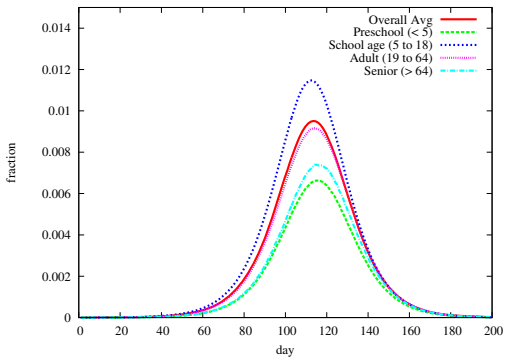


Fig. 7: Epidemic curves show subpopulation epidemics in the Delhi network when $R_0 = 1.35$. The Delhi population is partitioned to four groups based on age: preschool, school age, adult, and senior. Each dashed curve shows the fraction of people in that subpopulation infected on each day in the base case in Figure 6, where there is no intervention. The red curve shows the fraction of people in the whole Delhi population infected on each day in the same base case.

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 epidemic from an emerging disease. 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 as well.

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 7 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 7, among all subpopulations, only school age has an epidemic worse than the population average (red curve in figure). Closing schools can avoid disease transmissions between students within schools, which explains the high effectiveness of School Closure.

V. 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 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 characterizes the internal structure of a location. We define for each type of location an empirical value as their sublocation size. Please note that sublocation size is region specific value and should be adjusted based on local statistics when we model another area. Also we apply the gravity model to assign locations for activities. Based on observation in the real world [7], the gravity model suggests that the distance between one's home and work place or shopping center etc. follows an exponential distribution: $f(x;\lambda) = \lambda e^{-\lambda x}$ where $\frac{1}{\lambda}$ is the mean distance.

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 IV. Here we assume $R_0 = 1.35$. We point out, however, that the observations are similar for the sensitivity experiments with R_0 value being 1.40, 1.45, or 1.60.

A. Sensitivity to Sublocation Size

The sensitivity test results to various sublocation size are shown in Figure 8. Obviously, varying the sublocation size has a significant impact to either the disease spreading or the intervention to the spreading because it changes the contact density within locations. 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 comparing 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.

B. 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 9, 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 that in terms of epidemics, people's interaction pattern in a local place is more important than the location

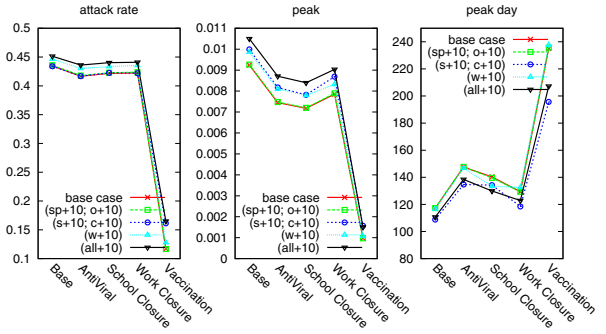


Fig. 8: Epidemics and policy efficacy in the Delhi network with various sublocation sizes ($R_0 = 1.35$). Six types of locations are modeled in the Delhi network: home(h), work-place(w), school(s), college(c), shops(sh) and other(o). In the base case, we define the sublocation sizes for those locations 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.

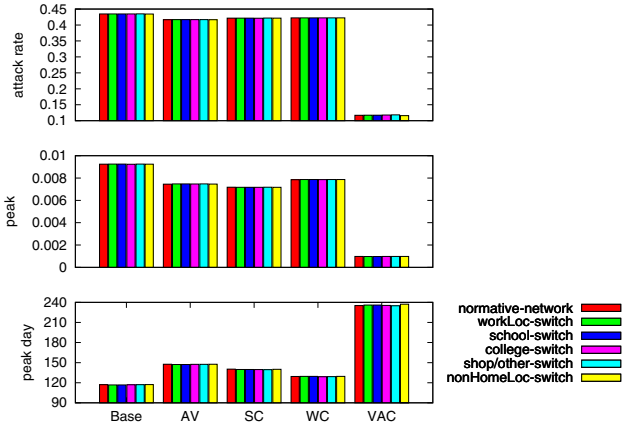


Fig. 9: Impact of location switches to the Delhi network under different public health intervention policy ($R_0 = 1.35$). Two people are selected randomly to exchange their daytime locations (named one switch) only when the locations are of same type. In the control cases in the figure, "workLoc-switch" means work-places are switched randomly between workers. Other legends is self-evident in the meaning. For each case, we conduct a large number of switches to assure system convergence. The location switches in all cases do not change the epidemic dynamics of the underlying networks however.

distribution in the city globally, given that the population density is unchanged.

VI. Comparison Study between the Coarse Network and the Refined Network

In our previous paper [11], 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 IV. 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

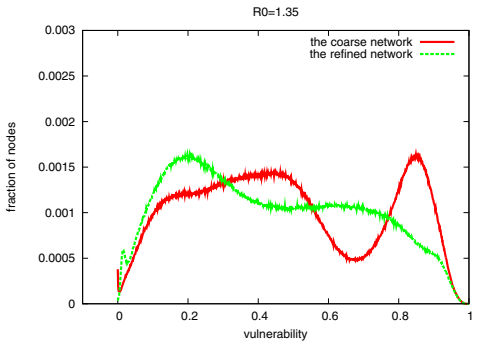


Fig. 10: The vulnerability histogram in the coarse network and the refined network with H1N1 when R0=1.35. The histogram shows very different structural properties for the two networks.

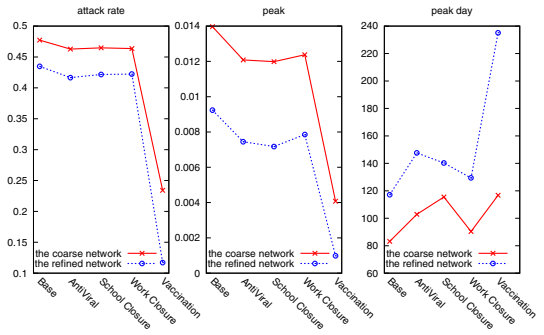


Fig. 11: Epidemics and policy efficacy in the coarse network and the refined network under H1N1 when R0=1.35. The epidemic parameter and the intervention strategy setups are same as the case in Figure 6. The two networks show significantly different epidemic dynamics, despite the calibration of R_0 to 1.35 for the two networks.

the exact location information of Delhi, so LandScan data is used which contains only the population density distribution over the region. More importantly, there were no activity survey data for India when we created the coarse network, only some aggregate statistics in the literature are available, including average time for school activities [35], [38] and work activities [30]. We compose simple yet reasonable activity sequences and calibrate each activity's duration based on the statistics we collected. 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.

The network structure profiling for the two networks are listed in Table III. Compared to the refined network, the coarse network has a much higher degree (76.99 v.s. 29.86) and lower edge weight (contact duration) on average. We deduce that such difference makes diseases spread easier in the coarse network[1]. Similarly, in regards of the clustering coefficient distribution, the refined network has a higher average clustering coefficient, which also helps hinder the disease spreading.

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 IV. We listed the results with R_0 1.35 in Figure 10, 11 to compare against the refined network. Similarly, the results with other R_0 values are omitted but they are also very similar to the case of R_0 1.35 here. Please refer to reference [41] for complete results.

The vulnerability distribution in Figure 10 reveals a clear difference between the two networks. Distribution of the refined network is generally flat, but that of 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 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 11, 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 much higher attack rate, higher peak and earlier outbreak dates than the refined network, conforming to our observation earlier.

Both networks indicate "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 that 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.

[1]Let's consider two simplified cases. Case 1, a seed node u has two contacts with durations $d1$ and $d2$. Case 2, a seed node u has one contact with duration $(d1 + d2)$. The expected number of secondary infections in case 1 is $(1 - (1-\tau)^{d1}) + (1 - (1-\tau)^{d2})$; that in case 2 is $1 - (1-\tau)^{d1+d2}$, where τ is the probability of disease transmission per unit of contact time. The expected number is almost the same in two cases, except that in case 1 is larger by a second-order difference: $(1 - (1-\tau)^{d1}) * (1 - (1-\tau)^{d2})$.

ACKNOWLEDGMENT

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.

REFERENCES

[1] M. Ajelli, B. Goncalves, D. Balcan, V. Colizza, H. Hu, J. Ramasco, S. Merler, and A. Vespignani. Comparing large-scale computational approaches to epidemic modeling: Agent-based versus structured metapopulation models. *BMC Infectious Diseases 2010*, 10(190), 2010.

	the coarse network	the refined network
Demographics	India Census 2001 [21]	India Census 2001 [21]+Household Microdata [14]
Locations	LandScan 2007 [23] School/College Statistics [13], [39]	MapMyIndia: information of real locations [24]
Activity	Delhi time use statistics [30], [35], [38]	2001 Thane Household Travel Survey [36]; India residential area activity survey by NDSSL

TABLE IV: Data used in construction of the coarse network and the refined network.

[2] N. Bailey. *The Mathematical Theory of Infectious Diseases and Its Applications*. Hafner Press, New York, 1975.

[3] Amlan Banerjee. *Understanding activity engagement and time use patterns in a developing country context*. PhD thesis, University of South Florida, 2006. Paper 2451. http://scholarcommons.usf.edu/etd/2451.

[4] A. Barrat, M. Barthelemy, and A. Vespignani. *Dynamical processes in complex networks*. Cambridge University Press, 2008.

[5] C. Barrett, R. Beckman, M. Khan, V.S.A. Kumar, M. Marathe, P. Stretz, T. Dutta, and B. Lewis. Generation and analysis of large synthetic social contact networks. In *WSC*, December 2009.

[6] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe. EpiSimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC'08*, pages 290–294, 2008.

[7] C.L. Barrett, R.J. Beckman, M. Khan, V. Kumar, M.V. Marathe, P.E. Stretz, T. Dutta, and B. Lewis. Generation and analysis of large synthetic social contact networks. In *Winter Simulation Conference (WSC), Proceedings of the 2009*, pages 1003 –1014, Dec. 2009.

[8] K. Bisset, J. Chen, X. Feng, VS A. Kumar, and M. Marathe. EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In *ICS*, pages 430–439, 2009.

[9] Keith R. Bisset, Xizhou Feng, Madhav V. Marathe, and Shrirang M. Yardi. Modeling interaction between individuals, social networks and public policy to support public health epidemiology. In *Winter Simulation Conference*, pages 2020–2031, 2009.

[10] D. L. Chao, M. E. Halloran, Valerie Obenchain, and I. M. Longini Jr. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, 6(1), 2010.

[11] J. Chen, F. Huang, M. Khan, M. Marathe, P. Stretz, and H. Xia. The effect of demographic and spatial variability on epidemics: a comparison between Beijing, Delhi, and Los Angeles. In *Proceedings of the 5th International conference on Critical Infrastructures*, 2010.

[12] V. Colizza, A. Barrat, M. Barthelemy, A. Valleron, and A. Vespignani. Modeling the worldwide spread of pandemic influenza: baseline case and containment interventions. *PLoS Medicine*, 4:95, 2007.

[13] Delhi Department of Planning. Economic Survey of Delhi 2005-2006, Section 15 . http:/delhiplanning.nic.in/.

[14] Sonalde Desai, Reeve Vanneman, and New Delhi National Council of Applied Economic Research. India human development survey (IHDS) 2005. Available online: http://dx.doi.org/10.3886/ICPSR22626.v8.

[15] S. G. Eubank. Scalable, efficient epidemiological simulation. In *ACM Symposium on Applied Computing*, pages 139–145, Madrid, Spain, March 2002.

[16] S. G. Eubank, H. Guclu, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 4:180–184, May 2004.

[17] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke. Strategies for mitigating an influenza pandemic. *Nature*, 442:448–452, 2006.

[18] Neil M. Ferguson, Matt J. Keeling, W. John Edmunds, Raymond Gani, Bryan T. Grenfell, Roy M. Anderson, and Steve Leach. Planning for smallpox outbreaks. *Nature*, 425:681–685, 2003.

[19] Timothy C. Germann, Kai Kadau, Ira M. Longini, and Catherine A. Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940, 2006.

[20] Health.india.com. Delhi swine flu update: 37 more cases, total 457. http://health.india.com/news/delhi-swine-flu-update-37-more-cases-total-457/, Feb 19th, 2013.

[21] India-Government. India census 2001 and 2011.

[22] Zhen Jin, Juping Zhang, Li-Peng Song, Gui-Quan Sun, Jianli Kan, and Huaiping Zhu. Modelling and analysis of influenza A (H1N1) on networks. *BMC public health*, 11 Suppl 1(Suppl 1):S9, January 2011.

[23] Oak Ridge National Laboratory. LandScan Data, Global Population Project at Oak Ridge National Lab. http://www.ornl.gov/sci/landscan/.

[24] MapMyIndia. Demographic and geo-spatial dataset for delhi, 2011.

[25] Unicef Media. Unicef: State of the world's children 2009. *Maternal and Newborn Care. New York, 2009.*, 2009. Available online at http://www.unicef.org/sowc09/report/report.php.

[26] Gautam I Menon and Sitabhra Sinha. Epidemiological Dynamics of the 2009 Influenza A(H1N1)v Outbreak in India. *preprint*, pages 1–5, 2010.

[27] Stefano Merler and Marco Ajelli. The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Processings of Royal Society*, 277(1681):557–565, 2010.

[28] Lauren Ancel Meyers. Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of The American Mathematical Society*, 44:63–86, 2007.

[29] Lauren Ancel Meyers and Nedialko Dimitrov. Mathematical approaches to infectious disease prediction and control. *INFORMS, Tutorials in Operations Research*, 2010.

[30] R L Narasimhan and R N Pandey. some main results of the pilot time use survey in india and their policy implications. *the International Seminar on Time Use Studies, 7-10 December.*, 1999.

[31] China. National Bureau of Statistics. National bureau of statistics database. http://www.stats.gov.cn/english/.

[32] Department of Environment and Forests of India. State of environment report for delhi, 2010. http://delhi.gov.in/, 2010.

[33] US Department of Transportation. National Household Travel Survey. http://nhts.ornl.gov/, 2009.

[34] Jon Parker and Joshua M. Epstein. A distributed platform for global-scale agent-based models of disease transmission. *ACM Transactions on Modeling and Computer Simulation*, 22(1), 2012.

[35] Delhi public school. class schedule of Delhi public school. http://dpsrkp.net.

[36] Nehra R.S. Modeling Time Space Prism Constraints in a Developing Country Context. Master's thesis, University of South Florida, 2004.

[37] Eubank S., Kumar V.S.A., Marathe M., Srinivasan A., and Wang N. Structure of Social Contact Networks and Their Impact on Epidemics. *In DIMACS Series in Discrete Mathematics and Theoretical Computer Science 70*, pages 179–181, 2006.

[38] the British school school at New Delhi. class schedule of the British school at New Delhi. http://www.british-school.org/home/the-school-day.html.

[39] University Grants Commission. India School/College Statistics. http://www.ugc.ac.in/.

[40] WHO. Summary of probable SARS cases with onset of illness from 1 November 2002 to 31 July 2003. http://www.who.int/csr/sars/country/table2004_04_21/en/index.html, retrieved 10-31-2008.

[41] Huadong Xia, Kalyani S. Nagaraj, Jiangzhuo Chen, and Madhav V. Marathe. Evaluating strategies for pandemic response in Delhi using realistic social networks. . *NDSSL Technical Report: 13-016.*, 2013.

[42] Yang Yang et al. The transmissibility and control of pandemic influenza A (H1N1) virus. *Science*, 326(5953):729–33, October 2009.