NDSSL Technical Report 10-111

August 23, 2010

Title: The Effect of Demographic and Spatial Variability on Epidemics: A Comparison between Beijing, Delhi, and Los Angeles

Authors: Jiangzhuo Chen
Fei Huang
Maleq Khan
Madhav Marathe
Paula Stretz
Huadong Xia

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 NSF Nets Grant CNS-0626964, NSF HSD Grant SES-0729441, NIH MIDAS project 2U01GM070694-7, NSF PetaApps Grant OCI-0904844, DTRA R&D Grant HDTRA1-0901-0017, DTRA CNIMS Grant HDTRA1-07-C-0113.

Network Dynamics and Simulation Science Laboratory
Virginia Bioinformatics Institute
Virginia Polytechnic Institute and State University

The Effect of Demographic and Spatial Variability on Epidemics: A Comparison between Beijing, Delhi, and Los Angeles

Jiangzhuo Chen, Fei Huang, Maleq Khan, Madhav Marathe,
Paula Stretz, and Huadong Xia

Network Dynamics and Simulation Science Laboratory
Virginia Bioinformatics Institute, Virginia Tech
1880 Pratt Drive, Blacksburg, VA 24061.
Email: {chenj,huangf,maleq,mmarathe,pstretz,xhd}@vbi.vt.edu

Abstract. A social network is a critical infrastructure for the propagation of an infectious disease in a population. It is important to study the structural properties of the social network for identifying feasible public health interventions that can effectively contain a potential epidemic outbreak. In this work, we focus on flu-like diseases and corresponding people-people social contact networks. We study such social infrastructures of three cities: Los Angeles, USA, Beijing, China and Delhi, India. These contact networks are different due to different construction methodologies and the fact that the populations inherently have very different demographic structures and activity patterns. We compare them in terms of static structural properties (such as clustering coefficient, degree distribution), as well as disease dynamics and efficacy of intervention (e.g., school closure). The study of our model robustness and the comparison between different contact networks can provide valuable insight on creating a global synthetic population and social infrastructure for studying public health problems.

1 Introduction

The structure of social contact networks influences the spread of infectious diseases in an urban region. An important goal of network science is the development of *structure to function theory* - identifying structural features of the social network that yield insights into the disease propagation. Effective pharmaceutical as well as non-pharmaceutical interventions (NPI) can be identified by analyzing the social contact networks. In this work, motivated by the recent pandemic caused by H1N1, we focus on influenza-like diseases and corresponding people-people social contact networks. We study such social infrastructures of three cities: Beijing, China, Delhi, India, and Los Angeles, USA. We compare them in terms of static structural properties, as well as disease dynamics and intervention efficacy.

We have generated synthetic populations and social contact networks for Beijing, Delhi, and Los Angeles using different methodologies. For Los Angeles,

we use census data, Dun & Bradstreet location data, and activity survey data. For Beijing and Delhi, we have only the LandScan population density data and limited census data. We develop a generic methodology that takes into account variable data availability and granularity across different regions of the world. This model, based on LandScan data, can be applied to generate a synthetic population, including individual demographics, home locations, and daily activities, for any area in the world. The Beijing, Delhi, and Los Angeles contact networks are different due to different construction methodologies and the fact that the three populations inherently have very different demographic structures and activity patterns. To construct the contact networks, we explicitly generated activity sequences for each individual in the population taking into account the variability in demographic and activity patterns for each city individually. Other works in current literature either ignore the detailed activities of the individual persons or use the same model for every city/country in the world, and thus lose the crucial demographic and spatial variability which has significant effect on network structure as well as disease dynamics.

For comparison, we have first computed major structural measures for the two networks, including degree distribution, clustering coefficient distribution, and vulnerability distribution. Second, we run simulations to compare the efficacy of widely accepted public health interventions on the epidemic progress in the three different populations. Our results highlight the importance of the spatial and demographic structure of the social contact network when designing effective interventions. For example, the distribution of school aged children varies widely between the three cities. This difference affects the efficacy of NPIs such as school closures. Structural analysis of the networks provides important cues in this regard. The results have an important implication, namely, guidelines developed by global health organizations, such as WHO, should be evaluated and adapted by each country based on specific demographic and spatial characteristics.

One challenge in generating a realistic synthetic population is that many important statistics or survey data are not available for regions outside the USA. Thus, a careful synthesis of data from various related or unrelated sources is necessary. LandScan data is a useful source for spatial distribution of the population. Ferguson et al. [17] used LandScan data to generate synthetic populations and model influenza transmission in Thailand and in a 100-km wide zone of contiguous neighboring countries. Like our model, their model explicitly incorporates household, schools, and workplaces. Thai census data was used for household size and age distribution. Households are randomly distributed following the density determined by LandScan data. In their model, a person is in contact with anyone visiting the same place; however, in a realistic scenario, people mix in small subgroups especially if the number of persons visiting the same place is large (say, more than 100). In our model, each location is divided into sub-locations and only the persons who are in the same sub-location are in contact. Further, they did not build any explicit edge and contact network, which can lead to loss of some crucial structural properties that can affect disease dynamics. Duration

of the contact between an infectious person and a susceptible person plays an important role in transmitting the disease to the susceptible person, and the duration depends on the activity type. In our model, contact duration for each contact has been generated and taken into account in the simulation of disease transmission; their model ignores individual contact duration.

In [15], Chao et al. also presented an individual-based simulation model of an influenza epidemic, where the individuals are members of social mixing groups, within which influenza is transmitted by random mixing. They divide the entire population into census tracts, which in turn are subdivided into communities of 500-3,000 individuals. The population is organized as a hierarchy of increasingly large but less intimate mixing groups. Workplaces and schools were created following census data. Long distance domestic travel was also considered. However, they also did not construct the explicit contact network, and the contact duration was not considered. Uniform contact probability was used for simulating disease transmission, which was tuned so that attack rates were similar to that of previous known influenza outbreaks such as Asian A (H2N2) and Hong Kong A (H3N2). Even though contact probability was tuned to international influenza outbreaks, population and survey data of the United States were used. A summarized comparison of our model with the models in [17] and [15] is given in Table 1. Here we would like to note that although we created explicit activity schedules for Delhi and Beijing, we created them based on an activity survey done in the United States. As a result, these activity schedules may not reflect real activities of the people in Beijing and Delhi. However, the strength of our model is that once such activity data becomes available, our model can be used more effectively.

Table 1. Comparison of our model with that in [17] and [15]

	Ferguson et al.[17]	Chao et al.[15]	Delhi / Beijing (our model)	Los Angeles (our model)
Explicit edge	no	no	yes	yes
Census data	yes	yes	yes	yes
LandScan data	yes	no	yes	no
Activity survey data	no	no	no	yes
Exact location data used	no	no	no	yes
Explicit activity schedule	no	no	yes	yes
Variable contact duration	no	no	yes	yes
Transmission depends on individual contact duration	no	no	yes	yes
International population	yes	no	yes	no
Individual-based model	yes	yes	yes	yes

Another model for generating a synthetic population along with assigned home and work locations is given in [18]. They built a synthetic population database. Individual level details have been included to support the infectious disease models. Cooley et al. [16] used this database to study the spread of seasonal influenza in North Carolina.

Our contributions. Our main contribution in this paper is two-fold: i) methodology for generating a coarse synthetic population and a social contact network for any international region from very limited census data and LandScan data; ii) a comparison of three different urban regions across the world in terms of structural properties and disease dynamics, and showing the effect of spatial and demographic variations. To the best of our knowledge, it is the first comparison of disease dynamics of three different urban regions across the world using an epidemic simulation methodology.

As we mentioned earlier, lack of necessary data makes it extremely hard to model population and contact networks accurately for many areas around the world. For most areas, such a model of population and a systematic simulation study of epidemic disease does not exist. The previous works in [17, 15] are the most robust and detailed models in the current literature. The discussion above and comparison with these models [17, 15] show the competitiveness of our model. Thus, it is reasonable to say our model and methods are as good as any other existing models.

The rest of the paper is organized as follows: in Section 2, we provide a summarized description of our models for generating synthetic populations and contact networks from limited census data and LandScan data. The details of our model are given in Appendix A and B. In Section 3, we present results of the simulation of disease dynamics in three generated networks and compare them. In addition to the disease dynamics, we also compare the structural properties of these networks. We conclude the paper in Section 4.

2 Generating a Synthetic International Population and Contact Network

In [12] we presented a method to a generate synthetic population and contact network for cities in the United States, including Los Angeles. In this paper, we focus on constructing population and contact networks for international cities Delhi and Beijing. Different types and sources of data lead to different models and methodology for generating synthetic populations and contact networks. Table 2 shows the data along with their sources that are used in construction of Beijing, Delhi, and Los Angeles.

In this section, we provide an overview of our model to generate synthetic populations for international cities using LandScan data and population survey data. The detail description of our model is given in Appendix A and B.

In the LandScan data, the area is divided into a 30 sec. $\times$ 30 sec. latitude/longitude grids. The data contains the population counts for each cell. The counts were apportioned to each cell based on likelihood coefficients which are

Table 2. Data used in construction of Delhi, Beijing, and Los Angeles Network

Network	Data	Source
Delhi	LandScan	Oak Ridge National Lab [5]
	India Census 2001	Government of India [4]
	Delhi school statistics	Delhi Department of Planning [3]
	College/university data	University Grand Commission, India [10]
Beijing	LandScan	Oak Ridge National Lab [5]
	China census data	National Bureau of Statistics of China [6]
	School data	Database Center of China Economy Website [1]
Los Angeles	US census data	US Census Bureau [11]
	Location data	Dun & Bradstreet [2] and NCES [7]
	Activity survey data	National Household Travel Survey [8]
	Street layout	NAVTEQ [9]

based on proximity to roads, slopes, land cover, nighttime lights, and other information. The LandScan data was compiled at Oak Ridge National Lab as a part of Global Population Project [5]. The census data we used consisted of: total population, age and gender distribution of the population, household size distribution, workplace distribution, number of schools of different types, and occupation distributions for different age groups.

First, we determine the cells (of LandScan data) that are within the boundary of the area of interest (e.g., Delhi, Beijing). To find the LandScan cells inside the area of interest, we are given a set of boundary points of a city (Delhi or Beijing). Using Bresenham's fast line drawing algorithm, boundary lines are computed from these given boundary points. Then the inside cells are determined using a flood-fill algorithm. The number of people in each cell in the LandScan data is converted into population density. This density serves as a probability of a household or a workplace being in this particular cell. Households, workplaces, and schools are generated following the distribution obtained from the census data and they are assigned a location using the LandScan density data. A list of households with assigned location ID, size, and location is created from the household size distribution. Similar lists for workplaces and schools are also formed. The total number of household, workplaces, and schools and their sizes are matched with the census data.

Synthetic populations are generated following the census data and each person is assigned a household and a daytime location, which can be a workplace, school, or a household (for persons that stay at home all day, for example, an unemployed person, housewife, etc). A list of persons with assigned ID, age, gender, and marital status is created from the given number of married and unmarried males and females for all age groups, which is obtained from the census data. When assigning people to households, a set of well-defined rules are followed to maintain a reasonable age gap and gender combination to a family; for example, an infant normally cannot live alone in a family of size one. Similar re-

spective rules were followed for assigning daytime locations (workplaces and/or schools) to the generated persons. When assigning a daytime location, distance to the location from home is also considered. A daytime location at distance d is selected randomly with probability following an exponential distribution: $f(x;\lambda) = \lambda e^{-\lambda x}$ where λ is the mean distance.

Next, we generate an activity sequence for each person. Once we have the synthetic population and their activity sequences, we can generate the contact network. An activity sequence is a set of activities, each including at least an activity type, a start time, a duration, and a location. We assume every individual has at most two activity types: home activity and another associated with his daytime location type. We define a person type for each individual based on his daytime location type, and this type is used to determine one's activity sequence. Each activity location is divided into sublocations. The sublocation model is a way of defining interactions among persons who visit the same location doing the same activity at the same time. Each person is assigned to a sublocation (within the activity location) randomly. The activity sequences and the sublocation model define a people-location (PL) bipartite graph, where people and locations are the vertices and there is an edge between a person P and location L if P visits L; time of the visit is a label of the edge. Then the contact network can be constructed from this PL graph. We define there a contact between two persons if they are in the sublocation with the same activity at the same time. As a result we have a contact network, where each person is a node and there is an edge between two persons if they are in contact with each other in some location. The contact network is generated from the activity list along with the assigned sublocations using a previously built simulation tool EpiSimdemics [13].

3 Comparison between Beijing, Delhi, and Los Angeles

In this section, we compare the three cities, first in terms of static properties, such as demographic structure of the populations and network structure of the social infrastructures. We are also interested in dynamic property comparisons. To this end, we simulate epidemic evolutions of an infectious disease on the three populations, with the same settings. We study the difference in the epidemic dynamics and possible sources of the difference. Finally, we use simulations to study how the same public health intervention strategies may have different effectiveness in containing the disease outbreak in three populations.

3.1 Population Demographics

The people in the three cities have significantly different demographics. The major statistics are in Table 3 and we plot the age distribution and household size distribution in Figures 1 and 2. We find that the Delhi population is much younger with much larger household sizes than the other two. Beijing has a smaller fraction of people of preschool or school age.

Table 3. Demographic statistics of three cities

City	Population size	Average age	Average household size	sex ratio (M/F)
Beijing	16,191,340	37.9	2.6	0.99
Delhi	12,905,750	25.6	9.1	1.22
Los Angeles	16,228,759	32.9	3.0	0.97

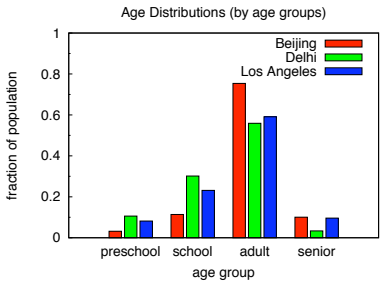

Fig. 1. Comparison of age distributions of Beijing, Delhi, and Los Angeles. Preschool: age 0 to 4; school age: 5 to 18; adult: 19 to 64; senior: above 64. Delhi population is younger than the other two cities.

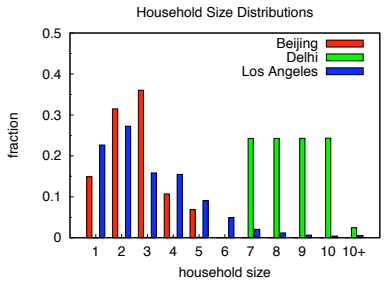

Fig. 2. Compare household size distributions of Beijing, Delhi, and Los Angeles. Delhi families are much larger than those in the other two cities.

3.2 Structural Properties of the Social Contact Networks

In this section, we discuss some structural properties of the three contact networks. Table 4 shows the sizes of the three contact networks: number of nodes, average degree of the nodes, maximum degree and minimum degree. Average degree in the Delhi network is the largest among these three cities and average degree in Los Angeles is the smallest. That is, on average, a person in Delhi gets in contact with more people than a person in Los Angeles. Figure 3 and 4 show degree distribution and clustering coefficient(CC) distribution of the three cities. These distributions show some differences among the cities. While the difference, with respect to degree distribution and CC distribution, between Delhi and Beijing is less, Los Angeles is more different than the other two.

The difference in the network structures of the Beijing, Delhi, and Los Angeles social networks comes from both the demographic difference of the real populations and the methodological difference in generating the synthetic populations and constructing the synthetic social contact networks.

The synthetic populations are generated based on statistical properties of the real populations. The demographic difference discussed in Section 3.1 reflects the real populations, which is a result of fundamental cultural, social, and economical differences, among others, between the urban areas in China, India, and the USA. Network structure is affected by the demographic structure. For example, since school type sublocations generally have larger sizes than work type sublocations, more students in a populations means more connections in the network. Also larger household size means more home type contacts. That explains why the Delhi network has higher average degree than the Beijing network, and both have much higher average degrees than the Los Angeles network.

For Los Angeles, we have detailed US census data, real locations, and sample activity schedules from survey. Our methodology in generating the US synthetic population guarantees that it is statistically indistinguishable from the real population. For Beijing and Delhi, we have only a few coarse statistics and LandScan population density data. We have to adopt another methodology which can only assure that the available distributions are observed. The people and locations in Delhi and Beijing are randomly located. This necessarily affects where people go and who they meet every day, which ultimately affects the structure of the contact networks.

We point out that although we used the same US activity survey data to create activity schedule templates for Beijing and Delhi, due to lack of data, the people in Beijing and Delhi have only home, work, and school type activities, while those in Los Angeles can also have shopping and other activities. Therefore, a person in Los Angeles has more opportunity to mix with other people. On the other hand, we set the sublocation size of Beijing and Delhi schools to 40, in contrast to a sublocation size 25 in Los Angeles schools. These differences also contribute to the structural differences in the three social contact networks.

Table 4. Sizes of the generated contact networks

Network	No. of nodes	Avg. deg.	Max Deg.	Min Deg.
Delhi	12,905,750	79.78	321	1
Beijing	16,191,340	66.77	313	1
Los Angeles	16,228,759	56.60	463	1

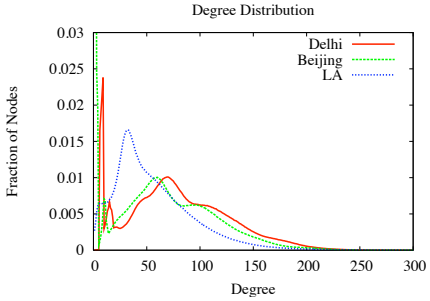

Fig. 3. Degree distributions of the three networks

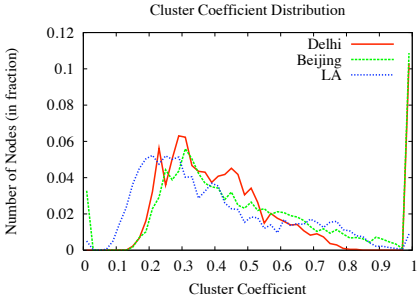

Fig. 4. Distribution of the clustering coefficient of the nodes.

3.3 Epidemic Dynamics and Intervention Efficacy

We compare the dynamics of infectious disease propagations in the three social contact networks and the effectiveness of various intervention strategies using simulations [14]. We assume that the disease is a strong influenza, which infects about 30% (called *attack rate*) of each of the three populations without any interventions. For the simulations in this section, we use a fast epidemic simulation tool that our group has developed, called *EpiFast*. For details about the epidemic simulations, interventions, and EpiFast, please see [14].

In Figure 5, we show the day-by-day average number of infections in the base case for the three cities. Beijing has an earlier outbreak; and Delhi has a higher peak and a shorter outbreak duration. The peak time of Beijing is about two months earlier than that of Los Angeles. This implies that Beijing needs to respond much more promptly in order to prevent a potential outbreak. The faster outbreak of the epidemic in Beijing and Delhi may be due to denser population and stronger mixing of people.

In Figure 6, we plot the distribution of individual vulnerability in three populations. The vulnerability of an individual is the probability of getting infected in a random epidemic. We estimate this measure using 1,000 simulation runs. It seems that Beijing has more people with very low vulnerability or very high vulnerability, while Delhi has more people with medium vulnerability. This implies that interventions targeting the most vulnerable people may have better performance in Beijing than in Delhi.

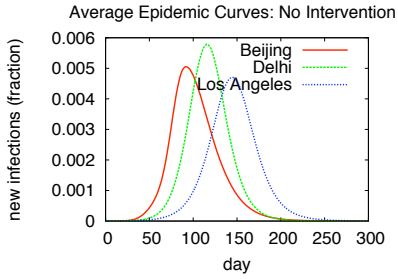

Fig. 5. Base case (no intervention) average epidemic curves.

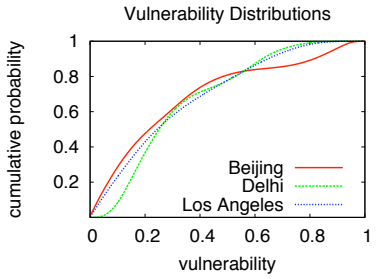

Fig. 6. Cumulative distribution functions of node vulnerability.

Next, we partition each population into subpopulations according to age: preschool, school age, adults, and senior. We plotted the epidemic curve for each subpopulation. This allows us to find out which group of people are particularly vulnerable (i.e., more frequently infected) in each population. By comparing Figures 7 and 8, we find that school age children are most vulnerable in general; but in Beijing their vulnerability is especially high. And the senior age group in Beijing are very resistant to the disease. This explains Beijing's vulnerability distribution shown in Figure 6.

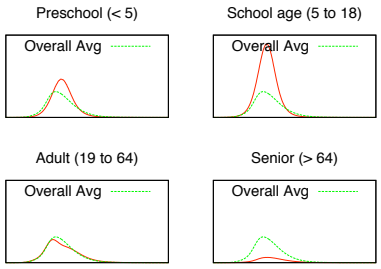

Fig. 7. Base case average epidemic curve of each age group in Beijing.

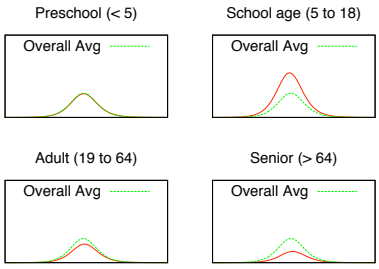

Fig. 8. Base case avg. epidemic curve of each age group in Los Angeles.

Now, we apply different public health intervention measures and compare their effectiveness in different cities. Interventions can be categorized into pharmaceutical interventions (PI's), e.g. vaccination and antiviral administration; and non-pharmaceutical interventions (NPI's), e.g. generic social distancing, school closure, and work closure [14]. PI's reduce the infectivity and vulnerability of the intervened people; while NPI's reduce people-people contact and therefore weaken the connectivity of the contact network. For vaccination we assume the vaccines are enough to cover 25% of the population. For antiviral we

assume 50% coverage. For each NPI strategy, we assume a compliance rate of 50% (i.e., 50% of the population will comply). We assume that vaccines are applied at the beginning of the epidemic; while the other interventions are applied when the infectious people in the population reaches 1%. For each setting, we run 25 replicates and take the average.

The epidemic curves are shown in Figure 9, and Figure 10-15 show the heatmaps of the densities of the infected populations of the three cities on the peak days. We find that vaccination is the most effective. With only 25% of the population vaccinated, the epidemic size decreases by more than 60% for Beijing, 80% for Delhi, and 97% for Los Angeles. Since the Beijing and Delhi social networks are more connected than Los Angeles, 25% vaccination suppresses the epidemic although it is not enough to contain the outbreak completely. In Los Angeles, the outbreak diminishes. For NPI's it seems that school closure is more effective than work closure. Despite the difference in school age subpopulation size and school type sublocation model among three cities, school closure appears to be similarly effective among them. For work closure intervention, its impact on the epidemic evolution is significant in Beijing in the short run, although the efficacy does not last and the outbreak returns.

The policy implication of these simulations is that vaccination seems to be the most effective intervention measure, school closure is the most effective among the NPI's, and that an epidemic is much more urgent for Beijing than for Delhi and Los Angeles.

Finally, we point out that our methodology for constructing Beijing and Delhi synthetic populations and contact networks can naturally make use of data of better quality and more details. As more and more data becomes available, the constructed synthetic population becomes closer and closer to the real one. The difference between such a population and the Los Angeles synthetic population due to methodology difference will diminish and the difference between the real populations will become the dominant source for the different epidemic dynamics presented in simulations. This means our simulations of epidemic evolutions and public health interventions will be more realistic and can provide better decision support for pandemic planning and controlling.

4 Conclusion

We presented a model to generate contact networks for Delhi and Beijing from limited and inadequate census data, and provided simulation results on dynamics of influenza-like disease in these networks. We also compared these results with the Los Angeles network, which has been generated from comparatively rich data sources. These results show the effects of spatial and demographic variation on disease dynamics and contact network structure. For more accurate results, the network methods need substantial improvements with additional details incorporated in the model. However, this model provided us useful insight toward understanding the structure of the contact network formed by the people of various cities and the differences between them caused by demographic differ-

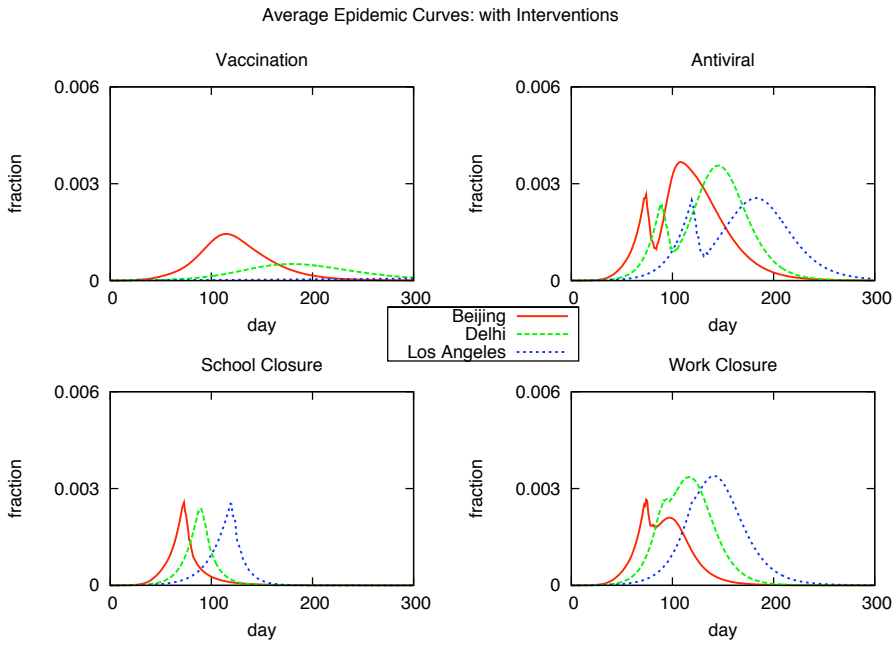

Fig. 9. Compare various PI and NPI strategies.

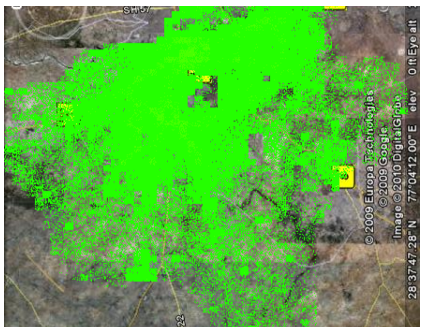

Fig. 10. Delhi – school closure with 25% compliance.

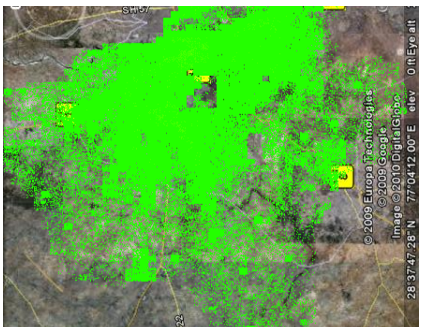

Fig. 11. Delhi – Vaccination with 25% compliance.

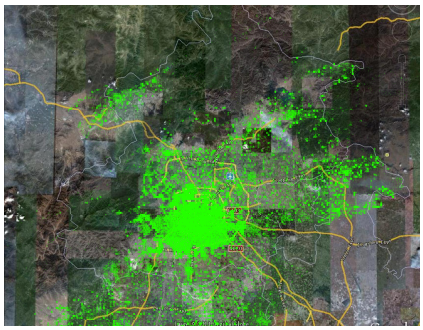

Fig. 12. Beijing – school closure with 25% compliance.

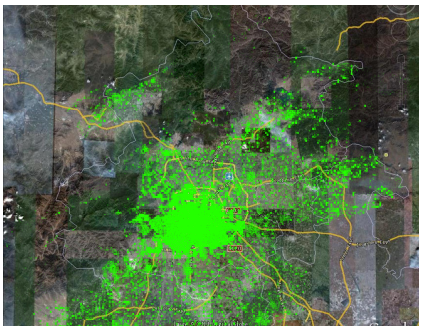

Fig. 13. Beijing – Vaccination with 25% compliance.

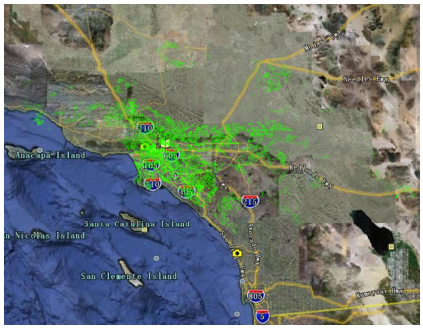

Fig. 14. LosAngeles – school closure with 25% compliance.

Fig. 15. LosAngeles – Vaccination with 25% compliance.

ences, as well as the differences in disease dynamics. Our future work includes continuous refinement of our model as more useful data become available.

Another interesting and important future work on these networks is to analyze sensitivity of the structures of these networks on disease dynamics. One particular experiment is to randomly switch the end points of the edges (shuffling the edges) such that degree of each node remains invariant. Then compare the disease dynamics in this shuffled network with that in the original network. Such an experiment can help us understand how disease dynamics can be affected by the network structures beyond degree distribution.

References

1. Database Center of China Economy Website, http://database.ce.cn/
2. Dun & Bradstreet, http://www.dnb.com/us/
3. Economic Survey of Delhi 2005-2006, Section 15, Delhi Department of Planning, http:/delhiplanning.nic.in/
4. Government of India, Office of the Register General and Census Commissioner, http://censusindia.gov.in/
5. Land Scan Data, Global Population Project at Oak Ridge National Lab, http://www.ornl.gov/sci/landscan/
6. National Bureau of Statistics of China, http://www.stats.gov.cn/english/
7. National Center for Education Statistics, http://nces.ed.gov/
8. National Household Travel Survey, http://nhts.ornl.gov/
9. NAVTEQ Maps and Traffic, http://www.navteq.com/
10. University Grants Commission, India, http://www.ugc.ac.in/
11. U.S. Census Bureau, http://www.census.gov/
12. Barrett, C., Beckman, R., Khan, M., Kumar, V., Marathe, M., Stretz, P., Dutta, T., Lewis, B.: Generation and analysis of large synthetic social contact networks. In: proceedings of the Winter Simulation Conference (WSC) (December 2009)
13. Barrett, C.L., Bisset, K.R., Eubank, S., Feng, X., Marathe, M.V.: Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In: Proceedings of the ACM/IEEE Conference on High Performance Computing (SC). p. 37 (2008)
14. Bisset, K., Chen, J., Feng, X., Kumar, V.A., Marathe, M.: EpiFast: a fast algorithm for large scale realistic epidemic simulations on distributed memory systems. In: Proceedings of the 23rd International Conference on Supercomputing (ICS). pp. 430–439 (2009)
15. Chao, D., Halloran, M., Obenchain, V., Longini, I.: FluTE, a publicly available stochastic influenza epidemic simulation model. PLoS Computational Biology 6(1) (January 2010)
16. Cooley, P., Ganapathi, L., Ghneim, G., Holmberg, S., Wheaton, W., Hollingsworth, C.: Using influenza-like illness data to reconstruct an influenza outbreak. Mathematical and Computer Modeling 48(5-6), 929–39 (2008)
17. Ferguson, N., Cummings, D., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S., Burke, D.: Strategies for containing an emerging influenza pandemic in Southeast Asia. Nature 437, 209–214 (September 2005)
18. Wheaton, W., Cajka, J., Chasteen, B., Wagener, D., Cooley, P., Ganapathi, L., Roberts, D., Allpress, J.: Synthesized Population Databases: A US Geospatial Database for Agent-Based Models. Methods Rep RTI Press 10(905) (May 2009)

A Detailed Method of Generating International Population

In this section, we describe the details of our model to generate synthetic populations using LandScan data and population survey data. In the LandScan data, the area is divided into a grid of cells with the size of each cell being $32'' \times 32''$. The data contains the number of population in each cell. First we determine the cells (of LandScan data) that are within the boundary of the area of interest (e.g., Delhi, Beijing). The number of people in each cell in the LandScan data is converted into population density. This density serves as a probability of a household or a workplace being in this particular cell. Households, workplaces, and schools are generated following the distribution obtained from the census data and they are assigned a location using the LandScan density data. Synthetic populations are generated following the census data and each person is assigned a household and a daytime location, which can a workplace, school, or a household (for persons that stay at home all day, for example, an unemployed person, housewife, etc). Then we generate activity sequence for each person. Once we have the synthetic population and their activity sequence, we can generate the contact network. The details of the methodology are given in the subsequent subsections.

A.1 Finding the LandScan cells inside the area of interest

We are given the cells of LandScan data and the boundary points of a city (Delhi and Beijing). Each boundary point is a pair of latitude and longitude. We want to find the cells inside the boundary. First determine a bounding box as follow.

1. From the boundary points, find the maximum and minimum latitude: max-lat and min-lat. Similarly, find max-long and min-long. These max-lat, min-lat, max-long, and min-long define a bounding rectangle.
2. Let there are $m \times n$ cells in this bounding rectangle. Let (x, y) be the coordinates of a point P in terms of latitude and longitude. We define *cell coordinates* of P to be (r, c) in terms of row and column, where $1 \leq r \leq m$ and $1 \leq c \leq n$.

Then find the cells inside boundary as below.

1. Convert coordinates of the boundary points from (latitude, longitude) to cell coordinates (row,column): if a boundary point (x,y) belongs to cell (r, c), its cell coordinate is (r, c).
2. Next, determine the cells that are on the boundary of the city. For each pair of consecutive boundary points (r_1, c_1) and (r_2, c_2), use Bresenham's line drawing algorithm to determine the cells on this line.
3. Using a flood fill algorithm, find the inside cells.

A.2 Generating household

Given: The household size distribution.
Output: A list of households with assigned Location ID (LID), size, and location.
Procedure:

1. Normalize the household size distribution data to the probability data of each household size. The sum of the probability data is 1.
2. For each household in a size group, assign household identifier (HID) sequentially beginning with 0, randomly assign a household size according to the household size probability. If the size is given as a range, assign the size uniformly randomly from the range.
3. Given a location type ID (TypeID) for household type, assign the LID of each household by LID = TypeID $\times$ 10000000 + HID.
4. For each household, randomly pick a cell within the boundary with the probability from population spatial distribution. Then pick a location within the cell uniformly at random.

A.3 Generating Workplaces

Given: The workplace size distribution; population size N; the daytime-location probability data, i.e., the probabilities that a person will be assigned to workplace, school, or household respectively.
Output: A list of workplace with assigned Location ID (LID), size, and location.
Procedure: The number of workplace can be estimated by integrating information from population size and the probability of daytime location as workplace.

1. Normalize the workplace size distribution data to the probability data of each workplace size. The sum of the probability data is 1.
2. For each workplace in a size group, assign workplace identifier (WID) sequentially beginning with 0, randomly assign a workplace size according to the workplace size probability.
3. Given a location type ID (TypeID) for workplace type, assign the LID of each workplace by LID = TypeID $\times$ 10000000 + WID.
4. For each workplace, randomly pick a cell within the boundary with the probability from population spatial distribution. Then pick a location within the cell uniformly at random.

A.4 Generating Schools

Given: The number of schools of each type, such as kindergarten, elementary school, junior high school, senior high school, college, graduate school, etc.
Output: A list of schools (of different types) with assigned Location ID (LID), size, and location.
Procedure:

1. For each school in each school type, assign school identifier (SID) sequentially beginning with 0, randomly pick a cell within the boundary with the probability from population spatial distribution. Then pick a location within the cell uniformly at random.
2. Given a location type ID (TypeID) for each school type, assign the LID of each school by LID = TypeID $\times$ 10000000 + SID.

A.5 Generating person ID, age, gender, and marital status

Given: Age ranges and for each range, the number of married and non-married males and females.
Output: a list of persons with assigned ID, age, gender, and marital status.
Procedure: For age range $[x, y]$, let M_u, M_m, F_u, and F_m be the numbers of unmarried male, married male, unmarried female, and married female respectively. Then follow the steps below.

1. For each person, assign person identifier (PID) sequentially begin with 0.
2. For a person in age range $[x, y]$, assign age uniformly at random from the range $[x, y]$.
3. For M_u persons in the age range $[x, y]$, assign gender to "male" and marital status to "non-married"; and so on.

A.6 Generating individual persons and assigning people to household

Given: Household, workplace, and school data; the age ranges and for each range, the number of males and females.
Output: A list of persons with assigned ID, age, gender and household
Procedure: Persons are generated household by household, i.e., sequentially generate a list of person for household 1, 2, 3, ... and so on. The number of persons generated for each household should be equal to the household size. When generating persons in each household, the following steps are followed.

1. Given the age and gender data, normalize the data to joint probability distribution data. For age range i, which spans $[x_i, y_i]$, let M_i and F_i be the numbers of male and female respectively. The normalized probability for M_i and F_i will be $M_i / \sum_i (M_i + F_i)$, and $F_i / \sum_i (M_i + F_i)$, respectively.
2. Randomly generate the age and gender for a person according to the age and gender joint probability data. For a person in age range $[x_i, y_i]$, assign age uniformly at random from the range $[x_i, y_i]$.
3. Repeat from Step 2 until all persons in the household are generated.
4. Send the list of persons in this household to a verification function, which justifies whether the generated persons follow a reasonable age gap and gender combination to a family. For example, an infant normally cannot live alone in a family of size one. If the verification function returns **FALSE**, repeat generating person for this family from Step 2; otherwise, continue the following steps.

5. For each person in the list, assign person identifier (PID) sequentially. (The first person in the first household begin the PID with 0.)
6. Repeat generating persons for all the households as above.

A.7 Assigning people to daytime locations

Given: Person data; daytime location data, i.e., workplace, school, and household; for each age, the daytime-location probability data, i.e., the probabilities that a person will be assigned to workplace, school, or household respectively.
Output: A list of persons with assigned daytime location.
Procedure: We assign the daytime location for the person sequentially from PID = 0. For each person, the following steps are followed.

1. Randomly generate the daytime location type L according to the daytime-location probability at the age of this person.
2. Let the location coordinates of the person's household H be (x_H, y_H). Then we can envision a home cell, where H is located, is surrounded by multiple rectangular rings of cells. The four edges of each ring have equal vertical or horizontal distance to the central home cell. Figure 16 illustrates an example of rings, where H is in home cell, cells on ring 1 are at distance 1 to the home cell and cells on ring 2 are at distance 2 to the home cell. Randomly select a ring at distance d, where $d = \lfloor x \rfloor$ and x follows an exponential distribution , i.e.,

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{if } x \geq 0, \\ 0 & \text{if } x < 0, \end{cases}$$

   where λ is the mean distance.
3. Search for the first location with type L on ring d with a clockwise direction. The search begins from the cell that is picked uniformly at random on the ring. If the target location type has a population capacity, then we should find the location not fully filled.
4. Repeat from Step 2 until L is found or a Step 2 has been continuously tried for R times.
5. If L has not been found after R attempts, continuously search L from the inner-most ring to the outer-most ring.
6. Assign the location, found by the above searching, to the person.

B Generating Contact Network

In this section we describe the procedures to create a social contact network for the synthetic population generated in Section A. To create contact network, we need to generate daily activities for each individual in the population. Each activity describes what the person is doing, at which location, from what time, and for how long.

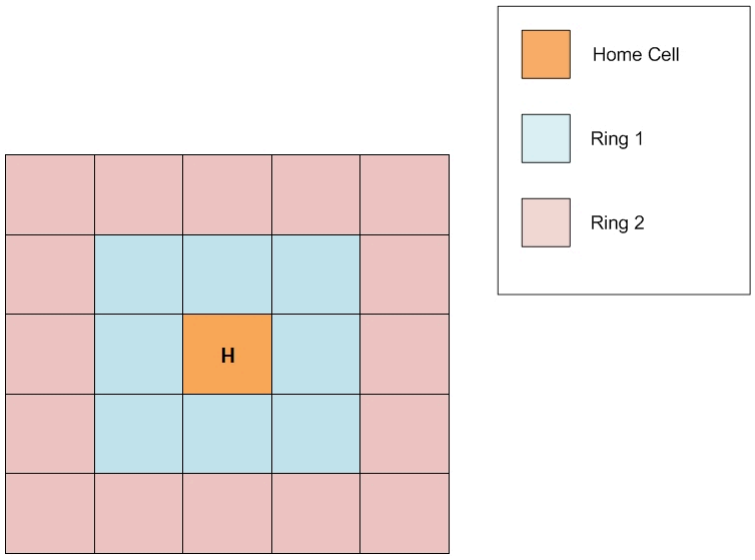

Fig. 16. Ring to the home location.

B.1 Generating Activity File

Here we describe a methodology to generate activities for the synthetic population. The activity lists are required for creating a social contact network for running EpiFast simulations. The list describes an activity sequence for each individual. An activity sequence is a set of activities, each including at least an activity type, a start time, a duration, and a location. Each individual will follow their sequence every day.

We assume that from the previous procedures we have generated for each individual the daytime location and location type, and the nighttime location. The nighttime location is always his home location. The daytime location type includes home, work, or one of several school types.

We assume that every individual has at most two activity types: home activity and another associated with the daytime location type. We define a `PersonType` for each individual based on the daytime location type.

For each `PersonType` we make up several activity sequences as templates, and define a distribution on them. Each activity sequence template includes activity types, start times, and durations for the activities. For now we will only have home activities and day-time activities corresponding to the individual's `PersonType`. We will consider shopping and other activities in the future.

To generate the activity list for each individual, based on the `PersonType`, we randomly choose an activity sequence template according to the distribution. For each activity in the sequence, if it is a home activity we assign home location of this individual as the activity location; otherwise we find the daytime location of this individual and assign it as the activity location.

Using this activity file, contact network is generated following the sublocation model as described below.

B.2 Sublocation Model

The sublocation model is a way of defining interactions among persons who visit the same location at the same time. Each activity location, say L, is divided into sublocations, say $L_1, L_2, \ldots$, and each person visiting location L is randomly assigned a sublocation in this location. Then we define there is a contact between two persons if they are in the sublocation at the same time. The activity list and the capacity of a sublocation for each type of activity location is given as input to the sublocation modeling. The details of the model is given below:

1. A 24-hour day is divided into 15-minute slots. For each location L determine the number of people visiting L for acticity A in each time slot t; let this number be denoted by $N(L, A, t)$.
2. For each A and L, compute $N_{max}(L, A) = \max_t N(L, A, t)$.
3. Let the capacity of sublocation for activity A at location L be $M(L, A)$. For each L and A, generate $N_{max}(L, A)/M(L, A)$ sublocations within location L.
4. For each activity in the activity file, assign a sublocation uniformly at random.

The activity list along with the assigned sublocations are fed to a simulation software EpiSimdemics [13] to generate the contact network. From the given input, EpiSimdemics identifies the persons who are in the same sublocation at the same time and their contact durations. In the resultant contact network, each person is a node, an edge exists between two persons if they are in the same sublocation at the same time, and the contact durations are the edge weights.