Modeling, Analysis and Comparison of Large Scale Social Contact Networks for Epidemic Studies

Huadong Xia

Dissertation submitted to the Faculty of the Virginia Polytechnic Institute and State University in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Computer Science and Applications

Madhav V. Marathe, Chair
Jiangzhuo Chen
Christopher L. Barrett
Henning S. Mortveit
Naren Ramakrishnan

Jan 28, 2014
Blacksburg, Virginia

Keywords: social contact network, networked-epidemiology, network analysis, modeling and simulation, socio-technique systems
Copyright 2014, Huadong Xia
Modeling, Analysis and Comparison of Large Scale Social Contact Networks for Epidemic Studies

Huadong Xia

(ABSTRACT)
Social contact networks represent proximity relationships between individual agents. Such networks are useful in diverse applications, including epidemiology, wireless networking and urban resilience. The vertices of a social contact network represent individual agents (e.g. people). Time varying edges represent time varying proximity relationship. The networks are relational – node and edge labels represent important demographic, spatial and temporal attributes. Synthesizing social contact networks that span large urban regions is challenging for several reasons including: spatial, temporal and relational variety of data sources, noisy and incomplete data, and privacy and confidentiality requirements. Moreover, the synthesized networks differ due to the data and methods used to synthesize them.

This dissertation undertakes a systematic study of synthesizing urban scale social contact networks within the specific application context of computational epidemiology. It is motivated by three important questions: (i) How does one construct a realistic social contact network that is adaptable to different levels of data availability? (ii) How does one compare different versions of the network for a given region, and what are appropriate metrics when comparing the relational networks? (iii) When does a network have adequate structural details for the specific application we have.

We study these questions by synthesizing three social contact networks for Delhi, India. Our case study suggests that we can iteratively improve the quality of a network by adapting to the best data sources available within a framework. The networks differ by the data and the models used. We carry out detailed comparative analyses of the networks. The analysis has three components: (i) structure analysis that compares the structural properties of the networks, (ii) dynamics analysis that compares the epidemic dynamics on these networks and (iii) policy analysis that compares the efficacy of various interventions. We have proposed a framework to systematically analyze how details in networks impact epidemic dynamics over these networks. The results suggest that a combination of multi-level metrics instead of any individual one should be used to compare two networks. We further investigate the sensitivity of these models. The study reveals the details necessary for particular class of control policies. Our methods are entirely general and can be applied to other areas of network science.
Acknowledgments

Chapters 3 - 7 are publications and technical reports authored in collaboration. Here I list the original titles and acknowledge the contribution of each of my co-authors:

Chapter 3: Implications of Demographic and Spatial Variability on Epidemic Study [36, 139]
Jiangzhuo Chen and Fei Huang implemented the Location assignment algorithm. Paula Stretz and Maleq Khan contributed expert knowledge of population synthesis. Madhav Marathe and Jiangzhuo Chen guided the model development and analysis.

Chapter 4: Evaluating strategies for pandemic response in Delhi using realistic social networks [144] and Chapter 5: A High Resolution Realistic Social Contact Network Construction Methodology
Kalyani Nagaraj analyzed Thane survey data and extracted activity templates. Madhav Marathe and Jiangzhuo Chen guided the model development and analysis.

Chapter 6: Computational methods for testing adequacy and quality of massive synthetic proximity social networks [138]
Christopher Barrett contributed expert knowledge of network modeling. Jiangzhuo Chen provided the guidance in the experiment design and model validation. Madhav Marathe guided the whole process of mathematical model development and quantitative analysis of simulation results.

Chapter 7: Comparison and validation of synthetic social contact networks for epidemic modeling [142, 143]
Jiangzhuo Chen provided the guidance in the experiment design and simulation results analysis. Madhav Marathe guided the whole process of metrics design and quantitative analysis of simulation results. Samarth Swarup contributed in results analysis and writing the manuscript.

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 5U01GM070694-11, NSF NetSE Grant CNS-1011769, NSF SDCI Grant OCI-1032677.
In addition, I would like to give a more personal note of gratitude.

I would like to address special thanks to my advisor, Professor Madhav Marathe, for his patience, visionary guidance, enthusiastic support and most of all the encouragement he has given me during my graduate studies. I would like also thank my co-advisor, Dr. Jiangzhao Chen, who worked closely with me on every aspect of this work and gave me helpful advice.

I would like to thank the professors on my committee, Professor Henning Mortveit, Professor Christopher Barrett and Professor Naren Ramakrishnan for their guidance in asking thoughtful questions along with insightful suggestions.

I thank professors and researchers at the Network Dynamics and Simulation Science Laboratory. Specifically, I thank Kalyani Nagaraj, Paula Stretz, Dr. Richard Beckman, Dr. Maleq Khan and Dr. Christopher Kuhlman for collaboration and/or advices with their expertise in different areas. I would like to thank my fellow graduate students, Yifei Ma, Guanghong Pei, Zhao Zhao, and Sichao Wu, for making my stay in NDSSL much more pleasurable.

I especially thank my dad, mom, and sister for their support, understanding and unconditional love during my graduate studies.
## Contents

1 Introduction
   1.1 Background .............................................................................. 1
   1.2 Research Questions and Contributions ...................................... 3
   1.3 Organization of the Dissertation .............................................. 6

2 Literature Review
   2.1 Network Modeling ...................................................................... 8
      2.1.1 Previous Work in Network Modification and Model Validation .... 10
   2.2 Computational Epidemiology Models .......................................... 11
   2.3 Social Contagion Applications: Other Test Signals for Network Structure . 12

3 Delhi-Network-V1: A Generic Network Construction Methodology .... 16
   3.1 Introduction ................................................................................ 17
   3.2 Materials and Methodology ......................................................... 20
      3.2.1 Data .................................................................................... 20
      3.2.2 Generating a Synthetic International Population and Contact Network 21
   3.3 Results ....................................................................................... 23
      3.3.1 Demographics, Mobility Patterns and Social Structures of the Synthetic Populations . 24
      3.3.2 Baseline Epidemic Dynamics .................................................. 26
      3.3.3 Intervention Efficacy .............................................................. 28
   3.4 Discussion .................................................................................. 30
3.4.1 Selection of $R_0$ in Simulations ............................................ 30

A Detailed Method of Generating International Population ................. 32
A.1 Finding the LandScan Cells Inside the Area of Interest ................. 32
A.2 Generating Households ................................................................. 34
A.3 Generating Workplaces ................................................................. 34
A.4 Generating Schools ................................................................. 35
A.5 Generating Person ID, Age, Gender, and Marital Status ................. 35
A.6 Generating Individual Persons and Assigning People to Household ..... 37
A.7 Assigning People to Daytime Locations ........................................ 38
A.8 Generating Contact Network ......................................................... 39
A.8.1 Generating Activity Files ......................................................... 39
A.8.2 Sublocation Model ................................................................. 40
A.8.3 Structural Properties of the Social Contact Networks ................. 41

B Model Validation: Sensitivity Test .............................................. 45
B.1 Method of Location Switch .............................................................. 50
B.1.1 Terminology Definition .............................................................. 50
B.1.2 Location Switch Strategy .............................................................. 51
B.1.3 Generalized Switch ................................................................. 51

4 Delhi-Network-V2: Evaluating Strategies for Pandemic Response in Delhi Using Realistic Social Networks .................................................. 55
4.1 Introduction ........................................................................... 56
4.1.1 Summary of Contributions ......................................................... 57
4.1.2 Related Work ........................................................................... 58
4.2 Data and Methodology: Network Generation ................................... 59
4.2.1 Data Collection ........................................................................... 59
4.2.2 Network Construction Methodology ......................................... 59
4.3 Results ...................................................................................... 67
4.3.1 Demographics and Daily Activity Pattern of the Synthetic Population 68
4.3.2 Graph Structural Properties of the Contact Networks . . . . . . . 69
4.3.3 Epidemic Dynamics and Intervention Policies . . . . . . . . . . . 72
4.3.4 Sensitivity Test for Our Synthetic Network Model . . . . . . . . 78
4.3.5 Validation, Assumptions and Limitations . . . . . . . . . . . . . 78
4.4 Conclusion . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 82

5 Delhi-Network-V3: A High Resolution Realistic Social Contact Network Construction Methodology 83
  5.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 83
  5.2 Data and Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
    5.2.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
    5.2.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 84
  5.3 Results . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 89
    5.3.1 Structure of Synthetic Components . . . . . . . . . . . . . . . . . . . 90
    5.3.2 Network Structure . . . . . . . . . . . . . . . . . . . . . . . . . . . . 92
    5.3.3 Epidemic Dynamics in Delhi . . . . . . . . . . . . . . . . . . . . . . . 94

6 Network Improvement: Computational Methods for Testing Adequacy and Quality of Massive Synthetic Proximity Social Networks 97
  6.1 Introduction . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 98
    6.1.1 Our Contributions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 99
    6.1.2 Related Work . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 100
  6.2 Network Generation: Data and Methodology . . . . . . . . . . . . . . . . . 100
    6.2.1 Data . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 101
    6.2.2 Methods . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 102
  6.3 Comparison of Static Structure and Epidemic Dynamics of Three Networks . 106
    6.3.1 Measuring the Features of Synthetic Components . . . . . . . . . . . . . 107
    6.3.2 Network Structure Comparison . . . . . . . . . . . . . . . . . . . . . . 113
    6.3.3 Uncertainty Reduction during Network Improvement . . . . . . . . . . . 116
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.3.4</td>
<td>Comparison in Network Dynamics and Effective Policy</td>
<td>119</td>
</tr>
<tr>
<td>6.3.5</td>
<td>Summary and Observations</td>
<td>121</td>
</tr>
<tr>
<td>6.4</td>
<td>Conclusion</td>
<td>123</td>
</tr>
<tr>
<td>7</td>
<td>Comparison and Validation of Synthetic Social Contact Networks across Geographical Regions</td>
<td>124</td>
</tr>
<tr>
<td>7.1</td>
<td>Introduction</td>
<td>125</td>
</tr>
<tr>
<td>7.2</td>
<td>Network Generation: Data and Methodology</td>
<td>127</td>
</tr>
<tr>
<td>7.2.1</td>
<td>Data</td>
<td>127</td>
</tr>
<tr>
<td>7.2.2</td>
<td>Methods</td>
<td>127</td>
</tr>
<tr>
<td>7.3</td>
<td>Comparison of the Two Cities</td>
<td>130</td>
</tr>
<tr>
<td>7.3.1</td>
<td>Population-level Comparison</td>
<td>130</td>
</tr>
<tr>
<td>7.3.2</td>
<td>Network Structure Comparison</td>
<td>133</td>
</tr>
<tr>
<td>7.3.3</td>
<td>Comparison of Epidemic Dynamics</td>
<td>135</td>
</tr>
<tr>
<td>7.4</td>
<td>Validation</td>
<td>137</td>
</tr>
<tr>
<td>7.5</td>
<td>Conclusion</td>
<td>139</td>
</tr>
<tr>
<td>8</td>
<td>Concluding Remarks</td>
<td>140</td>
</tr>
<tr>
<td>8</td>
<td>Bibliography</td>
<td>142</td>
</tr>
</tbody>
</table>
List of Figures

1.1 An example of a social contact network. ........................................ 2
1.2 The synthetic social contact network generation pipeline. ................. 4

2.1 Generic diffusion processes. ..................................................... 15

3.1 Heat map for Delhi population on a typical day. .......................... 22
3.2 Comparison of age distributions of Beijing and Delhi. ..................... 25
3.3 Comparison of household size distributions of Beijing and Delhi. ........ 25
3.4 Comparison of activity counts of residents in Beijing and Delhi. .......... 25
3.5 Comparison of travel distances of Beijing and Delhi. ..................... 25
3.6 Epicurves for base case (no intervention) when $R_0 = 1.35$. ............. 26
3.7 Base case average epicurve of each age group in Beijing when $R_0 = 1.35$. 27
3.8 Base case average epicurve of each age group in Delhi when $R_0 = 1.35$. 27
3.9 Distribution of node vulnerability ($R_0 = 1.35$). .......................... 28
3.10 The efficacy of various PI and NPI strategies to H1N1 in different cities. 29
3.11 In the table are $R_0$ values estimated for 2009 H1N1 for different countries. 31
3.12 Attack rate for different $R_0$. Multiple scenarios in the figure imply the comparison results are insensitive to $R_0$ value. .......................... 31

A.1 Ring to the home location. ....................................................... 39
A.2 Degree distributions of the two networks. .................................... 42
A.3 Contact duration distributions of the two networks. ........................ 42
A.4 Distribution of the clustering co-efficient of the nodes. .................... 42
4.15 Epidemic curves show subpopulation epidemics in Delhi-Network-V2. . . . . 76
4.16 Epidemics under various intervention strategies in Delhi-Network-V2. . . . . 77
4.17 A test of various combinations of compliances and trigger thresholds for the TLC strategy. ................................. 77
4.18 Epidemics and policy efficacy in Delhi-Network-V2 with various sublocation sizes ($R_0 = 1.35$). ................................. 79
4.19 Network structure perturbations for the Delhi Network ($G_P$) . ............. 80
4.20 Impact of location switches to Delhi-Network-V2. ................................. 81

5.1 Density heatmap of Delhi after we allocate people to their daytime locations. 88
5.2 Daily travel distance distribution for people in Delhi within a 24-hour-period. 91
5.3 Radius of gyration distribution for people in Delhi within a 24-hour-period. . 91
5.4 Temporal degree for randomly selected locations across a typical day. . ...... 91
5.5 Degree distribution of $G_{PL}$. The plot can also be understood as the visiting counts in all non-home locations. ................................. 92
5.6 Basic network structural properties of the people-people social contact network of Delhi, $G_p$. ................................. 93
5.7 Counts of labeled graphlets. ................................. ................................. 94
5.8 Epidemics under various intervention strategies in the Delhi network. . . . . 95
5.9 Subpopulation epidemics in Delhi-Network-V3. ................................. 96

6.1 Household structure comparison. ................................. 108
6.2 The layout of synthetic locations in the three networks. ................................. 109
6.3 Duration of several typical activities: home, work and school. ................................. 110
6.4 The temporal degree of important POIs. ................................. 111
6.5 Age mixing for different versions of Delhi synthetic populations. ................................. 112
6.6 The travel distance and radius of gyration for people in Delhi within a 24-hour-period. ................................. 113
6.7 Degree distribution for people-location networks. ................................. 114
6.8 Structural analysis of people-people contact networks. ................................. 115
6.9 Visualization for all 3-node and 4-node graphlets. ................................. 115
6.10 Distribution of featured graphlets. .................................................... 116
6.11 Distribution of of labeled graphlets. .................................................... 116
6.12 Basic switch operation. ................................................................. 118
6.13 Epidemic dynamics under various level of edge switches. .................... 118
6.14 Normalized epidemic dynamics under various level of edge switches. ...... 119
6.15 Vulnerability histogram in Delhi-Network-V1, Delhi-Network-V2 and Delhi-
Network-V3. ................................................................................. 120
6.16 Effects of vaccinating different subpopulations. .................................... 121

7.1 HH structure distribution of Los Angeles and Delhi. ............................ 128
7.2 Temporal degree of important POIs. .................................................... 132
7.3 Distribution of travel distance and radius of gyration. .............................. 133
7.4 Degree distribution for people-location networks. ................................. 133
7.5 Structural analysis of people-people contact networks. ......................... 134
7.6 Labeled graphlet distribution. ............................................................. 135
7.7 Vulnerability of Los Angeles and Delhi. .............................................. 136
7.8 Epidemics and policy efficacy in the Los Angeles network and Delhi-
Network-V2. ............................................................................. 137
7.9 Subpopulation epicurves of Los Angeles and Delhi. .............................. 138
List of Tables

1.1 Selected network metrics. .................................................. 7
3.1 Comparison of our model with several other models in the literature. .... 19
3.2 Data used in construction of Delhi and Beijing Networks. .................. 20
3.3 Demographic statistics and geo-location of two cities. ...................... 23
3.4 Profiles of social contact structures in the two cities. ....................... 24
3.5 Epidemic dynamics and the efficacy of various interventions. .............. 27
A.1 Sizes of the generated contact networks. ........................................ 41
4.1 Demographic statistics of Delhi in comparison with two other cities. .... 59
4.2 Demographics, location, and activity data used in the construction of Delhi-Network-V2. ......................................................... 60
4.3 Average structure properties of several city-scale contact networks. .... 71
4.4 Results of Targeted-Layered Containment in Delhi when $R_0$ is 1.35. .... 77
5.1 Demographics, location and activity data used in the construction of Delhi-Network-V3. .......................................................... 85
6.1 Demographic statistics and geo-location of Delhi. ............................ 101
6.2 Demographics, location and activity data used in construction of Delhi-Network-V1. 101
6.3 Demographics, location and activity data used in construction of Delhi-Network-V2. 102
6.4 Demographics, location and activity data used in construction of Delhi-Network-V3. .......................................................... 103
6.5 Data Size of the Output Synthetic Structures (in bytes). ..................... 103
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>Comparison of our analysis approaches to traditional approaches.</td>
<td>107</td>
</tr>
<tr>
<td>6.7</td>
<td>Measurements used in characterizing the static structure and dynamics.</td>
<td>122</td>
</tr>
<tr>
<td>7.1</td>
<td>Demographic statistics and geo-location of Delhi and Los Angeles.</td>
<td>125</td>
</tr>
<tr>
<td>7.2</td>
<td>Demographics, location and activity data used in construction of Delhi-Network-V2.</td>
<td>126</td>
</tr>
<tr>
<td>7.3</td>
<td>Demographics, location and activity data used in construction of the Los Angeles network.</td>
<td>126</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

1.1 Background

For decades, networks have been exploited to represent proximity interaction between individual agents. Such networks are called social contact networks, and have extensive applications in many areas, including epidemiology [15, 55, 83, 93, 136], sociology [6, 26, 85], economy [123], wireless networking [34], and urban resilience [31]. The vertices of a social contact network represent individual agents (e.g. people or locations). The edges represent time-varying proximity relationships. The network node and edge labels represent important demographic, spatial and urban attributes. We show an example in Figure 1.1, where the social network is represented by a people-location bipartite graph $G_{PL}$ and a people-people contact graph $G'_{P}$. Both $G_{PL}$ and $G'_{P}$ contain node and edge labels conveying rich information.

Such a social contact network usually cannot be obtained through direct surveys, because (i) most people do not know all the people they come into contact with during a day, and (ii) we cannot collect all the information available due to privacy and confidentiality.

Therefore, we seek to develop synthetic social contact networks. The word “synthetic” is used in two ways here. First, the network is not “real” but “statistically similar” to an idealized real social contact network for an urban region. We also use “synthetic” to mean that the network is synthesized by combining a variety of data sources in conjunction with mathematical models that abstract accepted social and behavioral theories. These theories tell us who, what, why and when of human movement and interactions in an urban (rural) region. A social contact network is constructed based on the specific dynamical process that we wish to study by assigning an edge between two individuals when they satisfy a desired proximity criteria. For example, proximity in case of Influenza would be different than when we study Smallpox or Ebola.
Figure 1.1: An example of a social contact network. It includes the people-location bipartite graph $G_{PL}$ and the people-people contact graph $G'_{P}$. In $G_{PL}$, a circle represents a person and a square represents a location, whereas an edge between a person and a location means that the person visits the location during the labeled time interval. The people-people graph $G_{P}$ is inferred directly from $G_{PL}$, and $G'_{P}$ is formed by combining edges from all subgraphs of $G_{P}$. In $G_{P}$ and $G'_{P}$, an edge $(u, v)$ with label $(t_1, t_2]$ represents temporal information of the contact between $u$ and $v$.

To this end, we start from collective census data and survey samples to create a disaggregated synthetic population, and then simulate people’s activities to estimate proximity contacts among them. One such process is shown in Figure 1.2, which contains several steps as follows. First, we create a synthetic population for the interested geographic region. The synthetic population is a set of synthetic people and households. Each synthetic individual belongs to a synthetic household, and is associated with realistic demographic variables recorded in the census, including gender and age. Such a synthetic population preserves the confidentiality of the individuals in the original data sets, yet produces realistic demographic attributes which collectively are statistically indifferent from the census. Second, each synthetic person in a household is assigned a set of activities to perform during a typical day, along with the times when the activities start and end. Those activities are generated based on time use survey data. Third, we assign an appropriate location for each activity and for each synthetic person. The locational data are retrieved from data sources such as land use patterns, tax data or commercial location databases. Locations are assigned to a person using a gravity model based on the distance between the locations and the person’s household address. Finally, we are ready to estimate proximity relationships. Each synthetic person is deemed to have made contact with a subset of other synthetic people simultaneously present at a location.

We emphasize that, the collection of data and the construction of synthetic social structures in each step is not fixed. We may have various ways to retrieve the goal in each step, so the
above framework is generic.

Following this generic framework, synthesizing social contact networks that span large urban regions is still challenging for several reasons including the spatial, temporal and relational variety of data sources, and data that are noisy and incomplete. We focus specifically the following issues:

- A variety of spatial, temporal and relational data sources are used in network synthesizing. Specific methods are required to process those irregular data.
- Data are often unavailable.
- The synthesized networks for the same population differ by the data and methods utilized to synthesize them. This may cause inconsistency within the same context.

This dissertation is focused on overcoming the above challenges and exploring large scale people-people interaction pattern in the real world. We are trying to address two main topics in this dissertation.

- A generic network construction methodology adapted for different data availability.
- The improvement of synthesized networks and the evaluation standards.

1.2 Research Questions and Contributions

To address the challenges proposed in last section, this dissertation undertakes a systematic study of synthesizing urban scale social contact networks within the specific application context of computational epidemiology. It is motivated by three important questions:

- How does one construct a realistic social contact network that is adaptable to different levels of data availability?
- How does one compare different networks for a given region, and what are appropriate metrics when comparing the relational networks?
- When does a network have adequate structural details for the specific application we have?

We study these questions by devising three sets of complex models in synthesizing the social contact network of Delhi, India. The first set of models result in what we call Delhi-Network-V1. It uses minimal data sources that capture the spatial and demographic features of
Figure 1.2: The synthetic social contact network generation pipeline. The figure is borrowed and modified from Madhav et al. [88]. The first two blocks represent the construction of synthetic population. The third block (in purple) represents the assignment of activities and activity locations to each synthetic person. The last block (in green) is about estimating proximity relationship based on people’s co-existence relation.
the underlying region. The second set of models result in Delhi-Network-V2. The models use a detailed layout of the built infrastructure as well as normative surveys to capture the movement patterns of individuals in a country such as India. These models are more intensive in terms of computing and data resources than the first set of models. However, Delhi-Network-V2 makes use of the activity survey from another India city, Thane, as a substitute for Delhi. This may cause mis-interpretation of mobility pattern for residents in Delhi. Our third set of models result in Delhi-Network-V3. It uses higher quality data targeted for Delhi, as well as incremental improvement in the methodology aspect. The third set of models appear to capture certain topological features better than the other two. Delhi represents a typical region in developing countries with poor data availability. Compared to US or European countries, data related to human mobility or land use are limited or coarser in these areas, which brings a major challenge in constructing a realistic network. Our effort in the three sets of models addresses the challenges of how to construct a realistic social contact network with different levels of data availability (especially when data is limited). Our results also suggest that we can iteratively improve the quality of a network by adapting to the best data sources available within the framework.

Based on the work of the three sets of iteratively improved models, we further explore the questions on network comparison and modification. Since no model is ever completely correct, the model modification and improvement in practice is a process of incorporating details into the model through iterative and varied use. However, we are faced with a challenge to evaluate the effects of network modification, given several urban scale networks. In this case, how does one capture the similarities and differences between them? The size, irregularity and the time varying nature of these networks create this challenging analytical question. The comparison is not simply a matter of comparing epicurves against real data, since it is very easy to match a sequence of numbers of infections by tweaking part of a large number of parameters. Our approach here, therefore, is to do a detailed comparison of structural and dynamical metrics on these networks to determine the differences between them. We propose a number of metrics, divided into four classes:

1. Metrics that capture the features of the population, built infrastructure and their layout,
2. Network level metrics that capture the structural features of the dynamic social contact network,
3. Metrics that capture the epidemic dynamics over the networks,
4. Policy metrics that capture the effect of controls.

We argue that these metrics matter to the network construction model because it pertains to several dimensions of network quality, including accuracy, precision, and impact on simulated epidemics. Table 1.1 lists how these metrics are related to the different quality dimensions. Some of the metrics are well studied in the literature by network scientists, although they
have not been studied in our specific context. Other metrics are new and inherently context specific; they are used to capture the specific aspects of social contact networks and to study epidemics. Computational considerations are important; we would like methods that can efficiently compute these metrics either exactly or approximately (if the problem is intractable). The structure and size of the networks make developing such methods challenging. Using these metrics, we systematically compare the three versions of the Delhi network. Our analysis reveals interesting differences between the networks, even though the networks are supposed to capture the interactions in the same urban region.

We further investigate the robustness of these methods and the algorithm sensitivity of the system. Specifically, we study the sensitivity of simulated epidemic outcomes to the algorithm of sublocation division and location assignment. Here sublocation means the typical connected group size within a location. Location assignment is the algorithm to assign people to different locations based on their home locations and daytime activities. Traditionally, a gravity model is used to describe the most likely distance between a person’s home and his other activity locations. The sensitivity study shows that the sublocation division is critical in forming the contact network structure and its dynamics. However, the epidemic outcome is quite robust to the location assignment algorithm. The analysis reveals the kinds of network details necessary for particular class of control policies.

1.3 Organization of the Dissertation

The dissertation is organized as follows. In Chapter 2, we conduct a literature review in the areas of network modeling and computational epidemiology. In the next three chapters, we introduce three social contact network construction methodologies. Chapter 3 proposes the generic yet coarse construction methodology referred as NDSSL-V1. We applied the NDSSL-V1 methodology to construct networks for two developing country urban areas, Beijing in China and Delhi in India. After that, we describe the improvement methodology NDSSL-V2 and NDSSL-V3 in Chapter 4 and 5. We argue that the networks are improved along the three versions of construction methodologies. To demonstrate this, we propose a systematic network comparison method, which involves a number of novel metrics in various levels. This is discussed in Chapter 6. The similar comparison techniques are used to compare networks constructed for different regions in Chapter 7.
Table 1.1: Selected network metrics. The metrics are selected with the consideration that they will measure the network and associated detailed structure in several quality dimensions: accuracy, precision and/or effects on epidemics. The candidate pool is massive and this list is necessarily limited.

<table>
<thead>
<tr>
<th>Categories</th>
<th>Metrics</th>
<th>Accuracy</th>
<th>Precision</th>
<th>Effect on Epidemics</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Entity Level &amp; Collective Level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Household Structure</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Location Layout</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Duration of Activities</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Number of Daily Activities</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Travel Distance</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Radius of Gyration</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Network Level</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$G_{PL}$</td>
<td>Temporal Degree of Locations</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>$G_{PL}$</td>
<td>Degree of $G_{PL}$</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td>$G_P$</td>
<td>Degree, Clustering Coefficient,</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Contact Duration, Shortest Path</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Graphlets Decomposition</td>
<td>-</td>
<td>-</td>
<td>✓</td>
</tr>
<tr>
<td></td>
<td>Network Entropy</td>
<td>-</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td><strong>Epidemic Dynamics</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pharmaceutical Interventions</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Non-Pharmaceutical Interventions</td>
<td>-</td>
<td>-</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>
Chapter 2

Literature Review

In this chapter, we conduct preliminary literature reviews for several fields related to the topic of this dissertation. First, we list major work in social contact network modeling. We then continue to present work of network modification and improvement in a more generic background (complex networks). We have focused our discussion of network modeling on computational epidemiology throughout the thesis. However, our work is generic enough to adapt to many other applications as well. We therefore introduce a variety of network applications in sections 2.2 and 2.3.

2.1 Network Modeling

As mentioned previously, we introduce three sets of network models here. The set of network models for Delhi-Network-V1 is based on very limited data but applicable to all national or sub-national region in the world, we refer here as NDSSL-V1. The network models used to construct Delhi-Network-V2 and Delhi-Network-V3 use more region specific data as well as region specific methods, which we refer to as NDSSL-V2 and NDSSL-V3. Please note that NDSSL-V3 has been used to generate networks for multiple US cities and is deemed as a set of mature models [52], while NDSSL-V1 [36] and NDSSL-V2 [144] are newly devised models. Among the three sets of models, NDSSL-V1 is aimed at generic networks, while NDSSL-V3 is aimed at producing best quality networks. We compare the two sets of models to other models in the literature.

Network models (or individual-based models) are required to rebuild micro properties for individual agents, rather than aggregate properties. However, most network models only make use of aggregate level summary data such as distributions of age and gender, and constrain those input data to disaggregated properties for individual agents with necessary assumptions. Only a few models take microdata as well as aggregated data as input. The two examples among them are NDSSL-V3 and the RTI [136] synthesized population database.
for USA. Both of the two models generate synthesized population based on the TRNSIMS population generator [125] or its variants. The idea is to collect household samples, replicate them to generate microdata records for all of the synthetic households using a statistical technique known as iterative proportional fitting [18]. The location generation algorithms for the two models are also very similar for work places, schools, etc. The two sets of models differ in several aspects. First, RTI models create residents of group quarters (i.e., college dormitories, prisons, military barracks, nursing homes and others) with a different algorithm. Second, RTI models assign people to different types of locations with different algorithms. They assign students to a closest public school that has capacity for students of that grade level, and assign students to private schools based on a gravity-like model. Work place assignment is based on the worker counts of each combination of census tract residence and census tract of work. Also, they create public transit systems and assign people to each bus/train based on distance and geospatial transit routes information. NDSSL-V3, on the other hand, uses a gravity model based on a physical distance value to assign people to all different locations. RTI is more realistic in this regard. Third, NDSSL-V3 creates explicit activity schedules for each person and creates the sublocation model to infer contacts between two arbitrary people in the same location. This type of information is missing in the RTI models and therefore no explicit network is built in the RTI models.

Both NDSSL-V3 and RTI models were originally designed for the US population and aimed at a high quality network because rich data are available for the US and European countries. However, in modeling synthetic population and networks in developing countries, data are often limited and a careful synthesis of data from various related or unrelated sources is necessary. The LandScan data set is a useful source for spatial distribution of the population and has been used in NDSSL-V1 to build the networks for Delhi and Beijing. Ferguson et al. [55] also uses 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 of NDSSL-V1, their model explicitly incorporates household, schools, and workplaces. The Thailand census data are 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 (e.g., more than 100). In NDSSL-V1, each location is divided into sub-locations and only the persons who are in the same sub-location are in contact. Furthermore, they did not build any explicit edges or a contact network, and therefore, some crucial structural properties that may affect disease dynamics are unavailable for analysis. 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 contact type. In our model, contact duration for each contact has been generated and taken into account in the simulation of disease transmission, whereas their model ignores individual contact duration.

Longini et al. [35] also present 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 is also considered. However, they did not construct the explicit contact network either, and the contact duration is not considered. Instead, uniform contact probability is used for simulating disease transmission, which is tuned so that attack rates are similar to that of previous known influenza outbreaks such as Asian A (H2N2) and Hong Kong A (H3N2). A similar model was also developed by the same group to study a 75\textit{km} × 75\textit{km} rural area in Thailand [83], where they consider subgroups in a location, similar to the concept of sublocation in our models.

Meta-population networks are a specific type of network models. These types of models are “mainly focused on the role of human mobility and involve a very accurate knowledge of the mobility fluxes between different sub-populations (transportation infrastructures, long-range airline connections, short range daily commuting pattern, etc.)” [81]. A node in such a network represents a subpopulation and the interactions within a subpopulation are modeled in other forms, such as compartments. Therefore, the meta-population network models are hybrid models and are typically applied to model specific scenarios very different from the above individual based models. The examples of meta-population networks include work from Vespignani group [39,40], Glass [66], and etc.

### 2.1.1 Previous Work in Network Modification and Model Validation

Much effort is being invested to develop more complex, more finely-grained computational models in many branches of science, coming along with developments in computing infrastructure and advances in quantitative experimental measuring techniques. Network improvement, on the other hand, is a relatively unexplored area. One pioneer work is from Eubank et al. [51] where they studied the sensitivity of network dynamics to different model input. The paper also proposes two criteria in evaluating the quality of the improved network: precision and accuracy. Precision means the variability in possible outcomes, i.e., how focused are the resulting networks. Accuracy measures how close the generated network is to the real world instance set.

Those measures are related to the general questions of validation and verification in mathematical and computer models. The belief is that each detail added renders the models more complicated and more difficult to calibrate [51,112], but also more faithful to the real world instance. The adequacy of tests is often the key criteria, meaning “No single test which serves to ‘validate’ a system dynamics model” (Forrester [58]). Many validation and verification methods are explored in this area [11,58,75]. Some methods in this area are utilized in our work to measure the quality of modified network, such as face validation (subject matter
methods) and docking test (quantitative methods).

In the area of social contact networks, measurements done in the last few years have suggested that certain network properties seem to be true for social networks. Some work in this area can also be used and compared to our work [80, 120, 121, 132]. Examples include the gravity model that describes people’s general mobility patterns [80, 132]; Barabasi and Marta et al. [62, 121] use a limited number of network motifs to represent most typical human mobility graphs.

2.2 Computational Epidemiology Models

Computational epidemiology studies the progress of an epidemic in a large population. In order to model such a procedure, we represent people’s health state with the following four states: Susceptible, Exposed, Infectious, and Recovered. An infectious disease is assumed to transmit between hosts (people) from an infectious host to a susceptible host. Diseases progress within a host following a certain epidemic model which describes a serial of progressive stages. For example, the SEIR model consider the health states flow within a host as follows:

\[ S \rightarrow E \rightarrow I \rightarrow R \]

Similarly, we have other epidemic models, including the SIR model, the SIS model, the SIRS model and the SEIS model. The development of computational epidemiology models is built upon certain epidemic models and has experienced several stages.

Traditionally, mathematical and computational modeling of epidemics has focused on aggregate models using coupled rate equations [9]. 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 public health policy of 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 [13, 94]. 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, 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 [35, 56, 57, 61, 113]. 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 [4, 38, 92]. Epidemic dynamics within a county are computed using an individual-based model. The dynamics over a network of counties are captured using coupled rate equations.

The final class of models use the most realistic representation of social contact networks [16, 50, 93]. In [16, 21, 22, 50] each individual in the United States is modeled with detailed demographic profiles and daily activities. Our study is based on this class of models, and we will give a detailed introduction to models in this sub-area in the next two sections.

2.3 Social Contagion Applications: Other Test Signals for Network Structure

Here we use epidemic diffusion processes to help compare and contrast the constructed networks. Epidemics propagating among the networked population works as a kind of test signal to explore the internal structure of the networks. However, epidemic diffusion is only one example of these kinds of test signals. We list a number of different fields to provide an overview of diffusion processes on networks. We point out that the networked model is suitable to model a set of generic reaction diffusion processes. The following diffusion processes can be deemed as sampled from those generic processes: the spread of ideas, fads, gossip, attitudes, and memes. We call these diffusion processes social contagion applications. Borgatti [27] and Kuhlman [76] et al. have reviewed a multitude of social contagion applications. We introduce some of them below.

Gossip. Gossip is typically shared between two persons or in a small group. It spreads by replication and does not pass the same link twice (meaning one does not tell the same story to the same person) [27]. This is similar to influenza and can be simulated with a similar model.

E-mail. Imagine an E-mail message that warns of an electronic virus within the network of an organization. It spreads through direct contacts and becomes pervasive on the intranet. The message may be passed to the same person multiple times, but it does not stimulate that person to forward the old message for the second time.

Emotional contagions. Emotions include happiness, fear, disgust, anger, and sadness. Studies [47, 68] have shown that human emotions are transferable among a crowd of people. Similarly, attitudes can also be spread from person to person [27]. The notion is that
individuals effect each other’s emotions or attitudes through interaction. Experiments on different types of emotions have been discussed in the research community, including fear and anger [95], anger and sadness [23], and stress between spouses [25]. Compared to influenza, emotions or attitudes may not remain stable and do not last a long period. People are not immune to the same emotions and may be susceptible to them for many times.

Financial contagions. These include contagions for financial instabilities, downturns and subsequent rebounds. These contagions are different from the ones introduced previously, because they often happen at multiple levels, including the company, institution, or country level, as well as the individual level [76]. For example, at the country level, the financial crisis in 1997 in East Asia was well known to spread across countries as a domino effect [145]. Similarly, the subprime mortgage crisis in 2007 started from the U.S. and reached worldwide in a short time [78]. At the individual level, financial contagion is found to affect investors [123]. A hybrid model is often used to simulate such a multi-level social contagion process.

Multiple contagions. Many studies have modeled diffusion processes as a single homogeneous flow, including our work in this dissertation. However, multiple interacting contagion processes are frequently observed. For example, the spreads of viruses and fear are often intertwined and co-evolved. Similarly, we may observe multiple contagions in the advertising of two or more competing products in a market. For example, Coke and Pepsi contend to attract consumers to use one company’s products over the other. Multiple contagions can be modeled as simultaneous contagions [48, 64, 71, 89], in addition to the interplay between those individual contagions.

Social media and online data. Many studies have shown that peer influence is ubiquitous among online social populations [7, 8, 26, 65, 77, 111, 119] through social media websites. Examples include tens of millions of Facebook users [127], whose behaviors are shown being affected by peers. Also, the mass protests in Spain are interpreted as results of group interactions through Twitter [63]. The behavior of online communities can often be used to understand people’s offline behavior. The benefits using the social media data are two-fold. First, the evolution of social media and online communities has helped researchers to simulate diffusion processes at unprecedented scales. Second, the online data is free of subjective errors occurred within the traditional data collection.

Most applications introduced above can be modeled with epidemiology models and implemented in our simulation system, either directly or with minor modifications. Gossip and E-mail are typical SEIR network flow. Emotional contagions basically follow the SI model. Financial contagions can be simulated with the hybrid model described in Section 2.2, where transmission of financial behaviors occurs at two levels: the country level where contagions spread across nations; and the individual level within each country. Multiple contagions can be properly simulated with multiple intertwined basic diffusion processes on a single-layer or multi-layer network. The interplay between multiple contagions could be implemented by careful design. For example, with the dual contagions of virus and fear, the spread of
fear will change the behavior of individual agents and consequently affect the spread of the virus. *Social media and online data* are inherently suitable for our simulation system, as they provide well defined quantitative measurement for social networks. Diffusion processes on online networks resemble those in the real populations in most of the above applications.

Please be advised that although these diffusion processes differ in diffusion behavior and specific parameters, most of them can be easily implemented and studied in a manner similar to the way we simulate epidemic dynamics. We select three typical diffusion processes to showcase the generality of our modeling, namely: DP1, DP2, and DP3. For simplicity but without losing generality, we assume all three diffusion processes follow the *SEIR* model. DP2 represents a typical influenza-like disease. It has an incubation phase distribution (1:0.3; 2:0.5; 3:0.2), meaning that 30% of the infected population has an incubation phase of 1 day, 50% with 2 days, and 20% with 3 days. It also has an infectious phase distribution (3:0.3; 4:0.4; 5:0.2; 6:0.1). DP1 represents a rumor-like information diffusion. Correspondingly it has a shorter infectious phase in general, with distribution (1:0.3; 1:0.4; 3:0.2; 4:0.1), and shorter incubation phase (0:0.5; 1:0.3; 2:0.2). DP3 is tentatively designed as another type of diffusion process dissimilar to DP2. It has longer infectious and incubation phases, with distributions (5:0.3; 6:0.4; 7:0.2; 8:0.1) and (2:0.3; 3:0.2; 4:0.5) respectively. We simulate those three diffusion processes with our EpiFast platform under various parameter settings (Figure 2.1).

To summarize, the epidemiology application utilized in our simulation can be replaced by many other social contagion applications. Therefore, methodologies we present in this dissertation can be used in many other areas to study different diffusion processes and build different types of networks.
Figure 2.1: Generic diffusion processes. We simulate three typical diffusion processes DP1, DP2, and DP3 on the same social contact network representing the population in Delhi urban area. Various parameters are explored, including transmissibility ("tau" in the figure), meaning the probability an infection will happen within a unit time between two neighbors on the network; and the asymptomatic probability of an infected person, range from 0.2, 0.4 to 0.6 (an asymptomatic person is assumed not infectious). The attack rate (the fraction of infected people after an epidemic outbreak) under different situation for the three processes are shown in the figure.
Chapter 3

Delhi-Network-V1: A Generic Network Construction Methodology

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

Abstract

Demographics and spatial variability have strong implications on epidemics. The same disease might impact different populations in very different ways due to those variations. However, comparison studies of epidemics and intervention strategies for populations across different regions were scarce if not none in the literature. One challenge is the unavailability of a generic methodology to model different populations with very different data input. In this work, we develop a novel network construction methodology. The methodology makes use of landScan data and very basic census information, and is open to incorporate region specific data, thus it is applicable to most areas in the world. We use the new method to construct social contact networks and study epidemics in two international cities: Beijing, China and Delhi, India. The contact networks in the two cities are different due to 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). We show that demographic and cultural differences in different regions greatly impact the epidemic dynamics. The results have an important implication that guidelines developed by global health organizations, such as WHO, should be evaluated and adapted by each country based on its own demographic and spatial characteristics.
3.1 Introduction

Demographics and spatial variability have strong implications on epidemics [36,142,144]. The same disease might impact different populations in very different ways due to those variations. For example, epidemic intensity and fatality rate for SARS in 2003 varied significantly across different regions in the world [60]. Consequently, effective pharmaceutical as well as non-pharmaceutical interventions (PI & NPI) differ geographically. However, little work has been proposed to model and predict such differences across different populations internationally.

One major challenge for the comparison study lies in the data availability for international populations. Influenza like diseases transmit among a population through people’s interactions. Those interactions form so-called social contact networks. To retrieve the social contact network for a region, we should model where people are during the course of the whole day. For this purpose, multiple sources of data are required, including census data, land use data and activity information local to the residents. A methodology to synthesize social contact networks for the US cities is already in place [15]. However, the data required in the method are very limited in developing countries, making network construction a difficult task in those areas. A method to deal with coarse level data is therefore required.

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. To construct the contact network, we explicitly generate an activity sequence for each individual in the population taking into account the variability in demographic and activity patterns for each city individually.

Based on the method, we construct and study social contact networks for two cities: Beijing, China and Delhi, India, with only the LandScan population density data and limited census data. The constructed networks are a representation of social structures in the two areas. Therefore, the two networks are different due to the fact that the two populations inherently have very different demographic structures and activity patterns.

With the networks ready, we could conduct comparisons of epidemics in the two regions, and study the impact of demographic and spatial variability. Our first set of analyses focus on comparing the two populations in terms of their demographics and activity patterns. The major structural measures for the two city networks, including degree distribution and clustering coefficient distribution, are also computed and analyzed. Second, we run epidemic simulations to compare the efficacy of widely accepted public health interventions on the epidemic progress in the two different populations. Our results highlight the importance of the spatial and demographic attributes of the social structures when designing effective interventions. For example, the fraction of school aged children varies greatly between the two cities. This difference affects the efficacy of NPIs such as school closures. Structural analysis of the social structures provides 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 its own demographic and spatial characteristics.

Our main contribution in this paper is three-fold: (i) a 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 two different urban regions across countries 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 different urban regions across the world using an epidemic simulation methodology. (iii) a methodology to inspect the robustness of the network construction model, which is presented in the supporting material.

Related Work. The modeling of the social contact network for a city is a way to simulate the complex social system. A careful synthesis of data from various related sources is necessary. LandScan data is a useful source for spatial distribution of the population. Ferguson et al. [55] 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 contact type. In our model, contact duration for each contact has been generated and taken into account in the simulation of disease transmission, whereas their model ignores individual contact duration.

In [35], 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.
Table 3.1: Comparison of our model with several other models in the literature.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Explicit edge</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Census data</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
<tr>
<td>LandScan data</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Activity survey data</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Exact location data used</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
<td>no</td>
</tr>
<tr>
<td>Explicit activity schedule</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Heterogeneous contact duration</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Transmission depends on individual contact duration</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>International population</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
<td>yes</td>
<td>no</td>
<td>yes</td>
</tr>
<tr>
<td>Individual-based model</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
<td>yes</td>
</tr>
</tbody>
</table>

Other models for generating a synthetic population along with assigned home and work locations are developed in [92, 118, 136]. They built a synthetic population database. Individual level details have been included to support the infectious disease models. Specifically, RTI [136] set up an online synthetic population viewer [135] for their synthesized US population database, providing an interactive map of a “synthetic population”. Each dot on the map represents a single household. The map helps us to visually see how age, income, race and household size vary in cities. However, such a high resolution synthesized population is typically not available in developing countries due to the data availability. In addition, similar to the above two models, the explicit network edges are not modeled and activity schedules are not used in these projects, which is often required in studying epidemics and polices quantitatively.

A summarized comparison of our model with the models in [35, 55, 92, 118] is given in Table 3.1. Here we would like to note that although we created explicit activity schedules for Delhi and Beijing, we created them based on some normative assumptions. 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.

As we mentioned earlier, lack of necessary data makes it extremely hard to model populations and contact networks accurately for many areas around the world. For most areas, such a model and a systematic simulation study of disease epidemics do not exist. The previous works in [35, 55] are the most robust and detailed models in the current literature. The discussion above and comparison with these models [35, 55] show the competitiveness of our model. Thus, it is reasonable to say our model and methods are at least as good as any other existing models.
3.2 Materials and Methodology

We describe a generic methodology in constructing populations and contact networks applicable to any regions in the world. As examples, we apply it to two international cities Delhi and Beijing.

3.2.1 Data

Table 3.2 shows the data along with their sources that are used in the constructions of the contact networks for Beijing and Delhi. Those data are very basic and we assume they can be obtained while modeling contact networks in most area in the world.

Table 3.2: Data used in construction of Delhi and Beijing Networks.

<table>
<thead>
<tr>
<th>Network</th>
<th>Data</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>India Census 2001</td>
<td>Government of India [110]</td>
</tr>
<tr>
<td></td>
<td>LandScan</td>
<td>Oak Ridge National Lab [30]</td>
</tr>
<tr>
<td></td>
<td>Delhi school statistics</td>
<td>Delhi Department of Planning [43]</td>
</tr>
<tr>
<td></td>
<td>Delhi school schedules</td>
<td>Delhi school websites [44, 117, 122]</td>
</tr>
<tr>
<td></td>
<td>College/university data</td>
<td>University Grand Commission, India [129]</td>
</tr>
<tr>
<td>Beijing</td>
<td>China census data</td>
<td>National Bureau of Statistics of China [100]</td>
</tr>
<tr>
<td></td>
<td>LandScan</td>
<td>Oak Ridge National Lab [30]</td>
</tr>
<tr>
<td></td>
<td>School data</td>
<td>Database Center of China Economy Website [42]</td>
</tr>
<tr>
<td></td>
<td>Activity statistics</td>
<td>China time use survey 2008 [97, 99]</td>
</tr>
</tbody>
</table>

LandScan is a global scale spatial population database and we use it as a substitute for real location information. There are several spatial population databases that meet our needs [82], including GPW, GRUMP, UNEP and LandScan. A comprehensive survey to these databases can be found in [82]. We decide to choose LandScan because it provides finest precision. In the LandScan data, the area is divided into a 30 sec. × 30 sec. latitude/longitude grids. The data contains the population count in each cell. The counts were apportioned to each cell based on likelihood coefficients which are 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 [30].

The census data we used consisted of: total population size, age and gender distribution, household size distribution, workplace distribution, number of schools of different types, and occupation distributions for different age groups. Those are most basic information and should be presented in a typical national census data set. More region specific data such as local school schedules [42, 44, 117, 122] and activity statistics [97, 99], as shown in Table 3.2, could be easily incorporated into the model as a improvement.
Generating a Synthetic International Population and Contact Network

In the following, we provide an overview of our model to generate synthetic populations for international cities using LandScan data and population survey data. The detailed description of our model is given in Appendix A and A.8.

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 (the daytime location can be household for a person who stays at home all day, e.g., 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 respective 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 $\frac{1}{\lambda}$ is the mean distance. Such a distance model is called gravity model [15].

Contacts among the population emerge from co-located relationships, i.e., two persons are in contact only when they both stay in a location at the same time. The contacts in locations are generated due to the collective effect of each individual’s stochastic activities during the course of the day. In our model, we create realistic disaggregated activity sequences for each individual to model such collective stochastic behavior. An activity sequence is a set of major activities for a person to take on a typical day, each activity includes at least an activity type, a start time, a duration, and a location. To retrieve the most representative activity sequence for each individual, the method in [15] makes use of the National Household Travel Survey (NHTS) [109] in USA. The survey collected the demographics of 169894 people and
their typical activities in a day. Those activity sequences are built into a classification tree, and they are assigned to each synthetic individual according to the person’s demographics based on the built classification tree. Unfortunately, similar survey data for China and India are not available when the paper was written. We did find some aggregate statistics in the literature though, including average time for school activities [44, 97, 117, 122] and work activities [96, 97, 99] for the two countries. We therefore compose simple yet reasonable activity sequences for the two cities, and calibrate each activity’s duration based on the statistics we collected. We assume every individual has at most two activity types: home activity and another activity associated with their daytime location. We define a person type for each individual based on their daytime location, and this type is used to determine one’s activity sequence. Note that the day time locations in the activity sequence is very similar to the concept of “mixing group” in Chao’s [35] and “places” in Ferguson’s [55]. Two types of mixing-groups/places are considered in their papers, including schools/daycare and work places. Our selection of daytime location types is similar to theirs in this regard. In Figure 3.1, we illustrate the heat map for Delhi population density after we distribute all synthetic people to their daytime locations. The population density distribution is consistent with our observation from other sources [138] where the population is distributed more sparsely as the physical distance is farther from the city center.

Once we have the activity sequence of each individual, we are ready to generate the contact network. The activity sequences and the associated locations define a people-location (PL)
Table 3.3: Demographic statistics and geo-location of two cities.

<table>
<thead>
<tr>
<th>City</th>
<th>population size</th>
<th>average age</th>
<th>household size</th>
<th>sex ratio (M/F)</th>
<th>area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>16,191,340</td>
<td>37.9</td>
<td>2.6</td>
<td>0.99</td>
<td>6,487 sq miles</td>
</tr>
<tr>
<td>Delhi</td>
<td>13,850,507</td>
<td>25.6</td>
<td>5.14</td>
<td>1.22</td>
<td>573 sq miles</td>
</tr>
</tbody>
</table>

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 assume each activity location is divided into sublocations. The sublocation model is a way of defining interactions among persons who visit the same location at the same time. Each person is assigned to a sublocation (within the activity location) randomly. We define a contact between two persons if they are present in the same sublocation 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. The contact network is generated from the activity list along with the assigned sublocations using a previously built simulation tool EpiSimdemics [16]. The formation of social contact structure is a stochastic process; therefore, the constructed network is only a sample of all possible contact networks in the real world. Nevertheless, the properties of the networks are stable across the network spaces, and the retrieved network sample is a good representative. We present a sensitivity test of the network construction method in the supporting material supporting this claim. The supporting material also contains the details of our model.

3.3 Results

We created two synthetic populations as well as associated social contact networks for Beijing and Delhi, based on methods discussed above. Because demographic structures and underlying social contact networks influence the spread of infectious diseases in an urban region, analysis of the structural properties of them yields insights into the disease propagation. To reveal the different epidemic dynamics in the two cities, we compare the two populations at various levels. First in terms of static properties, such as demographic structure of the populations and network structure of the social infrastructures. To compare the dynamic properties, we use a Monte Carlo method to simulate epidemic evolutions of an infectious disease on the two populations. We study the difference in the epidemic dynamics and possible sources of the difference, which show that how the same public health intervention strategies may have different effectiveness in containing the disease outbreak in the two cities.
Table 3.4: Profiles of social contact structures in the two cities. Here, edge weight means contact duration in seconds. Clustering coefficient is a measure of the degree to which nodes in a network tend to cluster together, it can be measured as the number of closed triplets in the network divided by the number of connected triplets of vertices.

<table>
<thead>
<tr>
<th>Network</th>
<th>average degree</th>
<th>average edge weight</th>
<th>average clustering coefficient</th>
<th>average shortest path</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>67</td>
<td>8022 seconds</td>
<td>0.50</td>
<td>6.68 hops</td>
</tr>
<tr>
<td>Delhi</td>
<td>76</td>
<td>9757 seconds</td>
<td>0.48</td>
<td>5.30 hops</td>
</tr>
</tbody>
</table>

3.3.1 Demographics, Mobility Patterns and Social Structures of the Synthetic Populations

The people in the two cities have significantly different demographics. The major statistics are in Table 3.3 and we plot the age distribution and household size distribution in Figures 3.2 and 3.3. We find that the Delhi population is generally younger, while Beijing has a smaller fraction of people of preschool or school age. Also, Delhi has higher fractions of large size households, including some up to 10+ family members, most households in Beijing on the other hand are smaller than 6 members.

The counts of activities across the day are shown in Figure 3.4. The distribution of different types of activities are consistent with the first principle observations, where home activities peak at night and other activities peak during the day. Compared to Delhi, Beijing has more people stay in a work location in anytime of the day, and less in schools. The differences are the combination of local people’s activity schedules as well as the demographic structure (Figure 3.3 and 3.3). There are a higher fraction of adults in Beijing thus more people are employed and stay at a work location.

Population movement patterns of the residents in the two cities also reveal varieties. In Figure 3.5, we plot the distribution of travel distances in the two cities in log-log scale. The two cities have a similar shape in the distribution. The majority of the trips in both cities are less than 10 km. For trips longer than 20 km, both cities have a power-law-like distribution. However, the distance values differ in the two, where all people in Delhi have a daily travel distance shorter than 50 miles and a number of people in Beijing travel longer than 50 miles. We believe this is partly due to the different areas of the two cities (refer to city borders in Table 3.3).

The above difference leads to different social contact structures. In Table 3.4, we list the profiles of the social contact networks of the two cities. A significant difference is the average degree of the two city networks. The average degree of Beijing network is 10 less than that of Delhi, and it has higher clustering coefficients and longer average shortest path, revealing a structure hindering the spread of an influenza like disease. For further details please refer to the supporting material.
Figure 3.2: Comparison of age distributions of Beijing and Delhi. Preschool: age 0 to 4; school age: 5 to 18; adult: 19 to 64; senior: above 64. Delhi population is younger than that of Beijing.

Figure 3.3: Comparison of household size distributions of Beijing and Delhi. Delhi families are much larger than those in Beijing.

Figure 3.4: Comparison of activity counts of residents in Beijing and Delhi.

Figure 3.5: Comparison of travel distances of Beijing and Delhi.
3.3.2 Baseline Epidemic Dynamics

The structural differences of the populations and contact networks have a strong implication for the epidemic dynamics in the two cities. To demonstrate this, we compare the dynamics of infectious disease propagations in the two social contact networks and the effectiveness of various intervention strategies using simulations [21]. We consider the 2009 H1N1 influenza, since it is the first global epidemic in the 21st century, and more information about the 2009 H1N1 is known; for example, the $R_0$ values in different cities in the world are known. Here we assume $R_0$ values in both cities are 1.35 (please refer to the discussion section for the selection of $R_0$). For simulation of disease dynamics, we use a fast epidemic simulation tool, called EpiFast [21], developed in NDSSL at Virginia Tech.

Due to the disparate contact structures of the two cities, the same type of disease may cause quite different consequences in each city. In Figure 3.6, we show the day-by-day average number of infections in the base case for the two cities with the assumption that $R_0 = 1.35$. The difference is significant. Under the same $R_0$ assumption, the network of Beijing has a much lower peak value but earlier outbreak (earlier peak day). Its attack rate is also slightly lower than that in Delhi (refer to Table 3.5 and Figure 3.10). We point out that the results are the consequence of different contact structures in the two cities. As shown in Table 3.4, Delhi has higher number of contacts (degree) and a less clustered structure (lower clustering coefficients), therefore a contact structure facilitating the diffusion of epidemics compared to Beijing.

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 shows which group of people are particularly vulnerable (i.e., more frequently infected) in each population. By comparing Figure 3.7 and 3.8, we find that the epidemic impact to subpopulations is quite consistent in the two cities, where school age children are most vulnerable; and seniors are least vulnerable in general. What differs is the subpopulations of adults in
Table 3.5: Epidemic dynamics and the efficacy of various interventions. We simulate influ-like
diseases and assume $R_0 = 1.35$. (AR, peak, peak day) are three crucial factors in characterizing an
epidemic outbreak, where AR is the acronym for “attack rate”. An attack rate is the cumulative
incidence of infection in a group of people observed over a period of time during an epidemic; peak
day is the day when the maximum number of new infections appears; peak is the fraction of new
infected people during the peak day.

<table>
<thead>
<tr>
<th>policy</th>
<th>Beijing</th>
<th>Delhi</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>AR</td>
<td>peak</td>
</tr>
<tr>
<td>no intervention</td>
<td>mean</td>
<td>0.451091</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>0.000269</td>
</tr>
<tr>
<td>vaccination</td>
<td>mean</td>
<td>0.230742</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>0.000399</td>
</tr>
<tr>
<td>antiviral</td>
<td>mean</td>
<td>0.439919</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>0.000598</td>
</tr>
<tr>
<td>school closure</td>
<td>mean</td>
<td>0.443036</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>0.000604</td>
</tr>
<tr>
<td>work closure</td>
<td>mean</td>
<td>0.437979</td>
</tr>
<tr>
<td></td>
<td>s.d.</td>
<td>0.000623</td>
</tr>
</tbody>
</table>

Beijing. The epicurve for adults in Beijing fits closely to the overall epicurve, while that for
Delhi is much lower than the overall epicurves, probably because of the dominant proportion
of adults in Beijing. Obviously, the different composition of demographic subgroups may
suggest a different way to contain the disease diffusion in the population.

![Figure 3.7: Base case average epicurve of each age group in Beijing when $R_0 = 1.35$.](image)

![Figure 3.8: Base case average epicurve of each age group in Delhi when $R_0 = 1.35$.](image)

In addition to the subgroup behavior, we also look into the difference of networks in indi-
vidual level based on vulnerability values. The vulnerability is the probability that a given
individual gets infected in a potential epidemic threat, and it is largely decided by the indi-
vidual’s position in the network. We estimate the vulnerabilities using 1,000 simulation
runs, an individual’s vulnerability is calculated as the fraction of simulations out of the 1,000
runs within which he/she get exposed. In Figure 3.9, we plot the distribution of individual vulnerability. Compared to Delhi, Beijing has a fraction of people with very low vulnerability (close to zero), and a fraction of people with very high vulnerability (close to 1). Specifically, seniors are least vulnerable people in both populations due to their low social behavior. The low vulnerability people in Beijing is likely to be seniors, given that the population sizes in the two groups are also similar.

![Figure 3.9: Distribution of node vulnerability ($R_0 = 1.35$).](image)

### 3.3.3 Intervention Efficacy

The epidemic difference in the two cities caused by their network structure differences has a strong implication for public health policy. The epicurve for Delhi has a higher peak and attack rate, Beijing has a little bit earlier yet mild (relatively) outbreak. As a result, Both Beijing and Delhi need to respond very quickly to the disease and Delhi may have to prepare for a more severe impact. In understanding the way to contain the epidemics in such large populations, we investigate two questions: 1) what is the difference in efficacy of the same intervention strategy in different regions? 2) what are the most effective intervention strategy given that multiple intervention strategies are available.

To answer these questions, we further conduct experiments to simulate the efficacy of different intervention policies. We apply different public health intervention measures and compare their effectiveness in different cities. Interventions can be categorized into pharmaceutical interventions (PIs), e.g., vaccination and antiviral administration; and non-pharmaceutical interventions (NPIs), e.g., generic social distancing, school closure, and work closure [21]. PIs reduce the infectivity and vulnerability of the intervened people; while NPIs reduce people-people contact and therefore weaken the connectivity of the contact network. For vaccination we assume the vaccine supply is sufficient 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 number of
The efficacy of various PI and NPI strategies to H1N1 in different cities. The PI interventions include antiviral and vaccination. The NPI intervention strategies include school closure and work closure. (1) attack rate: While vaccination is most effective among all strategies, it is slightly less effective to Beijing than to Delhi. (2) peak value: Work closure is better or equally effective in reducing the peak value than school closure in Beijing, probably due to the relative small ration of school population in the city. For Delhi, there seems no favorable incline between the two strategies. (3) peak day: Vaccination is still the most effective strategy in delaying the outbreak of the disease. In Delhi, school closure is much effective than work closure, while in Beijing, the two strategies seem similar.

The epidemic differences are shown in Figures 3.10. We find that vaccination is the most effective intervention. With only 25% of the population vaccinated, the attack rate decreases by more than 50% for all three cities. The other three interventions do not greatly reduce the attack rate but do help to either postpone the outbreak of disease or reduce the peak value (Figure 3.10). Generally, school closure and antiviral are the most effective to delay the outbreak of disease. School closure is especially effective for Delhi because Delhi has a largest proportion of school students.

Due to the different contact and demographic structure observed above, we hypothesize the effective public health policy might be different in the two cities. This is also verified in Figure 3.10. For Delhi, school closure is more effective than work closure in postponing the disease outbreak. However, school closure is not very helpful in Beijing, probably due to its small proportion of school population (Figure 3.2). Instead, work closure is more effective in reducing the attack rate and peak value, which is consistent with what we observed in the age-group based epicurve plots (Figure 3.7).

The policy implications of the epidemic simulations are that vaccination seems to be the most effective intervention measure, school closure and antiviral are generally the most effective to delay the outbreak of disease. From the analysis, we can see that the same policy might
have different efficacy in different cities. Importantly, we show it is possible to determine effective intervention strategies for a specific place based on demographic data and social structures, etc.

3.4 Discussion

We propose a generic methodology that construct social contact networks for a given region with basic census information and very limited activities statistics. The model resolution is a trade-off to generality here. Nevertheless, we point out that our methodology for constructing Beijing and Delhi synthetic populations and contact networks can naturally make use of higher quality and more detailed data. As more and more data becomes available, the constructed synthetic population can more closely resemble the real population. This means our simulations of epidemic spreading and public health interventions will be more realistic and can provide better decision support for pandemic planning and control.

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 relationships between social structure and epidemics. The epidemic simulations suggest that demographic and cultural differences in Beijing and Delhi impact the epidemic outcome as well as the efficacy of public health policies. Therefore, the public health policies designed for different countries should account for these differences. The public health policies by the WHO typically do not take these factors into account and are based on models for western countries or other places with data availability. Clearly, the situations could be improved by analysis of the data specific to the country studied.

Two main contributions of the paper is the network construction model as well as the study to public health policies based on epidemic simulations. To test the robustness of the methodologies used in these two topics, we have conducted sensitivity tests on related model parameters. Our first set of sensitivity test analyze the sublocation size and location assignment algorithm during network construction. The details and results can be found in the supporting material B. Our second set of sensitivity test examine the disease parameter $R_0$ during simulations, as discussed below.

3.4.1 Selection of $R_0$ in Simulations

We need to establish a consistent context for comparison of the epidemics in different cities. To this end, a disease calibration technique is often applied in the literature [67] so that the simulated disease has similar $R_0$ in different models. In this section, we also calibrate the transmissibility of H1N1 so that the simulated H1N1 has similar $R_0$ in the two cities. However, there is not a unique “precise” value for the H1N1 $R_0$, even though there is extensive research to explore the $R_0$ of the disease. This is not only because the value of $R_0$ is affected
by demographics and even the environment of the targeted population, but also because the methodologies used to measure them may produce errors of multiple aspects. Therefore, the $R_0$ of H1N1 we get from the literature varies from city to city and from model to model [106]. Some typical $R_0$ values are listed in Figure 3.11, which is representative $R_0$ estimations we found in the literature [74, 91, 146, 147] for the respective cities for the H1N1 disease. However, those $R_0$ values are actually over estimated [106]. Therefore, when we design the simulations, we choose simulated $R_0$ from the lower end of the estimated range: 1.35, 1.40, 1.45 and 1.60, and conduct all experiments in the paper with these four different $R_0$ values. All the results show that same conclusions apply for all of the four different $R_0$ values. As two examples, we plot the attack rates with two scenarios in Results section in Figure 3.12: the base case when no intervention is conducted and the case when school closure is applied. The relation between the attack rates of the two cities are insensitive to $R_0$ change in both scenarios. This suggests that the epidemic dynamic differences revealed in the simulations are independent of the disease parameters, but is a reflection of underlying network structure. To save space, we list only the results when $R_0$ is 1.35 in the paper, and omit the results when $R_0$ equals to 1.40, 1.45, 1.60 respectively. However, please note that the conclusions we draw when $R_0 = 1.35$ is applicable to all other three $R_0$ values.

<table>
<thead>
<tr>
<th>Country</th>
<th>Reproduction Number $R_0$</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>China</td>
<td>1.6809</td>
<td>[74]</td>
</tr>
<tr>
<td>India</td>
<td>1.45</td>
<td>[91]</td>
</tr>
<tr>
<td>Japan</td>
<td>2.0-2.6</td>
<td>[105]</td>
</tr>
<tr>
<td>USA</td>
<td>1.6 (range: 1.3-1.7)</td>
<td>[146]</td>
</tr>
<tr>
<td>Canada</td>
<td>1.31</td>
<td>[126]</td>
</tr>
<tr>
<td>Mexico</td>
<td>1.2, 1.4 to 1.6</td>
<td>[59]</td>
</tr>
</tbody>
</table>

Figure 3.11: In the table are $R_0$ values estimated for 2009 H1N1 for different countries. In 2011, Nishiura et al. [106] suggest that those $R_0$ values are over estimated, therefore we select a set of $R_0$ values from the lower end of the estimated range (1.35-1.60) and use them to calibrate the epidemic parameters in our study.

Figure 3.12: Attack rate for different $R_0$. Multiple scenarios in the figure imply the comparison results are insensitive to $R_0$ value.
Appendix 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 be 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 Algorithm 1.

We then find the cells inside boundary using Algorithm 2.
Algorithm 1: Find the bounding box in the area of interest.

**Input**: A list of boundary points; cells of LandScan data.

**Output**: A bounding box of LandScan cells that surrounds the area of interested.

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 \( 1 \leq c \leq n \).

Algorithm 2: Find the set of LandScan cells as a minimum coverage of the area of interest.

**Input**: the bounding box of LandScan cells; cells of LandScan data.

**Output**: The set of LandScan cells that overlaps with the area of interested.

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 Households

To generate households with a specified size and location, we make use of two inputs: the household size distribution and a set of LandScan grid cells within the area of interest. The description of the algorithm is in Algorithm 3.

The algorithm may produce errors around the boundary of the region. Due to the size of LandScan cells, the set of cells we retrieved from Algorithm 2 usually is not a perfect coverage of the area of interest, cells on the border may contain area outside of the region. Therefore, we may pick a location outside of the area of interest in Step 4. This is also a problem when we generate other places such as work places or schools in the following sections.

**Algorithm 3:** Generating Households

**Input:** The household size distribution; a set of LandScan grid cells within the area of interest.

**Output:** A list of households $h_i$ with assigned Location ID ($lid_i$), size $n_i$, and location $l_i$.

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

Generating workplaces requires a variety of data, including: 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. The description of the process is listed in Algorithm 4.
Algorithm 4: Generating Workplaces

**Input:** 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.

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 $\text{LID} = \text{TypeID} \times 10000000 + \text{WID}$.

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

The generation of schools is similar to that of workplace. In addition, there are several different types of schools, including kindergarten, elementary school, high school, and college. Each type of schools is generated separately but following the same pipeline. The details are described in Algorithm 5.

A.5 Generating Person ID, Age, Gender, and Marital Status

After all locations are generated and assigned with appropriate coordinates, we begin to create the synthetic population. Each synthetic person will be generated with an ID, age, gender, and marital status. The algorithm is listed in Algorithm 6. The input of the algorithm is basic statistics from census.
Algorithm 5: Generating Schools

**Input:** 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.

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 $LID = TypeID \times 10000000 + SID$.

Algorithm 6: Generating Individuals with Demographic Information

**Input:** 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.

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 $M_x$ 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

Algorithm 6 is applied to generated each person one by one. Actually, 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, Algorithm 7 is designed to replace Algorithm 6.

Algorithm 7: Generating Populations

**Input:** 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

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. 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.

In Step 4 we use a verification function as a filter to drop non-realistic households. For example, an infant normally cannot live alone in a family of size one, and the verification function will return **FALSE** for this kind of family. A direct impact of using the verification function is, it impose a selection procedure upon synthesizing individual people. The demographic distribution of the synthetic population may be skewed from the census statistics due to the selection. It seems more natural to construct the population according to distributions and then group them into households. This seems a better solution but it may cause some errors too, because the population cannot guarantee a perfect partition of households.
A.7 Assigning People to Daytime Locations

In addition to assigning people to households, we also need to assign them to locations for their daytime activities (e.g., going to school or work). The algorithm is described in Algorithm 8.

**Algorithm 8:** Assigning Daytime Locations

**Input:** 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.

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. 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 \( \lambda \) 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.

Figure A.1 illustrates an example of rings referred to in Step 2, 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. This assignment algorithm has some computational advantages, but has
shortcomings too. First, the 8 cells along the rectangular ring do not have same distance from the home cell. Second, a less evident error is that the cell itself usually is not a square on larger latitudes. The longitudinal distance of LandScan cells scale with \( \cos(\text{latitude}) \). In practice, this means that people travel a shorter East/West distance as you move away from the Equator.

\[ \text{Figure A.1: Ring to the home location.} \]

### A.8 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.

#### A.8.1 Generating Activity Files

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 epidemic 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 the home location. The daytime location type can be either 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 type, start time, and end time for activities in the sequence. 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.

**A.8.2 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 activity $A$ in each time slot $t$; let this number be denoted by $N(L, A, t)$.

2. For each $A$ and $L$, compute $N_{\text{max}}(L, A) = \max_t N(L, A, t)$.

3. Let the capacity of sublocation for activity $A$ at location $L$ be $M(L, A)$, which is referred as sublocation size in our paper. For each $L$ and $A$, generate $N_{\text{max}}(L, A)/M(L, A)$ sublocations within location $L$.

4. For each activity in the activity file, assign a sublocation uniformly at random.
Table A.1: Sizes of the generated contact networks.

<table>
<thead>
<tr>
<th>Network</th>
<th>No. of nodes</th>
<th>Avg. deg.</th>
<th>Max Deg.</th>
<th>Min Deg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>13,850,507</td>
<td>75.99</td>
<td>286</td>
<td>1</td>
</tr>
<tr>
<td>Beijing</td>
<td>16,191,340</td>
<td>66.77</td>
<td>313</td>
<td>1</td>
</tr>
</tbody>
</table>

The activity list along with the assigned sublocations are fed to a simulation software EpiSimdemics [16] 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.

A.8.3 Structural Properties of the Social Contact Networks

In this section, we discuss some structural properties of the contact networks we constructed for Beijing and Delhi. Table A.1 shows the sizes of the two contact networks: number of nodes, average degree of the nodes, maximum degree and minimum degree. Average degree in the Delhi network is the larger, which means, on average, a person in Delhi is in contact with more people than a person in Beijing. Figure A.2 to A.5 compare the distributions of several network structural properties of the two cities. These distributions show some differences among the cities.

The demographic and cultural difference accounts for the network structural difference revealed in the figures. For example, since school type sublocations generally have larger sizes than work type sublocations, more students in a population 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. Since the synthetic populations are generated based on statistical properties of the real populations. The network structural difference discussed reflects the difference between the real populations, which is a result of fundamental cultural, social, and economical differences, among others, between the urban areas in China and India.

Graphlets are another important feature that have been extensively studied in different applications of networks [62, 79, 84]. Note that our realistic social networks are not simply random graphs, but contain rich label information. Graph label refers to the properties of its nodes or edges, such as demographics and activities. Here we examine a bunch of 3-node and 4-node graphlets in terms of topology structure. The featured graphlets are visualized in Figure A.6, each of which represents an isomorphism class in terms of node rotations. Please note that all edges are reciprocal. The counts of these featured graphlets in the two people-people networks are shown in Figure A.7. From the histogram we can see that Beijing peaks at g3-1-9 while Delhi peaks at g3-1-7. Since g3-1-9 is an interaction triangle
Figure A.2: Degree distributions of the two networks.

Figure A.3: Contact duration distributions of the two networks.

Figure A.4: Distribution of the clustering co-efficient of the nodes.

Figure A.5: Distribution of the shortest path distance of the nodes.
between three adults while g3-1-7 is an interaction triangle between three kids, the different peaks straightforwardly show that Delhi has a younger population and kids contribute more in the social interaction activities. More interesting patterns could be found through the comparison of the graphlet counts.
Figure A.7: Graphlet counts in the two networks.
Appendix B

Model Validation: Sensitivity Test

Detailed and comprehensive data of a region is critical in constructing a high resolution network. However, not all data is available for us to prepare the Delhi network. We have to make assumptions in our model when necessary data is not retrieved yet or unlikely to be available. Two important assumptions in our model, made based on an educated guess, are the sublocation size and the location assignment algorithm. As introduced, people within a location are divided into connected subgroups in a network view. Let sublocation size be the largest subgroup size within a location; it reflects the internal structure of a location. We choose for each type of locations an empirical value for their sublocation size. Also in our model we apply a specific algorithm for assigning activity locations. For example, we assume that a person’s selection of his/her workplace or shopping center is based on the gravity model [15], i.e., further a place is from his home, less likely the place will be chosen. Therefore, for each activity of each person, we assign a distribution over all locations, with those close to the persons home having larger probabilities, and we randomly choose one from the distribution to be the location of that activity.

The synthesized networks may result in different structures due to such assumptions. To evaluate the influence of such choice to the quality of the constructed network, we select Delhi as a testbed and conduct sensitivity tests to measure the divergence in terms of epidemic output. Here we choose the same experimental settings as those in the section of Results.

To evaluate the possible errors introduced by a gravity model based algorithm, we compare it against a random location assignment algorithm. To this end, we switch people’s locations randomly and observe the corresponding changes through simulations. Since a different location distribution may result in a different people-location assignment (refer to Appendix A for details), the location switch can produce the same effect as varying the location distribution.

We start from the original Delhi Network built based on the gravity model and refer to it as the base case. The following switches result in location assignment of different degree of
randomness:

- **normative-network**: No switch. This is the original Delhi network where the location distribution is based on the population density. (base case)

- **school-switch**: We switch the school locations among all people who have school activities. That means, a student attending school A might go to school B instead after the switch, and will thus meet another group of people which introduces a new set of social relations.

- **college-switch**: We randomly switch colleges among college students and faculty.

- **workLoc-switch**: People who have work activities switch their work locations.

- **nonHomeLoc-switch**: People keep their own home location but all other locations are switched with another person with the same type of location.

Corresponding to the five switch operations, we generate five groups of social contact networks respectively. For the base case, there is only one contact network in the group. But for the other four groups, each contains 20 networks built by switching people’s activities with different random seeds. The detailed switching methodology is described in Appendix B.1.

For each group, we plot the averaged structure attributes of the 20 networks as well as their standard deviation as error bars in the Figures B.1 and B.2. The error bars for degree distributions in figure B.1 can hardly be seen, and the error bars for clustering coefficient distributions are very short also. Clearly, the structure differences are negligible for networks within the same group. Also, we can hardly tell the difference of structure attributes between the different groups. For degree distribution, the location switching brings almost no changes, although it does trigger some perturbations to the clustering coefficient distributions. We conclude that the switch of people’s locations doesn’t seem to impact the contact structures significantly.

This result conforms to our theoretic analysis quite well. The degree of a node represents its number of contacts, which come from two parts, one is from contacts at home, the other is from contacts in other locations. After switches, the node’s contacts from home remain unchanged. The node’s daytime locations are switched to another set of locations of the same type. The new locations may have a very different capacity which means they have a very different population size. However, people naturally form subgroups (sublocations) within locations. The size of sublocation is independent of the location capacity. This explains why the degree distributions of all five networks are the same. The similar inference can be applied to analyze the clustering coefficient distribution of the network similarity as we see from Figure B.2.

Similar to the trivial variations in the structural properties of these networks, very small differences in the epidemic dynamics on the switched networks are observed as shown in
Figure B.1: Degree distributions of the Delhi network under different activity location switch scenarios.

Figure B.2: Distribution of the clustering co-efficient of the nodes under different activity location switch scenarios.

Figure B.3: Distribution of edge weight under different activity location switch scenarios.

Figure B.4. We consider the spreads of disease under different scenarios, the base case with no interventions and four different public health strategies: antiviral (AV), school closure (SC), work closure (WC) and vaccination (Vax). Figure B.4 shows the attack rate, peak and peak day of the same H1N1 disease after simulations. The different strategies may show different efficacy in suppressing the disease spreading, but each of them has almost the same effect to all of the five networks with different location switch operations. It seems location assignments have limited impact on the results of the epidemic studies.

Figure B.4: Efficacy of public health policy under different location switch scenarios. All the networks in the figure have an almost overlapping attack rate, peak value and peak day under all four different types of interventions.
We now extend the above conclusion. In our model, a gravity model is used to assign people to their daily activity locations. While the gravity model has been proposed as a classic description of people’s mobility patterns, and is typically used to design people-location assignment in computational epidemiology, it is never rigorously tested. Inconsistency between real people’s mobility data and the gravity model has been observed and some competitive models are presented [62, 124]. The sensitivity test here tells us that which location a person is assigned to for their daily activity seems not that important for epidemic studies. That helps to remove our concern regarding the uncertainties in our gravity-based location assignment model.

The second model parameter in our sensitivity testing is the sublocation size. The sublocation size characterizes the mixing pattern of a location, so it varies for different types of locations, and is affected by cultural, economical and demographic factors. For example, a typical class size in a school in Beijing or Delhi is typically larger than that for Los Angeles because the educational tradition and population density are different in the three cities.

In the previous section, we select the parameter values for Delhi as, sublocation size in work places $w = 25$, in schools $s = 40$ and in college $c = 50$. We compare four networks using the various sublocation size: $(w, s, c), (w + 10, s, s), (w, s + 10, c + 10)$ and $(w + 10, s + 10, c + 10)$. Figures B.5 plots the structural properties of the four networks. As expected, with a larger sublocation size, people in the network are likely to gain more contacts for each activity and thus have statistically larger degree than those in the base case. By increasing all types of sublocation sizes by 10, we get the network $(w + 10, s + 10, c + 10)$, i.e., $(w:35; s:50; c:60)$ in the figures. The new network has a larger portion of high degree nodes than base case. Even if we increase only a couple of types of sublocation sizes $(w + 10, s, s)$ and $(w, s + 10, c + 10)$, the effect is observable, although less obvious. Similarly, for networks generated with various sublocation sizes, their clustering coefficient distributions are also perturbed.
Figure B.5: *The Delhi network* structural properties with different sublocation sizes. (i) On the left is the degree distributions of the Delhi networks generated with different sublocation sizes. The base case is (w:25; s:40; c:50), which means sublocation size for work place is 25, for school is 40, and for college is 50. The network (w:35; s:50; c:60) is generated by increase all sublocation sizes from base case by 10. The degrees of (w:35; s:50; c:60) are obviously larger than those of base case. (ii) On the right is the distribution of the clustering coefficient of the networks generated with different sublocation sizes. Generally, the network (w:35; s:50; c:60) has a larger portion of high clustering coefficient nodes than that in base cases (w:25; s:40; c:50).

Such changes in the structural features of the contact network produce changes in epidemic dynamics correspondingly. From Figure B.6 we can see that if we model the network with a different sublocation size, the predicted peak day as well as the attack rate will be very different.

To be fair, the sublocation size in the real world varies from case to case, even within the same type of places. For example, the size of classrooms in a school may vary from 10 to 100 to meet the needs of different types of classes. To simplify the discussion, we choose the most typical value for the sublocation size to represent all the sublocations of that type. Yet, with the simplified assumption, we can observe from the above experiments that the sublocation size is an important factor in constructing an accurate contact networks. In our future work of this project, we need to collect more data to improve the usability of our results. We have been conducting a survey to investigate the sublocation size distribution in real places, and the new results will be presented in our next paper.
Figure B.6: The attack rate of the Delhi network built with different sublocation sizes, compared under various PI and NPI strategies.

B.1 Method of Location Switch

In this section, we describe how to switch locations for the sensitivity test. The location switch randomly exchange two people’s daytime locations. The operation is conducted towards activity sequences before we feed them to EpiSimdemics to generate the contact networks. The procedure is as follows. First, we give a clear definition on activity sequence. After that, we describe our algorithm and formalize it as Generalized Switch.

B.1.1 Terminology Definition

- **Activity Tuple** is defined as a tuple (activity type, location, sublocation, start time, end time).

- **Activity Sequence (or Activity Tuple Sequence, represented as ATS)**: Each person’s activity is a sequence of activity tuples \([AT_1, AT_2, ..., AT_m]\), where \(AT_i\) is an activity tuple and it satisfies

\[
\begin{align*}
AT_{i}.endtime & \leq AT_{i+1}.starttime, \forall i < k \\
AT_{i}.starttime & < AT_{i}.endtime, \forall i \\
AT_1 & \geq 0 \land AT_k.endtime \leq 24 \times 60
\end{align*}
\]  

(B.1)
B.1.2 Location Switch Strategy

As described in section A.7, the locations and sublocations of non-home activities in the ATS are generated in advance and then assigned to each person based on the gravity model. If we don’t put on the constraint of the gravity model, we will generate ATS for the population in a larger random space. To explore the necessity of the gravity model constraint, we randomly shuffle locations and sublocations among the activity tuples for the populations.

Two preconditions should be satisfied for the switch. First, a location or sublocation should be switched to another location or sublocation of the same type. Otherwise, we will get unreasonable activities for the synthetic populations. Second, the mixing pattern in non-home locations should not be changed. That said, if a person visit three work places (locations or sublocations) on the day, after switching, he/she should be switched to three new work places that were visited by another single person. In this way, we won’t change the mixing pattern in non-home places.

To satisfy the preconditions, we should divide people into different groups. Within each group, the people have same types of activities and the number of activity tuples of that type is same to each other. Then we can shuffle people’s non-home locations/sublocations within the group without violating the preconditions.

Switching locations of a specific activity type among a group of people can be decomposed into a sequence of basic location swaps between two persons. The detailed algorithm is described in algorithm 9.

B.1.3 Generalized Switch

The activity sequences and sublocation model define a people-location (PL) bipartite graph. Observing the PL graph will give us clearer picture into the location switch strategy. As demonstrated in Algorithm 9, switching a group of people’s locations can be decomposed into a sequence of basic location swaps between two persons. The basic location swaps (indicated as * in algorithm 9) between two person as a whole can be modeled as a generalized switch in the people-location graph. The generalized switch is described in algorithm 10.

![Figure B.7: A generalized switch operates over two nodes A and B in the network.](image)
Figure B.7 illustrates how a generalized switch operates over two nodes A and B in the network. Edges between white nodes and A or B are unguarded edge set in the algorithm and the other edges belong to guarded edge set. Following the procedure, in Figure B.8, we extract the subgraph from the people-location bipartite graph that contains two persons $p_1$ and $p_2$ and all edges incident to them. We divide edges incident to $p_1$ into two sets, $E_1$ and $E_2$. $E_2$ is the unguarded set that contains all edges of the non-home activity (edges between black nodes). Similarly, we divide edges incident to $p_2$ into $E'_1$ and $E'_2$. Then the swap of $p_1$ and $p_2$’s location for their work activities is same as a generalized switch to $p_1$ and $p_2$ in the people-location graph.

Figure B.8: The generalized switch in the people-location graph. The swap of $p_1$ and $p_2$’s location for their work activities can be modeled as a generalized switch to $p_1$ and $p_2$ in the people-location graph.
Algorithm 9: Location switch of a population.

Input: ActType: Activity Type to be switched; 
\{p_1.ATS, p_2.ATS, ..., p_n.ATS\} : the set of ATS for each person, where p_i.ATS represents p_i’s activity sequence.

Output: switched ATS for the population \{p_1.ATS\_i', p_2.ATS\_i', ..., p_n.ATS\_n'\}

1. Let m be the maximum number of AT in a ATS;
2. for i = 1 → m do
   3. Let ATS set SS_i = ∅;
4. //Group people of the same type of activity together;
   5. for i = 1 → n do
      6. Let idx = 0;
      7. for AT_k ∈ p_i.ATS do
         8. if AT_k.activityType == ActType then
            9. idx + +;
         10. SS_idx = SS_idx ∪ p_i.ATS;
   11. for i = 1 → m do
      12. Let count = |SS_i|;
      13. Assume SS_i = \{p_1.ATS, p_2.ATS, ..., p_{count}.ATS\};
      14. Let permute(p_1, p_2, ..., p_count) = (p_1', p_2', ..., p_{count}');
      15. Decompose the above permutation into a sequence of swaps: 
          \[(p_{i1}, p_{i2})(p_{i3}, p_{i4})...(p_{it}, p_{it+1})\];
      16. //Randomly swap the locations of two persons with the same type of activity;
      17. for (p_i, p_j) ∈ \[(p_{i1}, p_{i2})(p_{i3}, p_{i4})...(p_{it}, p_{it+1})\] do
         18. for AT_k1 ∈ p_i.ATS do
            19. if AT_k1.activityType! = ActType then
               20. continue;
            21. find next AT_k2 ∈ p_j.ATS, s.t. AT_k2.activityType == ActType;
            22. //switch(loc, loc) indicates the basic location swap in Figure B.8;
            23. //It can be modeled as a generalized switch shown in Figure B.7;
            24. switch(AT_k1.location, AT_k2.location)*;
            25. switch(AT_k1.subLocation, AT_k2.subLocation)*;
      26. output (p_1.ATS, p_2.ATS, ..., p_n.ATS)
Algorithm 10: The generalized switch operation.

Input: $G = (V, E_1 \cup E_2)$, where $E_1$ is the guarded edge set and $E_2$ is the unguarded set; 
\[ \forall S \in N(A), (SA \in E_2 \land SB \in E_1) \lor (SB \in E_2 \land SA \in E_1) ; A, B \in U \subseteq V \]

Output: $G' = (V, E_1 \cup E_2')$

1. $E_2' = E_2$;
2. Let $N(A)$ be the nodes adjacent to $A$; $N(B)$ the nodes adjacent to $B$;
3. for $S \in N(A)$ do
   4. //Neglect edge AB;
   5. if $S \neq B$ then
      6. if $SA \in E_2 \land SB \notin E_2$ then
         7. $E_2' = E_2' \cup \{SB\} \setminus \{SA\}$;
      8. if $SA \in E_2 \land SB \in E_2$ then
         9. swap($SA, SB$) in $E_2'$;
   10. for $S \in N(B)$ do
       11. //Neglect edge BA;
       12. if $S \neq A$ then
          13. if $SB \in E_2 \land SA \notin E_2$ then
             14. $E_2' = E_2' \cup \{SA\} \setminus \{SB\}$;
          15. if $SB \in E_2 \land SA \in E_2$ then
             16. swap($SA, SB$) in $E_2'$;
       17. output $G' = (V, E_1 \cup E_2')$
Chapter 4

Delhi-Network-V2: Evaluating Strategies for Pandemic Response in Delhi Using Realistic Social Networks

Huadong Xia, Kalyani Nagaraj, Jiangzhuo Chen and Madhav Marathe

Abstract

We analyze targeted-layered containment (TLC) strategies to contain an influenza pandemic in Delhi, India. A key contribution of our work is a methodology for the synthesis of a realistic individual-based social contact network for Delhi 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 literature in combination 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. This time-varying, spatially explicit network has over 13 million people and more than 200 million people-people contacts. The network has several interesting similarities and differences as compared to similar networks for US cities.

As a second step, we use a high performance agent-based modeling environment to study how an influenza-like illness would spread over the Delhi network. We also analyze well understood pharmaceutical and non-pharmaceutical containment strategies to control a pandemic outbreak. The results suggest: (i) targeted layered containment strategies perform better than any of the individual interventions; (ii) the epidemic dynamics of the region are strongly influenced by the activity patterns and the demographic structure of its local res-
idents; and (iii) a high resolution social contact network helps in analyzing effective public health policies. To the best of our knowledge, this study is the first of its kind in the Indian sub-continent.

4.1 Introduction

Today’s densely populated urban regions facilitate swift transmission of infectious airborne diseases [137]. Additionally, urban contact networks in regions like India and China are witnessing rapid growth. Delhi, the National Capital Territory of India (NCT), is predicted to 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 [108]. In Beijing, the population rose from 12.9 million in 2000 to 18.8 million in 2010 [98]. The ever increasing density of these urban regions can further increase the risk of an unmitigated pandemic. Nevertheless, public health authorities around the world have focused on developing effective policies to control the spread of diseases — the close coordination among the authorities, use of data driven computational models and timely interventions have helped in controlling a number of recent outbreaks. Pharmaceutical as well as social distancing based interventions have proven effective in this regard.

Networked computational epidemiology is the 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 [53]. The basic approach has now become widely accepted in the epidemiology community. Researchers now agree that a better understanding of social contact network characteristics can provide novel insights into the disease dynamics and intervention strategies for effective epidemic planning.

A methodology to synthesize realistic social contact networks already exists for US cities. Contact networks for US cities are generated by composing the following 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 infer the spatio-temporal mobility patterns of individuals; (iii) user specified interaction criteria is used to estimate region-specific contact networks. The structure of the resulting social networks, calibrated to the above data, has been shown to influence the outcome of disease outbreaks in our simulated epidemic models [52,53].

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 (PUMS) for each Public Use Microsample Area (PUMA) modeled. The US National Household Travel Survey (NHTS) [109] captures the interdependence of people’s activities in the same household across all surveyed households in the United States. Data with a similar level of detail is not available for many other regions (including Delhi,
India), making it impossible to replicate the US network generation process for regions outside the US.

### 4.1.1 Summary of Contributions

Building on our earlier work, we construct a synthetic social contact network for Delhi. To overcome data limitations for Delhi, we developed several new methods, many of which are generic enough to be easily applied to synthesize networks for urban regions in other developing countries. Using a variety of data sources, demographic information for each person, a minute-by-minute schedule of their activities, and the locations where these activities take place are 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 synthetic individuals and $L$ is the set of locations. If a person $p \in P$ visits a location $\ell \in L$, there is an edge $(p, \ell, \text{label}) \in E(G_{PL})$ between them, where label is a record of the type of activity during this visit and its start and end time. To differ this network from Delhi-Network-V1 in Chapter 3, we call this version of the Delhi network as Delhi-Network-V2. The synthetic social contact network is: (i) spatially explicit – home locations, work locations, business locations, educational institutions, government institutions and other places of interest are explicitly represented; (ii) time varying – individuals carry out daily activities based on a normative day visiting potentially several locations, and in turn interacting with other individuals visiting the locations during the same time and (iii) labeled – both individuals and locations carry a range of attributes described in the subsequent sections. Note that it is impossible to build such a network by simply collecting field data. The use of generative models to build such networks is a unique feature of this work.

We then use high-performance agent-based simulations to study the spread of influenza-like illness over the synthetic social contact network of Delhi. We study the efficacy of various intervention strategies, including pharmaceutical and non-pharmaceutical interventions. We rank these strategies by order of their efficacy and discuss how these the outcome of the simulated intervention experiments compares with those reported for other cities in the world. Finally, we carry out a detailed sensitivity analysis to assess the robustness of our conclusions.

The remainder of the paper is organized as follows. Section 4.2 presents the new methodology to generate Delhi-Network-V2, and an overview of the data used for this methodology. In Section 4.3, we analyze structural properties of this network, and compare effectiveness of different public health interventions in an H1N1 epidemic in Delhi using simulations. To verify the robustness of our epidemiological findings against the uncertainty in our methodology, we propose a method to perturb network structures and study the sensitivity of their corresponding dynamical results in section 4.3.4. Section 4.3.5 summarizes our efforts in validating the synthesized networks. Finally, section 4.4 concludes with remarks on future
4.1.2 Related Work

Traditionally, mathematical and computational modeling of epidemics has focused on aggregate models using coupled rate equations [9]. In this approach, a population is divided into compartments according to an individual's health state (e.g., susceptible, exposed, infected, and recovered) and demographic group. The evolution of the infectious disease is characterized by ordinary differential equations. For analytical tractability, these models assume homogeneous mixing, which limits their use 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, that is the study of epidemic processes over explicit social contact networks. Research in this area can be divided into three distinct subareas.

Work in the first subarea aims to develop analytical techniques and computer simulations over classes of progressively sophisticated random graphs [13, 94]. 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 techniques 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) population density and is usually obtained using LandScan data and (ii) the demographic distribution of individuals within a population that is typically obtained from the regions census. A simple template is used to represent a community and these communities are joined hierarchically to obtain larger regions. See [35, 56, 57, 61, 113] for examples of this approach. These models can be extended to obtain hybrid models. In a hybrid model, counties maybe represented as nodes and edges are added between counties to capture the movement of individuals (see [4, 38, 92] for a comparative study). Epidemic dynamics within a county can then be computed using an individual-based model. The epidemic dynamics between counties are captured using coupled rate equations.

The final class of models use the most realistic representation of social contact networks; see [16, 50, 93]. In [16, 21, 22, 50] each individual in the United States is modeled with a detailed demographic profile and a daily activity-location schedule. Our synthetic social network for Delhi is constructed using this class of models.
Table 4.1: Demographic statistics of Delhi in comparison with two other cities.

<table>
<thead>
<tr>
<th>City</th>
<th>Population size</th>
<th>Average age</th>
<th>Average household size</th>
<th>sex ratio (M/F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Beijing</td>
<td>16.20M</td>
<td>37.9</td>
<td>2.6</td>
<td>0.99</td>
</tr>
<tr>
<td>Delhi</td>
<td>13.85M</td>
<td>25.6</td>
<td>5.14</td>
<td>1.22</td>
</tr>
<tr>
<td>Los Angeles</td>
<td>16.23M</td>
<td>32.9</td>
<td>3.0</td>
<td>0.97</td>
</tr>
</tbody>
</table>

4.2 Data and Methodology: Network Generation

In this section, we describe our methodology for constructing a high resolution social contact network for Delhi. The methodology builds on our earlier work done in the context of US cities [15, 52]. As mentioned earlier, an important challenge for synthesizing networks for developing countries is the availability of data. We use a number of imputation techniques to overcome this challenge.

4.2.1 Data Collection

Delhi in the paper refers to the National Capital Territory of India (NCT), including New Delhi and several adjacent urban areas. We model the population of Delhi based on the India Census of 2001 – it was the most recent census when we began the work. Delhi contained over 13 million people in 2001 and is one of the regions with highest population density in the world. The average population age is fairly young with a higher male to female ratio. Important population statistics are summarized in Table 4.1; and compared against 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 and used to construct Delhi-Network-V2 is listed in Table 4.2.

4.2.2 Network Construction Methodology

The procedure to construct Delhi-Network-V2 consists of the following broad steps: (i) synthesis of a baseline population with detailed individual structure that is statistically consistent with the true population; (ii) assignment of each individual a reasonable activity schedule; and (iii) assignment of locations for activities of each synthetic individual. In Section 4.2.2 and 4.2.2, we describe the new methods; other steps are based on the previous approach in [15] and are calibrated to a new data format or data alternatives summarized in Table 4.2.
Table 4.2: Demographics, location, and activity data used in the construction of *Delhi-Network-V2*.

<table>
<thead>
<tr>
<th>Data Set</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>Statistics for demographic variables such as age, gender, income, etc. in individual level and household level for Delhi residents.</td>
</tr>
<tr>
<td>Household Microdata: India Human Development Survey 2005 by the University of Maryland and the National Council of Applied Economic Research (IHDS2005) [46]</td>
<td>Micro samples for household structure, comes from 960 households in Delhi, including 4620 individuals. They describe each household sample: hh size, householder’s age, hh income, house types, animal care; they also depict each individual in the hh: demographic details, religion, work, marital status, relationship to head, etc.</td>
</tr>
<tr>
<td>Locations</td>
<td>It includes the following information for Delhi: (1) Ward-wise statistics for population and households in 2001; (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.</td>
</tr>
<tr>
<td>Activity</td>
<td>The data set collects a travel survey for residents in Thane. Thane is an India city similar to Delhi so we use the survey as a substitute for Delhi. Activity templates are extracted with CART, and assigned to the synthetic population with a decision tree.</td>
</tr>
<tr>
<td>2000 to 2007 school attendance statistics from UNICEF [128]</td>
<td>It collects the ratio of school attendance among school age kids in India.</td>
</tr>
<tr>
<td>India residential area activity survey by NDSSL</td>
<td>The survey focuses on the approximately 40% of adults in India who do not travel to work. People’s age, gender, and contact duration near their home are collected. The survey is conducted by our group NDSSL.</td>
</tr>
</tbody>
</table>
Figure 4.1: Flow chart for synthetic population generation. We create synthetic households by sampling from the joint distribution of demographic variables, in conjunction with household micro-samples. The rectangles represent data sets and the ellipses represent algorithms used in processing data. The algorithm “IPF” stands for Iterative Proportional Fitting, and is used to estimate the joint distribution table of demographic variables.

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

Due to several practical reasons, including anonymity and privacy, access to individual demographic information is not possible nor necessarily desirable. We thus build synthetic populations. A synthetic population is a set of synthetic people, each associated with demographic variables drawn from any of the demographics available in the census or similar aggregate resources. Joint demographic distributions can be reconstructed from the marginal distributions available in typical census data using an iterative proportional fitting (IPF) technique. Each synthetic individual is placed in a household with other synthetic people and each household is located geographically in such a way that a census of our synthetic population yields results that are statistically indistinguishable from the original census data, if they are both aggregated to the block group level. Synthetic populations are thus statistically indistinguishable from the census or any other aggregate data used; nevertheless since they are synthetic they respect privacy of individuals within the population. Note that, census tables are precisely constructed so as to respect privacy: our methods are a way to disaggregate these tables in a statistically sound manner. The synthetic individuals carry with them a complete range of demographic attributes collected in the census data. This include variables such as income level, age, etc.

In doing this, our method makes use of the summary statistics of interested demographic variables at a household level (India Census 2001 [72]) and a collection of household samples from India Human Development Survey 2005 by University of Maryland (IHDS2005 [46]). Household samples in IHDS2005 [46] comes from 960 households in Delhi, comprising 4620 individuals. They describe demographic attributes of each household sample and each individual in the household, as listed in Table 4.2. This is our first example of data imputation. PUMS or public use microdata is usually available for cities in the US but is currently not
Figure 4.2: The synthetic information viewer. Our laboratory has developed an interactive web-based graphical tool called synthetic information viewer to visualize the synthesized population. Figure shows a map of Delhi and its 114 wards. For example, clicking on ward No. 40 shows the statistics of synthetic households in the ward, which comprise 31,683 households and 163,246 individuals. The right panel displays the household size distribution for the synthetic households in ward No. 40.
available for India. Nevertheless, IHDS survey is extensive and provides this information. The data had to be manipulated to produce a PUMS like sample for the Delhi region. A flowchart of the method is illustrated in Figure 4.1. The synthesized population with realistic household structures is illustrated in Figure 4.2. Delhi is divided into 114 wards (an administrative region); a synthetic subpopulation is created for each ward separately using the above algorithm.

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

If available, raw travel survey data for the Delhi would have afforded a direct implementation of the activity assignment method described in [15] as the next step toward building the Delhi synthetic network. Such data was not easily accessible at the time of this study. However, due to the availability of detailed summary statistics of the 2001 Thane, India travel survey in literature [10, 102], we devise a discrete-time simulation to generate and assign activity schedules to the Delhi synthetic population. Thane is a city in the western state of Maharashtra, India. A quick comparison based on census data [72] reveals the high degree of similarity of the demographic structure and religious/cultural habits between the two India cities, therefore we consider Thane to be 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 the sociodemographic profile of mobile adults (adults that recorded at least one trip) and people recording no trips, as well as travel data statistics in the form of empirical frequency distributions of trip start times and trip durations of the survey sample population. Statistics of personal and household trip rates are split by mode of transportation, household size and individual worker status. The literature also briefly describes trip frequency, activity characteristics, and time use characteristics of students younger than 16 years old, students older than 15 years old, and mobile adults. Detailed trip chaining analysis is also reported for commuters (adults reporting at least one work-based trip). All trips reported in the survey began at home and ended at home. Based on the Thane survey statistics reported in [102] and [10], the activity assignment process described in Algorithm 11 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 (namely, age and gender). For adults, this is achieved by sampling from the commuter status and demographic distribution
of adults in the survey population reported in [10]. The algorithm classifies synthetic adults as commuters (adults reporting at least one work related trip), non-commuters (mobile adults with no work related trips), or zero trip makers. We further assume all adult non-commuters below the age of 23 have school related activities and classify them as college attendees. Since the literature reports commuter status statistics only for adults, we make the following assumptions about individuals aged 17 years or less, henceforth referred to as kids. A kid under the age of 6 years is assigned the same sequence of activities as an adult from the same household having no work related or school related activities. Kids 6 to 10 years old are classified as primary school attendees, non-school goers making at least one trip in a day, or zero trip makers. Similarly, kids 11 to 17 years old are classified as secondary school attendees, non-school goers that make at least one trip in a day, or zero trip makers. These assumptions are made based on observations from the real world. 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 [128] as well as 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 [10]: home to school and school to home. As a result, we assign the activity sequence home – travel – school (college) – 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 non-commuters. Close to all non-commuting adults report exactly two trips in a day [10], 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 available literature provides no information on non-commuter kids in the survey, we assume that the above frequency distribution of activity sequences of non-commuting adults applies to non-commuter 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) conditional on the time left until the end of the day. 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 11: Assign Activities

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

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

1. for each synthetic individual $i$ do
   2. $\left[\xi, \text{actCLASS}_i\right] = f_1(\text{age}_i, \text{gender}_i, \xi)$; /* assign activity class */
   3. $\left[\xi, \text{actSEQ}_i\right] = f_2(\text{actCLASS}_i, \xi)$; /* assign activity sequence */
   4. for each activity $j$ in $\text{actSEQ}_i$ do /* generate detailed schedule */
      5. $\left[\xi, \text{startTime}_{i,j}\right] = g(\text{actSEQ}_i, \text{activity}_j, \text{endTime}_{i,j-1}, \xi)$
      5. $\left[\xi, \text{endTime}_{i,j}\right] = h(\text{actSEQ}_i, \text{activity}_j, \text{startTime}_{i,j}, \xi)$

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

Location creation, assignment and contact network estimation

Locations are where people conduct their activities (including household activity). They decide how people are distributed in the geographical space of the city. The data set of MapMyIndia [87] contains land use statistics for Delhi, including geo-coordinates for various classes of points of interest (POI). This includes, work locations, malls and shopping centers, recreational places, official buildings, etc. MapMyIndia provides one of the most extensive POIs for Indian cities. Although the data was collected for mapping services, it provides just the kind of information we need. We extracted those coordinates and assigned people to those locations for their daytime activities. Schools, colleges, shopping centers and other places are also considered as 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. We do not have a complete data set for real home coordinates. However, the city of Delhi is divided into 114 wards and we know the number of households in each ward (Figure 4.2), which helps us precisely distribute home addresses over the whole city.

After we assign people to locations based on their activities, we then capture their geo-spatial positions over the course of a day. In Figure 4.3 we illustrate such travel routes for three members in a typical family.
Figure 4.3: The daily travel routes for all three members in a family. The routes for different members use different line style. Father’s routes are shown as solid thick lines: home→work place→other place→home; mother’s are solid thin lines: home→work place→mall→home; son’s are dash lines: home→school→home.

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

A people-people contact network $G_P$ can be inferred from $G_{PL}$. Potential people-people contacts occur when two persons coexist in the same location at the same time. If a location is large, however, two people will not meet even if they are there simultaneously. Therefore, we measure people-people interactions within a location via its sublocation structure. For example, a sublocation could be a room in a building, and people in a room are considered to be in contact with each other.

The sublocation size is considered the average size of sublocations at the peak time within a given location. For each type of location, we select an empirical number as the sublocation size. The sublocation size is an important parameter characterizing the interactions of people within a location. We will discuss potential errors due to a biased sublocation size in section 4.3.4.

Contacts in residential area

The above methodology has been applied to generate several other cities in the world [15,36]. However, as an unusual social-economic phenomenon in India, about 40% of the population does not travel on a daily basis and stays around their residential area for the whole day. The data is verified from two independent sources, a nationwide household survey conducted for
India [46], and the travel survey we retrieved from [102]. Reference [102] reports that 40% of people in Thane survey do not travel, excluding 32% of commuters, 12% of Non-commuters, and 16% of school kids. This is an important socio-cultural difference between developed nations such as US and developing nations such as India.

This motivates us to model the interactions among those people who stay home as their percentage is far from being negligible. We conducted a survey in Delhi and several other cities nearby, collecting data on “at-home people” within a residential area (Table 4.2). Since these people claim they do not travel on a daily basis, we assume they are in contact only with those people within their own locality. The survey gives us the typical number and duration of contacts for people in different demographic groups, and who they are in contact with. The data suggests that residential contacts tend to be volatile and mix homogeneously within same age groups. We further assume that those contacts are highly clustered like any typical social network. With this information, we model contacts between people in residential area as follows. First, we extract probabilistic distributions of contact number and contact duration for each age/gender group from the survey. Second, given each person’s age and gender, we get a number by sampling the contact number distribution and use that number as the degree of this person. Consequently, we get a degree list for all the people in a residential area. Third, given the degree list, we use a configuration model with the added feature of preserving triads [133] to generate a random network. The model is modified slightly so edge weights are calibrated following the contact duration distribution, and the homogeneous mixing is preserved too. We call the generated network a residential network, and edges in the network are deemed residential contacts.

We repeat the above process for each residential area in the city to get a set of residential networks. We then incorporate these residential networks into Delhi-Network-V2 by simply putting all the edges in the residential networks into Delhi-Network-V2.

4.3 Results

We generate Delhi-Network-V2 (both $G_{PL}$ and $G_P$) for Delhi using data listed in Table 4.2 and the methodology outlined in section 4.2. In the following, we conduct a detailed analysis of the synthetic Delhi population and Delhi-Network-V2. We then use EpiFast [21], a powerful epidemic simulation platform, to study the effect of various intervention strategies on the spread of influenza-like diseases in Delhi.
4.3.1 Demographics and Daily Activity Pattern of the Synthetic Population

The individuals in the synthetic population are synthesized by aggregating members from those representative household samples based on the distributions of household level demographics. In Figure 4.4, we compare the individual level demographics of the synthetic population and the true population of Delhi. The linearity of points in the Q-Q plot in Figure 4.6a suggests that the age distribution of the synthetic population matches that of the true population. However, there is a deviation of the age distribution for males (and consequently, females) in the synthetic population from that of the reported distribution of males (and of females) in the census (Figure 4.6b), especially among the younger age groups. The deviation is due to the micro household samples while generating the households, but is relatively small and can be neglected for our purposes here.

Figure 4.5 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: home, work, school, etc. The aggregated activities have a bias towards “home”. For any given time, there are more people at home than at all other locations. This is due to the special economic phenomenon in India where about 40% of people do not travel on a daily basis. As discussed earlier, this 40% comprises of stay-at-home women, the elderly and a portion of young people. This phenomenon is observed in the Thane survey and is assumed to be true for Delhi as well. Most people work or study during the day, and almost all people stay at home late night. This unique cultural feature is quite different from a typical city in the US.

Figure 4.4: The comparison of age-group counts. The left side depicts that for Delhi from the India Census 2001 [72]. The right side depicts that for the synthetic population based on micro sample data and household level statistics. Visually we see the synthetic population conforms to the real statistics quite well. QQ-plots in Figure 4.6 show us a clearer visualization.

Figure 4.5: Temporal 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.
Next, we turn our attention to the metrics of people’s mobility. People’s mobility pattern is a highly relevant factor to epidemic spreading among a population. In Figure 4.7 and 4.8 we show two measures in this regard, the daily travel distance and the radius of gyration. The daily travel distance is the summation of all trip lengths of a person in a typical day. Assume a person visited $N$ places on a day (repetitive visits to a same place are counted multiple places), then the radius of gyration of the person is defined as:

$$R_g = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (r_k - r_{\text{mean}})^2}$$

where $r_{\text{mean}}$ is the mean position of the $N$ visited places. The travel distance depicts the length of the movement, and the radius of gyration depicts the area covered by a person’s movement. From the figure, we see that the distribution of the two metrics both reveal a truncated power-law like curve. This is consistent with the findings in the literature [62].

### 4.3.2 Graph Structural Properties of the Contact Networks

Metrics related to structural analysis of $G_{PL}$ and $G_P$ are shown in Figure 4.10. In the bipartite people-location graph $G_{PL}$, there are 13.85 million people and 1.11 million locations. Degree distribution is plotted in log-log scale in Figure 4.9. A large part of the degree sequence follows a power law distribution, which conforms to findings in other work [53]. For $G_P$, we plot the distribution of node degrees, clustering coefficients and contact durations.
Obviously, the degree distribution in $G_P$ is totally different from that of $G_{PL}$. To get a better understanding of the epidemic implication of those measurements, we compare these structural properties of Delhi-Network-V2 against those of the Los Angeles network in Table 4.3. Different from theoretic assumptions such as power law degree distribution, the degree distribution of Delhi-Network-V2 $G_p$ peaks around a degree of 20. The average degree is about 30; this is a relatively small number based on our other study of US cities [36]. Compared to the Los Angeles network, the average edge weight (representing accumulated contact duration) is longer, and the average clustering coefficient is significantly higher. Those structural features suggest residents in Delhi tend to stay with a few fixed acquaintances for a long time instead of meeting many unfamiliar people for a short time. Such a contact structure leads to implications regarding the pandemic spreading in the population.
(a) Degree distribution of $G_P$. The degree of a vertex in $G_P$ depicts the number of contacts of the corresponding person over the course of the day. The average degree is 29.86.

(b) Duration distribution of people-people contacts (edge weight in $G_P$). The average contact duration is about 363 minutes.

(c) Shortest path distribution of $G_P$. The shortest path between two vertices (from the same connected component) is the minimum number of hops between them over the network. The average shortest path is 6.4 hops.

(d) The clustering coefficients (CC) in $G_P$. CC of a vertex is given by the proportion of links between the vertices within its neighborhood divided by the number of links that could possibly exist between them. The average CC of Delhi-Network-V2 is 0.546.

Figure 4.10: Network structure profiling for Delhi-Network-V2 ($G_P$). $G_P$ may contain multiple connected components, those distributions are aggregate of distributions from all connected components.

Table 4.3: Average structure properties of several city-scale contact networks.

<table>
<thead>
<tr>
<th>people-people network</th>
<th>No. of edges</th>
<th>number of nodes</th>
<th>Avg. degree</th>
<th>Avg. edge weight (minute)</th>
<th>Avg. CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>the Delhi network</td>
<td>206,787,386</td>
<td>13,850,507</td>
<td>29.86</td>
<td>363</td>
<td>0.546</td>
</tr>
<tr>
<td>the Los Angeles network</td>
<td>459,273,880</td>
<td>16,228,759</td>
<td>56.60</td>
<td>141</td>
<td>0.389</td>
</tr>
</tbody>
</table>
Figure 4.11: Visualization for labeled graphlets. Here each graphlet represents an isomorphism class. The node labels include k, t, a and s, which represent preschool kids, school students (most are teenagers), adults and seniors respectively. Here kids are aged between 1 to 5, students are between 6 to 18, adults are between 19 to 60 and seniors are older than 60.

Graphlets are another important feature that have been extensively studied in different applications of network science [62, 79, 84]. $G_{PL}$ and $G_P$ are relational networks; labels on their edges and nodes capture important demographic and other social features. Here we examine a number of 3-node and 4-node graphlets as visualized in Figure 4.11. Note that each graphlet represents an isomorphism class in terms of node rotations, and all edges are reciprocal. The counts of these featured graphlets in the Delhi network are shown in Figure 4.12. Similarly, we compare the graphlet distribution to that in the Los Angeles network. A significant structural difference of the two networks is revealed in terms of the graphlet decomposition. From the histogram we can see that Los Angeles peaks at g3-1-9 while Delhi peaks at g3-1-7 and g4-2-0. Since g3-1-9 is an interaction triplet between three adults while g3-1-7 and g4-2-0 represent intensive interactions among three or four kids, the different peaks straightforwardly show that Delhi has a younger population and kids contribute more in the social interaction activities than those in Los Angeles. More interesting patterns could be found through further comparisons.

### 4.3.3 Epidemic Dynamics and Intervention Policies

#### Epidemic model

We now turn our attention to dynamics on the social contact network. We focus on influenza like illness for this. We run epidemic simulations on the Delhi contact network to study
Figure 4.12: Counts of labeled graphlets. Corresponding graphlets are visualized in Figure 4.11.

epidemic dynamics and the effects of public health policies in the Delhi population. We simulate the spread of H1N1 [69]. The disease progression within the individual-based model follows the standard Susceptible-Exposed-Infectious-Recovered (SEIR) model. Three key parameters to the epidemic dynamics are the basic reproduction number, $R_0$; incubation period distribution and infectious period distribution. (i) $R_0$ is the number of secondary cases one case generates on average in a previously unaffected population. The $R_0$ of H1N1 is studied extensively: for India, 1.45 is estimated [91], estimations in other regions range from 1.20 to 1.68 [59, 74, 105, 126, 146]. Not a unique “precise” value is agreed upon. To address the variations in different estimates of $R_0$ of H1N1 in the literature, we choose a set of values: 1.35, 1.40, 1.45, and 1.60. The range of these values covers most estimates for $R_0$ of H1N1 found in the literature. (ii) The incubation period is the interval between exposure to an infectious disease and the appearance of the first signs or symptoms and usually lasts between 1-4 days for seasonal influenza [33]. (iii) The infectious period is the interval during which infected individuals can spread the disease to susceptible individuals. The period typically lasts 3-5 days [33]. The incubation and infectious periods in our model are described using discrete probability distributions since the actual length of these two periods vary on different individuals. In the experiments, we select the discrete distribution of incubation period as “1:0.3 2:0.5 3:0.2”, meaning that an exposed individual stays in incubation status for 1 day with a probability 0.3, 2 days with a probability 0.5 and 3 days 0.2. Similarly, we select the distribution of infectious period as “3:0.3 4:0.4 5:02 6:0.1”. We therefore design 4 different disease models with four different $R_0$ values and the fixed distributions of incubation and infectious periods.

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 4.13. The distributions for other $R_0$ are very similar to that of $R_0 = 1.35$ (omitted here to save space), indicating 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 $R_0$.

![Vulnerability histogram for nodes in Delhi-Network-V2](image)

Figure 4.13: Vulnerability histogram for nodes in Delhi-Network-V2. This is based on 10,000 random epidemic simulations with $R_0 = 1.35$. In the figure, the fraction of low vulnerability nodes are more than the fraction of high vulnerability, where about 40% of people have a vulnerability lower than 0.35.

The vulnerability distribution of the people varies from 0 to 1, biased toward the left side. Quite a few people have a vulnerability close to 0. 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 4.5.

**Optimal intervention strategies during epidemic spreading**

Using a high resolution contact network modeled for Delhi, we are able to achieve a better understanding for the epidemics and effectiveness of different intervention policies. We simulate four public health policies frequently applied in the real world, including pharmaceutical interventions (PI) and non-pharmaceutical interventions (NPI). PI includes *Antiviral* and *Vaccination*; NPI includes *School Closure* and *Work Closure*. The simulation results when $R_0 = 1.35$ are presented in Figure 4.14. 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 insensitivity to $R_0$ here and in the following sections suggests that the ordering of the policy effects remain the same regardless of the $R_0$ value of a disease.

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

School Closure and Antiviral have their pros and cons. Antiviral will help reduce the attack rate more than a School Closure, but a School Closure works better in reducing the maximum number of cases on any day (peak), and in delaying the occurrence of the peak. School Closure, however, is better in all three parameters (attack rate, peak population, and peak day) than Work Closure. By dissecting into the subpopulation structure and comparing their epidemic dynamics, we could gain insights on controlling disease spread. In Figure 4.15 we plot the epidemic curve, which shows the fraction of people infected on each day, for each of the four subpopulations (preschool, school age, adult, and senior). As observed from Figure 4.15, among all subpopulations, only school age has an epidemic worse than the population average (green curve in the figure). Closing schools can avoid disease transmissions between students within schools, which explains the high effectiveness of School Closure.

Targeted-Layered Containment

In the last section, we simulate different intervention strategies separately to test their efficacy. In real life, however, several types of interventions are typically conducted simultaneously. Therefore we consider combinations of a set of interventions, including targeted interventions and general interventions. Such a combination of various interventions is called targeted-layered containment (TLC) [67].

We consider the TLC policy with a combination of the following interventions, and examine multiple levels of compliance with the interventions and infection rate thresholds for initiating
interventions. Vaccination is excluded, since it is often unavailable at the early phase of epidemics.

- **Targeted-Anti-Viral**: diagnosed individuals are applied anti-viral therapy (under some compliance).

- **Targeted-Stay-Home**: diagnosed individuals are suggested to stay at home (under some compliance).

- **School-Closure**: students’ school activities are removed (under some compliance and initiated with specific infection-rate threshold).

- **Work-Closure**: work places are closed (under some compliance and initiated with specific infection-rate threshold).

For the disease parameters, we assume the asymptotic rate is 20%, and the ascertainment rate is 75%, therefore, the probability an infected individual is diagnosed is 60%. The initiating threshold for all interventions are 0.1% and the compliance is 60%. The comparison between each type of intervention and TLC (the combination of all interventions) are shown in Figure 4.16. TLC is the combination of all other interventions, therefore obviously and reasonably, the most effective among all interventions.

To examine the robustness of the results to model assumptions, we conduct sensitivity analysis as follows. For each intervention within the TLC, we vary the compliance levels to be 30%, 60% and 90%; and vary the initiating threshold to be 0.01% and 0.1%. The results are shown in Table 4.4; a more straightforward visualization is plotted in Figure 4.17.
The results indicate that the time when the TLC is initiated is more important than people’s compliance to the interventions. Compared to the baseline case where no intervention is conducted, a TLC initiated when the infected rate is 0.1% brings a significantly lower attack rate. While a TLC initiated when the infected rate is 0.01% does not reduce the attack rate much. The reason is likely due to the susceptible population size. When an epidemic is alleviated by an early TLC, not many people become immuned. When the disease eventually does outbreak later, most people are still susceptible and become severely infected.

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

Table 4.4: Results of Targeted-Layered Containment in Delhi when $R_0$ is 1.35.

<table>
<thead>
<tr>
<th>compliance (%)</th>
<th>threshold(%)</th>
<th>average attack-rate(fraction)</th>
<th>average peak (fraction)</th>
<th>average peak-day</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA (baseline)</td>
<td>NA</td>
<td>0.434817</td>
<td>0.00924687</td>
<td>117.6</td>
</tr>
<tr>
<td>30</td>
<td>0.01</td>
<td>0.432875</td>
<td>0.00901022</td>
<td>136.967</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.416667</td>
<td>0.00714037</td>
<td>140.633</td>
</tr>
<tr>
<td>60</td>
<td>0.01</td>
<td>0.433024</td>
<td>0.00902515</td>
<td>156.933</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.417798</td>
<td>0.00713408</td>
<td>164.2</td>
</tr>
<tr>
<td>90</td>
<td>0.01</td>
<td>0.433257</td>
<td>0.00904592</td>
<td>168.03</td>
</tr>
<tr>
<td></td>
<td>0.1</td>
<td>0.419244</td>
<td>0.00728672</td>
<td>176.8</td>
</tr>
</tbody>
</table>
4.3.4 Sensitivity Test for Our Synthetic Network Model

Delhi-Network-V2 was constructed using aggregated and noisy information — this aspect is inherent to the process of generating such networks. Given that the structure of the network crucially affects the disease dynamics and interventions, we carry out a detailed sensitivity analysis. Two important model parameters are the sublocation size and the location assignment algorithm. As described earlier, people within a location are divided into connected subgroups in a network view. Let sublocation size be the average subgroup size within a location; it reflects the internal structure of a location. We define for each type of locations an empirical value for their sublocation size. Please note that sublocation size is a 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 observations in the real world [15], the gravity model suggests that the distance between one’s home and work place or other activity locations follows an exponential distribution: $f(x; \lambda) = \lambda e^{-\lambda x}$ where $\frac{1}{\lambda}$ is the mean distance. We choose the same experimental settings as those in Section 4.3. Here we assume $R_0 = 1.35$. We point out, however, that the observations are similar for the sensitivity experiments with the $R_0$ value being 1.40, 1.45, or 1.60.

The sensitivity analysis regarding various sublocation size are shown in Figure 4.18. The effect of varying sublocation size has a significant impact on the spread of the disease as well as interventions. Second, changing the sublocation size of some specific types of locations changes the structure of the network. For example, in the baseline network, closing schools is more effective in delaying disease spreading as compared to closing work locations. For the network constructed after we increase sublocation size of work places (w+10 in Figure 4.18), however, the effect of closing work places is as significant as that of closing schools.

To quantify the role of location assignment process, we switch locations for two randomly chosen people with the same type of activities. The results are summarized in Figures 4.19) Figure 4.20. The results indicate that network structure, dynamics and policies are quite robust to the location assignment process.

4.3.5 Validation, Assumptions and Limitations

Model validation is an important issue when building complex models for social systems. It is becoming increasingly well accepted that for such models, predictive validity is of limited use. In large social simulations, such predictions usually take the form of postdictions of historical information, e.g. matching the epidemic curve to historical data. Although useful in certain cases, it can also inadequate for a number of reasons. First note that any measured real world data is incapable of capturing the range of possible outcomes only those modes that happened in the real world appear in the measured data. Thus the process of postdiction alone is inadequate. Second, high dimensional models such as the ones discussed here have
enough latitude to fit them to relatively sparse data that is collected in the field. One of our goals was to develop models that have explanatory power. Another goal was to use the models for counterfactual analysis such as assessing layered containment strategies in the event of a pandemic. Substantial effort has been made to validate our complex models with these goals in mind. We discuss this briefly below.

1. Our overall process for producing synthetic populations and networks has been published in [14, 15]. The resulting synthetic population is guaranteed to be statistically identical to original data that was used to synthesize it. This includes, census and micro-survey data [43, 44, 46, 72], location data from MapMyIndia [87], activity and time use surveys [96, 114], transport networks [10] and residential surveys. Thus for example, total population and fraction of individuals in various age groups, ratio of males to females, number of school going children, number of workers in our synthetic population matched survey data. Similarly, the number of schools, colleges, their locations, population density all were matched.

2. Data-driven social, behavioral and epidemic theories were used to develop the procedural components of our model. This includes: (i) location assignment models based on generalization of gravity models [15, 28, 49, 102, 136], (ii) homophily-based assignment of activity templates to individuals [15, 19, 28], (iii) within and across host disease progression models [106, 107, 126], (iv) models of social distancing and compliance [15, 67].

3. Further validation of the synthesized network was undertaken. This included structural properties of the time varying network [52, 53], visitation of individuals to locations [62], temporal variation of location occupancy [53], homophily properties in

![Figure 4.18: Epidemics and policy efficacy in Delhi-Network-V2 with various sublocation sizes ($R_0 = 1.35$). Six types of locations are modeled in Delhi-Network-V2: 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.](image)
Figure 4.19: Network structure perturbations for the Delhi Network ($G_P$). Only CC and shortest path distribution change slightly due to the location switching.
4.20: Impact of location switches to Delhi-Network-V2. We simulate under different public health intervention policies ($R_0 = 1.35$). Two people are selected randomly to exchange their daytime locations (named one switch) only when the locations are of the same type. In the control cases in the figure, “workLoc-switch” means work-places are switched randomly between workers. Other legends are self-evident in their meaning. For each case, we conduct a large number of switches to assure system convergence. However, the location switches in all cases do not change the epidemic dynamics of the underlying networks.

subnetworks [90], structural motifs [19, 121] etc. Note that the structural properties were not a part of model input.

4. Finally structural validation has been undertaken [19, 138]. This includes: a comparison of mobility results with recently published mobility data and laws [62], epidemic dynamics and interventions over structured networks [67].

Despite this, the models nevertheless have a number of limitations; future work will seek to address these limitations. Important limitations include:

1. Detailed location data in residential areas was not used. Specifically, detailed information about precise location of each residential building was not used in the modeling process. Such data is available but expensive. We are investigating open source solutions to overcome this issue.

2. Delhi as well as other urban regions in developing nations have slums. Slums were not well represented in our populations; although they are picked up at an aggregate level. Slum populations have several unique cultural and social attributes. Furthermore, the slums are usually very dense agglomerations of households.

3. Thane activity survey was used as the best extant representation of the activities undertaken by individuals in Delhi. Delhi specific activity surveys should be used in future work.

4. Finally, Delhi has a lot of migrant workers; moreover the population is increasing rapidly. We have assumed that the population is fixed over the course of the epidemic.
4.4 Conclusion

The social contact network plays an important role in infectious disease epidemiology. In this paper, we described a methodology to generate high-resolution urban scale social contact networks for regions in developing countries. We focused on the national capital territory in India that comprises of New Delhi and the surrounding areas. A key challenge in developing a high resolution social contact network in India is the availability of data sources. We discussed how imputation methods developed in statistics can be used to overcome this challenge.

We then used $EpiFast$ – a high performance computational epidemic modeling framework to study how an influenza like illness can spread through the region. Detailed computational experiments were carried out to assess the efficacy of various social and pharmaceutical interventions.

As a part of the paper, we have begun investigating a fundamental question in networked epidemiology — how much and what details need to be captured when constructing synthetic social contact networks. As a first step, we have developed a number of structural and dynamical measures of social contact networks. These measures serve as a basis to compare two synthetic networks. Future work will investigate the role of these structural measures on network dynamics and more importantly on interventions.
5.1 Introduction

Today’s typical urban area is accumulating a higher population density than ever, which creates a challenge in controlling epidemics. Influenza-like diseases spread through people–people interactions, therefore a high density urban area creates an ideal fabric for the diffusion of infectious diseases among a population.

To attack this challenge, networked epidemiology provides a computational solution for understanding the diffusion process. In this solution, people–people interactions are modeled as a social contact network where nodes represent people and edges represent the proximity between them.

Given the scale, such a network has to be synthesized. The idea is to figure out people-location relationships during the course of an entire day. In doing this, we collect various data sources: demographic census data, travel surveys related to people’s behavior, and city land use data. Using this data and building on a previously proposed method [15], we have created a high resolution network for Delhi.

Our methods have been used to create synthetic populations and networks for major U.S. cities [15]. The work of this paper extends the previous models in three distinct ways: (i) New data and models are incorporated. We extend our work to build the network for Delhi, a city which has a very different demographic structure, cultural environment and
infrastructure than a U.S. city. For example, community activities in U.S. cities are not critical in shaping the local contact structure. However, we find they are not negligible in Delhi, therefore, we create another layer of the social contact network (the residential network) to model those community activities. The work of this chapter verifies that the methods are fairly open and generic after customization to the region-specific data. (ii) We use the network constructed for Delhi to study the social contact structure in the city. We have already had two sets of methodologies in place for this region – Delhi-Network-V1 and Delhi-Network-V2 – but this one has the highest accuracy in terms of the data used. To the best of our knowledge, it is the most detailed social contact network ever developed for an Indian city. (iii) Divergence from Delhi-Network-V1 and Delhi-Network-V2: Activity is assigned to a whole household instead of an individual, which preserves the correlation between household members’ behavior. Herein, we refer the Delhi network created in this chapter as Delhi-Network-V3.

5.2 Data and Methods

5.2.1 Data

The data we use to create Delhi-Network-V3 are listed in table 5.1. Multiple data sources are referenced, including demographics, locations and activity aspects. Please note that Delhi-Network-V3 makes use of same demographics and location data sources as Delhi-Network-V2. Both of them make use of India residential area activity survey by NDSSL too. The differences lie in that we collect household travel survey for Delhi specifically in Delhi-Network-V3 (DDMS: Delhi Demographics and Mobility Survey by MapMyIndia and NDSSL in the table), and use a household travel survey for Thane as a substitute in Delhi-Network-V2. In the DDMS survey, we have activities of a household instead of independent individuals. Therefore we have real disaggregated structure of people’s activity, and even better we know the correlation between family members (for example, an infant will be together with at least one of adults in the same household).

5.2.2 Methods

We use “first principles” based methods to model the social contact network for an urban region. The methodology starts by building a synthetic population, then adds their mobility in a simulated geo-spatial territory. The procedure can be divided into four broad components: synthetic population; activities assignment based on surveys; location assignment for activities, and social contact network construction.
Table 5.1: Demographics, location and activity data used in the construction of *Delhi-Network-V3*.

<table>
<thead>
<tr>
<th>Data</th>
<th>Statistics</th>
<th>Sources</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>13.85M population</td>
<td>India Census 2001 [72]</td>
<td>no relevance between demographic variables</td>
</tr>
<tr>
<td></td>
<td>2.67M households</td>
<td>Household Microdata: India Human Development Survey 2005 [46]</td>
<td>limited coverage</td>
</tr>
<tr>
<td>Locations</td>
<td>2.64M non-home locations</td>
<td>MapMyIndia Dataset [87]</td>
<td>ambiguous location types</td>
</tr>
<tr>
<td>Activity</td>
<td>samples: 8484 respondents from 2366 households</td>
<td>Delhi Demographics and Mobility Survey by MapMyIndia and NDSSL</td>
<td>Household based travel survey with detailed household member demographics</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>00-07 school attendance statistics from the UNESCO Institute of Statistics (UIS) [128]</td>
<td>collective statistics</td>
</tr>
<tr>
<td></td>
<td>37 samples</td>
<td>India residential area activity survey by Network Dynamics and Simulation Science Laboratory (NDSSL) at Virginia Tech</td>
<td>coverage and background data for surveyed people is limited</td>
</tr>
<tr>
<td>Resultant Network</td>
<td>13.85M nodes 207M edges</td>
<td>-</td>
<td>generated using additional social theories.</td>
</tr>
</tbody>
</table>
Synthetic Population

For the reasons of privacy and expense, the real population with individual demographic information is not available for us to use. Our purpose is to make use of census data and surveyed household samples to construct a synthetic population. Each individual in the population carries a set of demographic attributes such as age, gender, and income, and belongs to a synthetic household. The collective distributions of those demographic variables for the whole population show no significant statistical difference from the census data.

In doing this, our method replicates surveyed household samples as the synthetic households, and in this way maintains a realistic household structure calibrated with actual world data. All the people in those synthetic households naturally form individuals in the synthetic population. The probability that a household sample is selected to replicate is calculated by the joint distribution of household-level demographic attributes, such as household size, householder’s age, household income, et al. Most census data sets contain only marginal distributions for these demographic attributes. A joint distribution table for these variables can be constructed from the marginal distributions using an iterative proportional fitting (IPF) technique [18].

An ideal synthetic population keeps the distributions of demographic attributes consistent with the census data for either the individual level or household level. In the method above, the household level demographic distributions are achieved (with permitted errors) while the individual level is not guaranteed. To obtain a synthetic population with realistic individual attributes, increasing the amount and representativeness of the household samples will help. In practice, we always get satisfying results at both levels (see an analysis of Delhi-Network-V2 in Chapter 4).

The Delhi synthetic population in this chapter is created based on the India Census of 2001 and the India Human Development Survey of 2005, as shown in the “Demographics” row of Table 5.1. This synthetic population is the same as what we created for Delhi-Network-V2 in Chapter 4.

Mobility Model: Activities

To understand where people are during the course of a typical day, it is important for us to know their activity schedules. In our model, an activity schedule is a sequence of daily activities for each person. Such data are typically recreated from activity survey information.

A set of activity templates for households is determined from the surveyed sequences. Each household template includes the activities each household member performs and the time of day they are performed. Such a household activity template is associated with one or multiple combinations of household-level demographic attributes. A classification tree can then be built by choosing appropriate demographic variables, including household size, householder’s
Each synthetic household is matched with a template based on the classification tree.

In the previous two versions of Delhi networks (Chapter 3, Chapter 4), we built unique sets of activity templates for each respective network. They were assigned to each individual without consideration of activity correlation between family members. In this new version of the Delhi network, we have a higher quality data set and can model the activities with a more detailed structure.

**Mobility Model: Location Assignment**

Activities are conducted in geographical sites using commercial data sets and/or open databases. For Delhi-Network-V3, we purchased a high quality geospatial data set from MapMyIndia (MMI data set). The data set is described in Table 5.1. Approximately 2.64 million non-home locations are provided with exact coordinates on the map, along with infrastructure information such as roads, railways, bridges, rivers, and etc.

We classify those locations into the following types: home locations, work places, schools, colleges, shopping places and others. These locations are divided into “anchor” locations and “non-anchor” locations, where “anchor” locations include home locations, work places and schools. Home locations in the MMI data set are not complete and therefore we synthesize them as follows. The number of home locations can be derived by the number of households ward by ward (a ward is a subdivision of the city, of which there are 114 in Delhi). The exact address of those home locations are decided by distribution along streets, roads, trails in the residential area, etc.

Every individual in the model will be assigned a set of locations at which their daily activities are conducted. First, the home location for a synthetic household is selected randomly within the boundary of the given ward. Second, we select anchor locations for the household members based on a “gravity” model. In the model, a non-home anchor location is selected using a probability function that depends on the travel time or distance between the location and the previous anchor location (home by default). The probability function is defined as:

\[ p(D|O) \propto e^{b \cdot \text{Dist}(O,D)}, \]

where \( D \) is destination location, \( O \) is the original location, \( \text{Dist}(O,D) \) is the distance or travel time between them, and \( b \) is a calibration constant. The MMI data set suggests the average travel distance for single trip is 7.75 miles. We calibrate the parameter \( b \) so that the average distance out of the above distribution equals 7.75 miles. Finally, we fit in “non-anchor” locations such as shops between “anchor” locations. The selection of “non-anchor” locations also follows the same gravity model.

In Figure 5.1, we plot the population density after people are assigned to corresponding non-home locations.
Network Generation

The above three steps create a synthetic population with realistic household structures. Each person is scheduled with activities minute by minute on a daily basis and locations where activities take place. This information produces a temporal people-location graph $G_{PL}$. It is a bipartite graph with two types of nodes, people set $P$ and location set $L$. If a person $p \in P$ visits a location $l \in L$, there is an edge $(p, l, \text{label}) \in E(G_{PL})$ between them, where the label carries edge information (such as activity type, duration, start and ending time) or node information (demographics, etc.).

The people-location graph $G_{PL}$ can be easily transformed into a people-people contact network $G_{PP}$. We assume each activity location is divided into sublocations. The sublocation model is a method for defining interactions among persons who visit the same location at the same time. Each person is assigned to a sublocation (within the activity location) randomly. We define a contact between two persons if they are present in the same sublocation 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. The contact network is generated from the activity list along with the assigned sublocations using a previously built simulation tool: EpiSimdemics [16].
Residential Network

The above methodology has been applied to generate several other cities in the world [15, 37]. However, an unusual social-economic phenomenon in Delhi results in about 40% of the population not having a formal job and remaining close to their residential area for the entire day. This data is verified from two independent sources, a nation-wide household survey conducted for India [46], and the travel survey we retrieved from [102]. Reference [102] claims 40% of people do not travel, together with another 32% of commuters, 12% of non-commuters, and 16% of school kids.

This motivates us to model the interactions among those people who stay home as they are not negligible. We conducted a survey in Delhi and several other cities nearby (India residential area activity survey in Table 5.1), collecting data regarding those “at-home people” within a residential area. Since these people claim they do not travel on a daily basis, we assume they are in contact only with those people within their own locality. The survey gives us the typical number and duration of contacts for people in different demographic groups, and with whom they are in contact. The data suggests residential contacts tend to be volatile and mix homogeneously within the same age groups. We further assume those contacts are highly clustered like any typical social network. With this information, we model contacts between people in residential area as follows. First, we extract probabilistic distributions of contact frequency and duration for each age/gender group from the survey. Second, given each person’s age and gender, we get a sample based on the distribution of the contact number and use that sample as the degree for this person, giving us a degree list for all the people in a residential area. Third, given the degree list, we use a configuration model with the added feature of preserving triads [133] to generate a random network. The model is modified slightly so edge weights are calibrated following the contact duration distribution, and the homogeneous mixing is preserved as well. We call the generated network as a residential network, and edges in the network are deemed as residential contacts.

By repeating this process for each residential area in the city, we get a set of residential networks. We then incorporate these into the Delhi network by simply adding all the edges in the residential networks into the Delhi network.

5.3 Results

With the previous methods, we create a realistic social contact network, called Delhi-Network-V3, differentiated from the other two versions of the Delhi network (Chapter 4, 5). Since our methods are first principles based, we investigate the internal structures of synthetic components, including the synthetic population, mobility patterns and activities statistics, as well as structural properties of the network. Our observations are as follows: (i) The synthetic components such as synthetic population demographic structure can be evaluated.
by comparing with real data. We argue that realism of those synthetic components increases our confidence in the quality of the constructed networks. (ii) By measuring the clustering coefficients, degree, shortest path of the Delhi-Network-V3 (together with our previous studies of Delhi-Network-V1 and Delhi-Network-V2), we show that realistic social contact networks share some common properties with other types of random networks, such as highly clustered structure, small world effect, etc. They also reveal structure features that differ from other complex networks in the literature. For example, the degree distribution of the people-people network of Delhi has multiple modes rather than power law distribution. (iii) We use the labeled graphlets to explore the local structure of the network. Such measures are closely related to the local topology structure of the network and the label information is helpful to infer epidemic dynamics.

5.3.1 Structure of Synthetic Components

Delhi-Network-V3 is very similar to Delhi-Network-V2 in many aspects, including synthetic populations and locations (both synthetic and real locations). Please see Chapter 4 for more details and only focus on the synthetic components with new structural properties in this one.

People’s mobility pattern is a highly relevant factor to epidemic spread in a society. The above features of synthetic population, location layout and activity statistics co-evolved together shape the mobility pattern of people. Delhi-Network-V3 has an improved result for the mobility pattern due to the new activity data set applied. In Figures 5.2 and 5.3 we show two measures in this regard, the daily travel distance and the radius of gyration. The daily travel distance is the summation of all trip lengths of a person in a typical day. Assuming a person visited \( N \) places on a day (repetitive visits to a same place are counted as multiple places), the radius of gyration of the person is defined as:

\[
R_g = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (r_k - r_{mean})^2}
\]

where \( r_{mean} \) is the mean position of the \( N \) visited places. The travel distance depicts the length of the movement, and the radius of gyration depicts the area covered by a person’s movement. From the figures, we see that the distribution of the two metrics both reveal a power-law like curve. This is consistent with the literature.

The other metric shown here is the temporal degree of typical locations. Real social networks are dynamic; even simple measures such as degree distributions are temporal functions. In Figure 5.4, we select three typical types of location: work place, home location and schools. For each type of location, we randomly select five locations, then plot their temporal degree. The peak degrees of work locations or schools are around noon, while for home locations it is in the morning and at night, which is consistent with our observations and experiences in the real world.
Figure 5.2: Daily travel distance distribution for people in Delhi within a 24-hour-period.

Figure 5.3: Radius of gyration distribution for people in Delhi within a 24-hour-period.

Figure 5.4: Temporal degree for randomly selected locations across a typical day. From the left to right, the location type is work, home and school.
5.3.2 Network Structure

The people-Location Graph $G_{PL}$ of Delhi-Network-V3 reveals a power law like degree distribution for its location-people network, as shown in Figure 5.5. This feature is also verified frequently in the literature [53], suggesting an potential innate properties for large scale people-location graphs.

![Degree distribution of $G_{PL}$](image)

Figure 5.5: Degree distribution of $G_{PL}$. The plot can also be understood as the visiting counts in all non-home locations.

Similarly, we analyze the people-people graph $G_P$. First, we show the regular structural measures for Delhi-Network-V3 in Figure 5.6, including degree, clustering coefficients and edge weight distribution. Some measures share common profiles with other complex networks [12, 104], such as highly clustered structure, and very limited number of hops for the shortest path between any two nodes in the network. However, some features seem unique to the social contact network. The contact duration distribution has multiple modes, which is driven by the various groups of people with different types of activities. The degree distribution is not like the power law distribution seen in other complex networks. We suggest that our first principles based networks are very different from most existing network models and cannot be produced based on traditional approaches such as ERGM generative models.

Graphlets are another important feature that have been extensively studied in different applications of network science [62, 79, 84]. Note that our realistic social networks are not simply random graphs, but contain rich label information. Graph label refers to the properties of its nodes or edges, in our example demographics and activities. Here we examine a collection of 3-node and 4-node graphlets as visualized in Figure 4.11. Please note that each graphlet represents an isomorphism class in terms of node rotations, and all edges are reciprocal. The counts of these featured graphlets in the Delhi network are shown in Figure 5.7. From the histogram we can see that Delhi-Network-V3 peaks at g3-1-7 and g4-2-0. Since both g3-1-7 and g4-2-0 represent intensive interactions among a groups of kids, the peaks straightforwardly show that Delhi has a young population and kids contribute greatly to the social
Figure 5.6: Basic network structural properties of the people-people social contact network of Delhi, $G_P$. 
interaction activities. More interesting patterns are revealed through further investigation.

5.3.3 Epidemic Dynamics in Delhi

To demonstrate epidemic dynamics on *Delhi-Network-V3*, we run epidemic simulations as follows:

Using our high resolution contact network modeled for Delhi, we are able to achieve a better understanding of both epidemics and the effectiveness of different intervention policies. We simulate an H1N1 like influenza with $R_0=1.35$. Four public health policies are designed, including pharmaceutical interventions (PI) and non-pharmaceutical interventions (NPI). PI includes *Antiviral* and *Vaccination*; NPI includes *School Closure* and *Work Closure*. The simulation results are presented in Figure 5.8.

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

School Closure and Antiviral interventions have their pros and cons. Antiviral will help postpone the outbreak day (peak day) better than a School Closure, but a School Closure works better in reducing the maximum number of cases on any day (peak). School Closure and Antiviral, however, are both better in all three parameters (attack rate, peak population, and peak day) than Work Closure. The reason for the low performance of Work Closure in Delhi could be attributed to the low level of work-related trips (40% of the adults do not
travel to work on a daily basis) and a young population.

Figure 5.8: Epidemics under various intervention strategies in the Delhi network. The simulation assumes $R_0 = 1.35$, and includes a base case where no intervention is conducted. Here we use the tuple (attack-rate, peak, peak-day) to characterize epidemic dynamics. For Vaccination and Antiviral, we randomly choose 25% of the population to apply corresponding pharmaceutical treatments. School Closure and Work Closure are applied for 3 weeks, and the compliance of targeted people is 60%. All interventions are initiated when 0.1% of the nodes in the city are infected.

By dissecting the subpopulation structure and comparing their epidemic dynamics, we can gain insights into controlling the disease spread. We divide the population into four subpopulations with different age ranges: preschool, school age, adult, and senior. In Figure 5.9, we plot the epidemic outcome for each subpopulation. Among all subpopulations, both school age students and kids have an epidemic outcome worse than the population average (blue and green curves in the figure). Adults, on the other hand are more resistant to epidemic diffusion, possibly because of the low social activities of adults in Delhi. Residential contacts seem to impose only slight effects on adults and seniors (the two major subpopulations being modeled with residential contacts). Closing schools can avoid disease transmissions between students, which explains the high effectiveness of School Closure for that subpopulation, while Work Closure’s impact is most directly and effectively felt by the adults subpopulation.
Figure 5.9: Subpopulation epidemics in *Delhi-Network-V3*. We assume $R_0 = 1.35$, the Delhi population is partitioned into four groups based on age: preschool, school age, adult, and senior. The red curve represents the average value for the whole population, and other curves show the percentage of people in a subpopulation infected on each day.
Chapter 6

Network Improvement: Computational Methods for Testing Adequacy and Quality of Massive Synthetic Proximity Social Networks

Huadong Xia, Christopher Barrett, Jiangzhuo Chen and Madhav Marathe

Abstract

Social contact networks play an increasingly important role in computational epidemiology. In this paper we focus on massive social contact networks that cover an urban region comprising millions of individuals and billions of time varying interactions. It is impossible to obtain such networks by simply measuring social interactions. As a result, such networks are often synthesized using a diverse set of real-world data. The synthesis method can be viewed as a complex stochastic process that outputs one realization of such a network. We first synthesize a social visitation network, representing people visiting locations during different time intervals. We then project it to synthesize a people-people contact network. The resulting networks are extremely large, dynamic and unstructured. Any meaningful description of such networks is usually done in terms of structural properties.

Building on our earlier work, we synthesize three versions of social contact networks for the National Capital Territory (NCT) of India with increasing details. The differences of the three networks come from the quality of data collected and the corresponding models applied. Two important questions arise when synthesizing such massive dynamic social
contact networks: (i) How does one compare the networks that span the same region? and (ii) When is the synthesized network adequate? To address them, we compute a number of network metrics, some of which are classical, while others capture the semantics of social contact networks. These metrics are used to study the similarities and differences between two networks representing the same urban region. For question (ii), we study our ability to understand the dynamics and control of epidemics. Dynamical metrics that capture the joint interaction between the local dynamical process and the network structure are presented and used to analyze the NCT networks.

6.1 Introduction

This paper focuses on social contact networks in the context of computational epidemiology. In contrast to mean field based models, network-based models aim to understand the spatio-temporal spread of diseases through a population by studying the epidemic diffusion processes over a network. It is now well accepted that social contact networks are usually not homogeneous. The irregular structure of the network critically affects the disease dynamics. Furthermore, it also affects the efficacy of methods that are developed to contain the spread. Here we focus on the epidemiology of airborne infectious diseases in large urban regions. A good example of this is seasonal influenza. While modeling social contact networks that span such large regions, an immediate difficulty is that such networks cannot be obtained by simple measurements. We resort to developing methods that synthesize such networks from available data; much of the data is aggregate, at a point in time, and often procedural. Moreover the networks are massive, irregular, dynamic and unstructured. In earlier papers, we proposed a method for synthesizing such social contact networks [15, 52]. Our work represented the first such attempt at synthesizing massive social contact networks. Informally speaking, the method can be viewed as a complex stochastic process which yields a random realization of a synthetic social contact network. We will often use the term “synthetic network” without being explicit about this aspect. Nevertheless, this aspect is important to keep in mind throughout the paper. Another important point to remember is that the method follows a fundamentally different philosophy than the ERGM models [70]. We will discuss this later in the paper.

Since our work appeared, several authors have proposed methods for synthesizing urban-scale social contact networks. Each of these methods results in an individual-based network. Individuals correspond to nodes, and edges capture the time varying interactions between individuals. If one fixes on an urban region, each such method produces a network with the same number of individuals. In other words, the cardinality of the vertex set is the same, but often, the similarity stops there. The structural properties of the networks are quite different, because of the data sources used and the algorithmic methods that are used to synthesize the networks using these data sets.
6.1.1 Our Contributions

Given the importance of network structure for epidemic dynamics and control, we are faced with two important challenges. First, given two such urban scale networks, how does one capture the similarities (and the lack) between them? To do this, one needs network metrics that can shed light on the local as well as global structures of the networks. The second challenge is that of adequacy: When is a given social contact network deemed adequate from the standpoint of the questions at hand? Here adequacy refers to developing robust decision making in the context of epidemics. The size, irregularity and the time varying nature of these networks create these challenging questions.

In this paper, we study these two questions by using three sets of complex models for synthesizing the social contact network of NCT, India. The first set of models result in what we call Delhi-Network-V1. It uses minimal data sources that capture the spatial and demographic features of the underlying region. The second set of models result in Delhi-Network-V2. The models use a detailed layout of the actual infrastructure as well as normative surveys to capture the movement patterns of individuals in a city. These models are more intensive in terms of computing and data resources than the first set of models. However, Delhi-Network-V2 makes use of activity survey from another India city, Thane, as a substitute for Delhi. This may cause mis-interpretation of mobility pattern for residents in Delhi. Our third set of models result in Delhi-Network-V3. It uses higher quality data targeted to Delhi, as well as partial improvements in the methodology aspect. The third set of models appear to capture certain topological features more accurately than the other two.

We then describe a number of metrics, divided into four classes: (i) metrics that capture the features of the population, actual infrastructure and their layout, (ii) network level metrics that capture the structural features of the dynamic social contact network, (iii) dynamical metrics that capture the epidemic dynamics over the networks, and (iv) policy metrics that capture the effect of controls. Some of these metrics are well studied in the literature by network scientists, although they have not been studied in our specific context. Other metrics are new and inherently context specific; they are used to capture specific aspects of social contact networks and to study epidemics. Computational considerations are important and we are interested in methods that can efficiently compute these metrics either exactly or approximately (if the problem is intractable). The structure and size of the networks make developing such methods challenging.

Using these metrics we analyze the three networks described above. Our analysis reveals interesting differences between the networks, even though the networks are supposed to capture interactions in the same urban region. The analysis also reveals the kinds of details necessary for a given class of control policies.
6.1.2 Related Work

Traditionally, mathematical and computational modeling of epidemics has focused on aggregate models using coupled rate equations [9]. In this approach, a population is divided into subgroups (compartments) according to an individual's health state (i.e., 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 the use of these models for spatially sensitive processes.

In recent years, high-resolution individual-based computational models have been developed to support planning, control of, 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 two distinct sub-areas.

The first sub-area aims to develop analytical techniques and computer simulations over classes of progressively sophisticated random graphs [13, 94]. 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 is to obtain closed form analytical results.

The second sub-area focuses on the most realistic representation of social contact networks; see [16, 22, 32, 35, 55, 92, 93, 118]. In [16, 22] each individual in the United States is modeled with a detailed demographic profile and daily activities. Massive data with diverse sources are generated with these models and therefore new methodologies and theories are urgently needed to analyze and understand those data as well as the associated large scale networks. Our synthetic social networks for NCT is constructed using this class of models.

6.2 Network Generation: Data and Methodology

To study the intervention strategies for pandemic response, it is important that we create a faithful representation of the people-people contact network in a region. In this section, we describe three sets of models for constructing a realistic social contact network for NCT (National Capital Territory, Delhi, India). One makes use of basic census data and aggregate statistics readily available for most regions. This method is generic, but it produces a coarse level network due to limitations of the input data. The latter two methods uses more faithful and/or higher resolution data sets, and yields a more detailed network.
Table 6.1: Demographic statistics and geo-location of Delhi.

<table>
<thead>
<tr>
<th>City</th>
<th>Population Size</th>
<th>Average Age</th>
<th>Household Size</th>
<th>Sex Ratio (M/F)</th>
<th>City Borders in longi/lat (left,top,right,bottom)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Delhi</td>
<td>13.85M</td>
<td>25.6</td>
<td>5.14</td>
<td>1.22</td>
<td>(76.833,28.892,77.35,28.4)</td>
</tr>
</tbody>
</table>

Table 6.2: Demographics, location and activity data used in construction of Delhi-Network-V1.

<table>
<thead>
<tr>
<th>Data</th>
<th>Statistics</th>
<th>Sources</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>13.85M population, 2.67M households</td>
<td>India Census 2001 [72]</td>
<td>no correlation between demographic variables available</td>
</tr>
<tr>
<td>Locations</td>
<td>56k</td>
<td>LandScan 2007 [30]</td>
<td>noisy data</td>
</tr>
<tr>
<td>Activity</td>
<td>NA</td>
<td>School/College Statistics [43,129]</td>
<td></td>
</tr>
<tr>
<td>Resultant Network</td>
<td>13.85M nodes, 526M edges</td>
<td>Delhi time use statistics [44,96,122]</td>
<td>data is ambiguous and irregular</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-</td>
<td>complex, not describable from pre-defined network models,</td>
</tr>
</tbody>
</table>

6.2.1 Data

In the paper, Delhi and NCT are used interchangeably to refer to the National Capital Territory (NCT) of India. It is the capital of India, including New Delhi and several adjacent urban areas. It is one of the most populous urban areas in the world. We generate a synthetic population of 13.85 million individuals and 2.67 million households, based on India Census 2001 data. Details regarding NCT are summarized in Table 6.1.

Multiple data sources are used to construct the coarse and the detailed networks for NCT. The data we use to construct the networks are listed in the “Sources” column of Tables 6.2, 6.3 and 6.4.

There are several differences between the three data sets. First, Delhi-Network-V2 and Delhi-Network-V3 requires more information to synthesize. For example, micro household samples are used in Delhi-Network-V2 but are not used in Delhi-Network-V1. As another example, subjective surveys regarding individual behavior in residential areas are not used to synthesize Delhi-Network-V1.

Second, some data sets used to synthesize Delhi-Network-V1 are less accurate than those used to construct Delhi-Network-V2 and Delhi-Network-V3. Examples include the activity surveys and location data sets. In Delhi-Network-V1, LandScan data is used to infer locations, while Delhi-Network-V2/Delhi-Network-V3 use real location information directly from MapMyIndia data set. LandScan data contains population counts for each small area, approximately 1km resolution (30'' x 30''), but does not have real location data. MapMyIndia
Table 6.3: Demographics, location and activity data used in construction of Delhi-Network-V2.

<table>
<thead>
<tr>
<th>Data</th>
<th>Statistics</th>
<th>Sources</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>13.85M population</td>
<td>India Census 2001 [72]</td>
<td>no correlation between demographic variables available</td>
</tr>
<tr>
<td></td>
<td>2.67M households</td>
<td>Household Microdata: India Human Development Survey 2005 [46]</td>
<td>limited coverage</td>
</tr>
<tr>
<td>Locations</td>
<td>2.64M non-home locations</td>
<td>MapMyIndia Dataset [87]</td>
<td>type of locations is ambiguous</td>
</tr>
<tr>
<td>Activity</td>
<td>samples: 14,428 respondents from 3,505 households</td>
<td>2001 Thane Household Travel Survey [102]</td>
<td>no predefined data model, only collective statistics available (we don’t have access to surveyed individual data)</td>
</tr>
<tr>
<td></td>
<td>NA</td>
<td>00-07 school attendance statistics from the UNESCO Institute of Statistics (UIS) [128]</td>
<td>collective statistics</td>
</tr>
<tr>
<td></td>
<td>37 samples</td>
<td>India residential area activity survey by Network Dynamics and Simulation Science Laboratory (NDSSL) at Virginia Tech</td>
<td>coverage and background data for surveyed people is limited</td>
</tr>
<tr>
<td>Resultant Network</td>
<td>13.85M nodes 207M edges</td>
<td>-</td>
<td>generated using additional social theories.</td>
</tr>
</tbody>
</table>

utilizes rich urban planning and land use statistics, therefore providing more precise data for locations in Delhi.

The data sets used in the three networks share some common traits. As shown in Table 6.2 to 6.4, all three utilize data from diverse sources; the input data are often noisy and unstructured. These features impose challenges to network synthesis and analysis. The network construction methodology is described in Section 6.2.2. Statistics of the three generated networks are summarized in Table 6.5 and Section 6.3.2. From Table 6.5 we can see that massive scale synthetic structures are generated along with the three resultant networks. To analyze networks of such a large scale, we have to develop novel metrics, and Section 6.3 shows our preliminary efforts to accomplish this task.

### 6.2.2 Methods

In this section, we provide an overview of three sets of methods for generating synthetic populations and networks. The methods are based on our earlier work [15], and consists of the following broad steps: (i) synthesize a baseline population with a detailed individual structure and the same aggregate statistical properties of the real population; (ii) assign each individual a reasonable activity schedule; (iii) create locations in the region where synthetic people perform their activities. In the following section, we describe the differences between
Table 6.4: Demographics, location and activity data used in construction of *Delhi-Network-V3*. The differences between *Delhi-Network-V2* and *Delhi-Network-V3* are indicated with *.

<table>
<thead>
<tr>
<th>Data</th>
<th>Statistics</th>
<th>Sources</th>
<th>Characteristics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>13.85M population</td>
<td>India Census 2001 [72]</td>
<td>no correlation between demographic variables available</td>
</tr>
<tr>
<td></td>
<td>2.67M households</td>
<td>Household Microdata: India Human Development Survey 2005 [46]</td>
<td>limited coverage</td>
</tr>
<tr>
<td>Locations</td>
<td>2.64M non-home locations</td>
<td>MapMyIndia Dataset [87]</td>
<td>location type ambiguous</td>
</tr>
<tr>
<td>Activity</td>
<td>samples: 8484 respondents from 2366 households*</td>
<td>Demographics and Mobility Survey by MapMyIndia and NDSSL.*</td>
<td>Household based travel survey with detailed household member demographics*</td>
</tr>
<tr>
<td>Resultant Network</td>
<td>13.85M nodes 207M edges</td>
<td>-</td>
<td>generated using additional social theories.</td>
</tr>
</tbody>
</table>

Table 6.5: Data Size of the Output Synthetic Structures (in bytes).

<table>
<thead>
<tr>
<th>Synthetic Structure</th>
<th><em>Delhi-Network-V1</em></th>
<th><em>Delhi-Network-V2</em></th>
<th><em>Delhi-Network-V3</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Population File</td>
<td>373M</td>
<td>388M</td>
<td>388M</td>
</tr>
<tr>
<td>Location File</td>
<td>62M</td>
<td>80M</td>
<td>80M</td>
</tr>
<tr>
<td>Activity File</td>
<td>2.7G</td>
<td>1.1G</td>
<td>1.6G</td>
</tr>
<tr>
<td>Residential Network</td>
<td>-</td>
<td>2.1G</td>
<td>2.1G</td>
</tr>
<tr>
<td>the People-People Network</td>
<td>11G</td>
<td>5.6G</td>
<td>6.2G</td>
</tr>
</tbody>
</table>
the three methods in dealing with different input formats.

**Delhi-Network-V1**

*Delhi-Network-V1* uses LandScan data and population census data. The detailed method can be found in Appendix 3-A.

First, we use LandScan data to reconstruct location information for the city. LandScan is a global scale spatial population database [30] where the world is divided into $30' \times 30'$ latitude/longitude grids, with population counts in each cell. The counts are apportioned to each cell based on likelihood coefficients which are calculated using proximity to roads, slopes, land cover, night time lights, and other information. We determine the cells (of LandScan data) that are within the boundary of the region under consideration. A probability is assigned to each cell by normalizing the counts of the cell in the LandScan data. The probability is used to locate households, workplaces and schools in this particular cell. Households, workplaces, and schools are generated following the distribution obtained from the census data and placed at locations using the LandScan data.

Second, a synthetic population is generated from census data: total population, age and gender distribution of the population, household size distribution, workplace size distribution, number of schools of different types, and occupation distributions for different age groups. Each synthetic person is assigned demographics including age, gender, and marital status, a household, and a daytime location, which can be a workplace, school, or a household. When assigning people to households, a set of well-defined rules are followed to maintain a reasonable age gap and gender combination for each family. When assigning a daytime location, the distance from home to the location is considered. A daytime location with distance $d$ is selected randomly with probability following an exponential distribution: $f(d; \lambda) = \lambda e^{-\lambda d}$ where $\frac{1}{\lambda}$ is the mean distance. Such a distance model is called the gravity model [15].

Next, we generate an activity sequence for each person. An activity sequence is a set of major activities a person conducts on a typical day. Each activity includes an activity type, a start time, a duration, and a location. To retrieve the most representative activity sequence for each individual, we make use of a generic activity sequence template library retrieved from a U.S. survey [109], but customized with local activity types and activity time statistics [44,96,122]. These activity sequences are associated with the sample demographics, and are assigned to synthetic individuals according to their demographics.

Finally we synthesize the contact network. The activity sequences and the associated locations define a labeled people-location (PL) bipartite graph $G_{PL}$, where people and locations are the vertices and there is an edge between person $P$ and location $L$ if $P$ visits $L$; time of the visit is the label associated with the edge. The people-people contact network $G_P$ is constructed from the PL graph as follows. We assume each activity location is divided into sublocations. The sublocation model is a way of defining interactions among persons
who visit the same location at the same time and are in spatial proximity. Each person is assigned to a sublocation (within the activity location) randomly. There is an edge between two people if they are in the same sublocation at the same time.

**Delhi-Network-V2**

The method builds on our earlier work described in [15] but accommodates region-specific data as listed in Tables 6.2 and 6.3. It does differ from the method used to create Delhi-Network-V1. In Delhi-Network-V1, synthetic households are created by first generating synthetic individuals matching the census; these individuals are combined to create synthetic households which match the aggregate census data. In Delhi-Network-V2, a collection of household samples from a survey taken in Delhi [46] are used to help construct households with more realistic correlational structure. For example, the correlational structure of the number of kids and adults in each household is better captured in Delhi-Network-V2.

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 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. Our selection of the samples follows the joint distributions of household level demographic variables such as householder’s age and household size. The distribution of each variable is provided in census data, but such a joint distribution is usually unavailable, and we apply the iterative proportional fitting (IPF) procedure to estimate them [18]. IPF is an iterative algorithm for estimating cell probabilities in a contingency table such that the marginal totals remain fixed [15].

The other major difference between the two methods is the creation of a residential network. People who do not have a formal job and stay at home are assumed to interact only with their family in Delhi-Network-V1, yet we know that in the real world social contacts do exist among those people when they walk around their residential area. In Delhi-Network-V2, we conducted a survey in Delhi and several other cities nearby, collecting data regarding those “at-home people”. Since those people claim they do not travel, we assume they are in contact only with people within their own community. The survey gives us the typical number and duration of contacts for persons in different demographic groups, all of whom do not travel to work or school. With this information, we then use a configuration model [133] to randomly generate new contacts for each person in the community. Those contacts form a contact network which we call the *residential network*. We then incorporate it into Delhi-Network-V2.
**Delhi-Network-V3**

The method is very similar to that used to construct Delhi-Network-V2. This is not surprising given that the data input for Delhi-Network-V2 and Delhi-Network-V3 are identical for most sources. Residential networks are also created in Delhi-Network-V3. The only difference between the input for the two networks are the activity survey files, as shown in table 6.3 and 6.4. We have collect activity survey from 8484 respondents in 2366 households. The classification tree are utilized to retrieve activity templates from this well formatted survey data set, and to assign activity templates to the synthetic population. Different from Delhi-Network-V1 and Delhi-Network-V2, the activity templates are collected for households, not for individuals. It means that, we utilize the classification tree to match synthetic households against surveyed households, and assign activity templates for all members in the synthetic households once and for all. Details of the V3 method can be found in Chapter 5.

**6.3 Comparison of Static Structure and Epidemic Dynamics of Three Networks**

Even though the three networks are modeled for the same region, they differ largely due to the differences in data and methods. We focus on two basic questions: (i) How do we compare several synthetic networks that span the same region? and (ii) When are the details contained in the synthesized network adequate?

The first question is relevant to the comparison of the structural properties of different networks. Given that we have a first principle approach to network synthesis, we compare the networks at various levels. (1) We start from node-based and set-based information (labeling structure), which includes household structure, infrastructure location layout and individual activity; (2) We also validate the detailed network structure against collective statistics, e.g., generated individual demographics are matched to meet the surveyed demographic distribution; (3) We study the network level features, including network structural properties and entropy; (4) Finally, we reach the epidemic dynamics level: policy effects on the networks. We argue that these metrics matter to the network construction model because it pertains to several dimensions of network quality, including accuracy, precision, and impact on simulated epidemics. We emphasize here that, unlike a traditional approach such as using ERGM models, realistic social contact networks are irregular, unstructured and dynamically changing, and therefore difficult to measure or describe with any single metric. Only when a network outperforms another in adequate metrics will we say it has a higher quality. Table 6.6 lists the novelty of our approaches in contrast to traditional approaches.

The second question is fundamental in the area of computer modeling. “The heterogeneity of the natural world ensures that any model that strives for realistic representation must contain large quantity of empirical input” [112]. However, we cannot incorporate all the
Table 6.6: Comparison of our analysis approaches to traditional approaches.

<table>
<thead>
<tr>
<th>Dissimilarity in</th>
<th>Traditional Approach</th>
<th>Novel Approach</th>
<th>Major Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>static structural properties</td>
<td>described with simple models: homogeneous mixing; scale free networks; etc.</td>
<td>described with multidimensional features: household structure, geo-spatial distribution of population, gyration radius of people, temporal feature of people-location graphs, age mixing pattern, etc.</td>
<td>Real social contact networks contain clear yet complex structures which cannot be described with a single theoretic model. For example, while people-location graphs show a clear power law like degree distribution, people-people graphs have a uni-peak distribution.</td>
</tr>
<tr>
<td>epidemic dynamics</td>
<td>differential equations; deduced from simple network structural properties</td>
<td>agent-based large scale simulations of epidemics and interventions</td>
<td>People’s behavior and demographic structure have significant impact on epidemic dynamics. The different output between the two networks demonstrates that details contained in Delhi-Network-V1 are not adequate.</td>
</tr>
</tbody>
</table>

details into the model. It is important to know when new input ceases to improve the final results. At that point, we can claim that the details contained are adequate. To build a realistic social contact network and use it to predict epidemics and intervention effects, we propose two criteria of adequacy. First, does the new input decrease the uncertainty in the system? Reduced uncertainty improves the precision of the final results. Second, does the new input significantly change the simulation results of epidemics and intervention policy? If yes, then this means that the new detailed structure needs to be investigated further. If the answer is no then it means that the additional details were probably not needed for the specific policy questions at hand.

6.3.1 Measuring the Features of Synthetic Components

Synthetic components refer to synthetic population, existing infrastructure and their layout. The quality of those components directly impact the accuracy of the constructed social contact networks.

Population

We create two synthetic populations for the three networks, wherein Delhi-Network-V2 and Delhi-Network-V3 use the same underlying synthetic population. Both synthetic populations are built based on India census 2001 data, therefore, their similarity is high. However, the households in Delhi-Network-V1 are generated by randomly placing synthetic individuals
together with the constraint of household size distribution. The population used in Delhi-Network-V2 and Delhi-Network-V3 is generated using IPF based on the surveyed household samples (see Section 6.2). The distribution of household structure is shown in Figure 6.1. Assuming samples are not biased too much, household structure in Delhi-Network-V2/Delhi-Network-V3 should be more realistic, because it is calibrated with real world data. For example, the portion of single-parent-with-children family is unusually high in Delhi-Network-V1 (or V1 in abbreviation), which is unlikely in a traditional society in India. This is corrected to a low value in Delhi-Network-V2/Delhi-Network-V3 (or V2/V3 in abbreviation).

The members in a household of V1 are gathered randomly; in V2/V3 they are put together based on the variety of member demographics characterized by household samples. Obviously, the random space of the synthetic population for V2/V3 is reduced greatly. To demonstrate this, we calculate entropies for the two combinations of households, where we estimate the probability of a household structure as its percentage of all types of households. The entropy is 1.37 for V1, 1.02 for V2/V3, indicating reduced uncertainty in the latter.

**Location Layout**

Locations are where people conduct their activities. They determine how people are distributed in the geographical space of the region. Delhi-Network-V2/Delhi-Network-V3 uses the MapMyIndia [87] location database, a high resolution dataset, to reconstruct synthetic locations. This database contains land use statistics in the NCT of Delhi, including the coordinates for multiple types of real locations where people work, study, shop and entertain. So the synthetic locations in V2/V3 are accurate. However, such data is costly and not always available for all regions. V1 presents a generic model using the globally available geospatial database LandScan. In Figure 6.2 the layout of each set of locations is visualized to show the difference.
Figure 6.2: The layout of synthetic locations in the three networks. The two layouts are generally similar but differ in details. In the peripheral area, the locations are located around streets in 6.2b, but such textural features are missing in 6.2a due to modeling precision: LandScan data used in 6.2a only contains a $30'' \times 30''$ resolution.

Activity Statistics

_Delhi-Network-V1_ makes use of generic survey templates, with the total time allocated to different types of activities calibrated with locally sampled data (section 6.2). In contrast, _Delhi-Network-V2_ makes use of real survey data from another Indian city, Thane. While it does not capture city specific activities, the survey does capture important details pertaining to the activity sequence. _Delhi-Network-V3_ uses demographics and a mobility survey for residents right in Delhi, so the input is of the highest quality. The different sources of the data sets seems to produce totally different activities in the three networks. Here, we pick three metrics of the activity statistics to show the different patterns in the three synthetic networks.

For the first, we measure the typical duration of different types of activities. Figure 6.3 shows some interesting patterns between the three networks. _Delhi-Network-V1_ has large duration variation, with values frequently ranging from 1 to 20 hours. On the other hand, _Delhi-Network-V2_ contains highly clustered duration values: the most frequent duration for home is around 16-18 hours, 8-10 for work and 5.5 hours for school activities. The pattern of _Delhi-Network-V3_ is similar to that of _Delhi-Network-V2_, only that the variation of duration is a little bit larger than that for _Delhi-Network-V2_: the most frequent home duration is 12-17 hours, 8-10 for work and 5-6 for school activities.

The second metric is the temporal degree of typical locations. Real social networks are dynamic; even simple measures such as degree distributions are temporal functions. In Figure 6.4, we select three typical types of location: work place, home location and schools. For each type of location, we randomly select five, then plot their temporal degree. The three networks show some similarity visually, e.g., the peak degree for work locations or schools
Figure 6.3: Duration of several typical activities: home, work and school. The different patterns in the three networks are significant: the variation of durations in *Delhi-Network-V1* is large and the frequent duration is scattered from 1 to 20 hours; in *Delhi-Network-V2* durations are centered around a specific value for all three activities. *Delhi-Network-V3* is similar to *Delhi-Network-V2*, but the variation of duration is slightly larger.

are around noon, while for home locations, it is in the morning and at night, which is very consistent with our experience. Nevertheless, we observe some major differences between the three networks. The degree value changes sharply in *Delhi-Network-V1*, but remains stable and changes slowly in *Delhi-Network-V2* and *Delhi-Network-V3*. This shows that people in V2/V3 are involved in an activity within a period roughly the same as others conducting the same activity, while for V1 variance of activity time is larger. V2 and V3 reveal high similarity, but differ greatly in school activities. In India, elementary school students and middle school students share the same campus but take classes at different times, one set in the morning and the other in the afternoon. *Delhi-Network-V2* models the situation so we see that school activity distribution contains two phases, one group of activities which happen before 12:30 pm and the other group of activities of middle school students happens after 12:30 pm. *Delhi-Network-V3*, although it uses a survey taken in Delhi, does not discern between the two groups, thus the difference.

The last metric we calculate for activity statistics is the average number of activities a person conducts over a typical day. For *Delhi-Network-V1*, the number is 5.3, while for *Delhi-Network-V2* it is 2.1, and 2.45 for *Delhi-Network-V3*. That means the activity schedule in *Delhi-Network-V1* is busier, and each individual on average will meet more people.

**Age-group Mixing Pattern**

The age-group mixing pattern is another interesting metric, which reveals how people interact with each other. In the Figure 6.5, we see that the mixing patterns in three synthetic
Figure 6.4: The temporal degree of important POIs. Each figure plots the temporal degree varying over 24 hours for 5 randomly selected locations, where x-axis shows the time in hours, and the y-axis shows the number of people at that time x. The first row (7.2d to 7.2f) depicts three types of locations in Delhi-Network-V1, the second row depicts locations for Delhi-Network-V2, and the third row those of Delhi-Network-V3.
populations are significantly disparate. \textit{Delhi-Network-V1} has strong interaction within each age-group, but very weak interaction between different age-groups. \textit{Delhi-Network-V2} and \textit{Delhi-Network-V3} show strong interaction between age-groups, but they also differ for within age-group mixing: \textit{Delhi-Network-V2} has a relatively weak interaction within the age-group of kids (age 0-5); while \textit{Delhi-Network-V3} has strong within age-group interactions for most age-groups. Our observation is consistent with people’s activity in the three synthetic populations. In \textit{Delhi-Network-V1} and \textit{Delhi-Network-V3}, kids are sent to day care centers where they meet peers of the same age group, while kids in \textit{Delhi-Network-V2} are assigned the same sequence of activities as an adult from the same household having no work related or school related activities (Chapter 4). This explains why kids in \textit{Delhi-Network-V2} have weak inner group mixing but a (relatively) strong inter-group mixing; while kids in the other two networks reveal opposite behaviors.

![Figure 6.5: Age mixing for different versions of Delhi synthetic populations.](image)

\textit{Delhi-Network-V1} has strong interaction within each age-group, but very weak interaction between different age-groups. \textit{Delhi-Network-V2} and \textit{Delhi-Network-V3} show strong interaction between age-groups, but they also differ for within age-group mixing: \textit{Delhi-Network-V2} has a relatively weak interaction within the age-group of kids (age 0-5); while \textit{Delhi-Network-V3} has strong within age-group interactions for most age-groups.

**Mobility Pattern**

People’s mobility pattern is a highly relevant factor to epidemic spread in the society. The above features of synthetic population, location layout and activity statistics, co-evolved together to shape the mobility pattern of people. To measure the difference between the mobility patterns created by the two methods, we propose two metrics: the daily travel distance and the radius of gyration. The daily travel distance is the summation of all trip lengths of a person in a typical day. Assume a person visited \( N \) places on a day (repetitive
visits to a same place are counted as multiple places), then the radius of gyration of the person is defined as:

\[ R_g = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (r_k - r_{\text{mean}})^2} \]

where \( r_{\text{mean}} \) is the mean position of the \( N \) visited places. The travel distance depicts the length of the movement, and the radius of gyration depicts the area covered by a person’s movement. From Figure 6.6, we see that the distribution of the two metrics for all three networks reveal a power law like curve. This is consistent with the findings in the literature. However, the three networks differ in actual values. \textit{Delhi-Network-V1} contains more long distance travels, and the portion of long and short radius gyration is higher than that of V2/V3. The former is consistent with our observation in section 6.3.1 that people in V1 have a busier schedule and travels more frequently.

![Figure 6.6: The travel distance and radius of gyration for people in Delhi within a 24-hour-period. The travel distance depicts the length of the travel, and the radius of gyration shows the area covered.](image)

### 6.3.2 Network Structure Comparison

The differences or errors that reside in the raw data are conveyed to the structure data we created for population and locations, and further propagate to the constructed network. The comparison of the following structural difference of the three versions of networks clearly show the trend for us.

**People-location network** \( G_{PL} \)

The degree distribution of the people-location graph is plotted in Figure 6.7, wherein \textit{Delhi-Network-V1} contains some odd peaks due to data flaws in input. Both \textit{Delhi-Network-V2} and \textit{Delhi-Network-V3} reveals a power law like degree distribution for its location-people
network. As verified frequently in the literature [53], we believe they are more accurate than Delhi-Network-V1.

People-people network $G_{PP}$

The people-people networks $G_{PP}$ are significantly different in terms of degree distribution, edge weight and clustering coefficient as shown in Figure 6.8. Interestingly, we can easily tell the source of differences through our previous analysis. To give an example, the presence of residential networks in Delhi-Network-V2/ Delhi-Network-V3 introduces several differences in the structural properties. First, we assume the people who do not work have no interactions with neighbors in Delhi-Network-V1 and are responsible for the peak of low degree and the high peak on the right side of clustering coefficient distribution. However, since the portion of low degree people gain social contacts modeled from the residential contact survey, those low degree and highly clustered people disappear in V2 and V3.

While the structure of Delhi-Network-V1 is totally different from that of V2 and V3, the difference between V2 and V3 is much smaller. The activity data used in V3 seem only slightly modify the results when compared to V2.

Graphlets are another important feature that have been extensively studied in different network applications [62, 79, 84]. We examine all isomorphism classes for 3-node and 4-node graphlets in terms of topology structure. The featured graphlets are visualized in Figure 6.9, each of which represents an isomorphism class. Please be advised that we deem all edges as reciprocal. The counts of these featured graphlets in the three people-people networks are shown in Figure 6.10. Again, the distributions of V2 and V3 are very similar, while the difference between V1 and V2/V3 is much more significant. From the histogram we can see that V2/V3 contains a majority of graphlets in larger amounts, except for graphlet g4-1 and g4-4. All three graphlets g4-1, g4-2 and g4-4 contain a 4-node-circle, the difference being that graphlet g4-2 is more clustered than the other two. The fact that Delhi-Network-V2/ Delhi-Network-V3 contains more g4-2 but less g4-1 and g4-4 suggests their local connectivity is
Figure 6.8: Structural analysis of people-people contact networks.

Figure 6.9: Visualization for all 3-node and 4-node graphlets. Here each graphlet represents an isomorphism class.

tighter than that of Delhi-Network-V1. Similarly, the number of triangles (g3-1) in Delhi-Network-V1 is less than those in V2/V3, suggesting a similar conclusion, that the clustering coefficients of the latter are generally higher.

Our realistic social networks are not simply random graphs, but contain rich label information regarding demographics and activities. In Figure 6.11, we count the above graphlets with various node labels in the three networks. The corresponding visualizations of those labeled graphlets are displayed in Figure 4.11. Similarly, each graphlet here represents an isomorphism class in terms of rotation relation. As opposed to unlabeled graphlets, Delhi-Network-V1 contains a majority of labeled graphlets with larger amounts. Exceptions are g3-0-0, g3-1-0, g3-1-7, g4-2-0 and g4-5-0. All these graphlets contains all or partial nodes with “kid” labels.

Not so surprising is the high similarity between V2 and V3, the counts of almost all labeled graphlets are very close for V2 and V3. The result, together with previous metrics, verify the two networks resemble each other. Since the two networks are created from the same data sources except the activity survey data, the activity patterns for Thane (used in V2)
6.3.3 Uncertainty Reduction during Network Improvement

We have proposed quite a few metrics for comparing irregular synthetic components during the construction of the networks in Section 6.3.1. Although our analysis shows that we have continuously reduced the uncertainty and improved the quality of those synthetic components from V1 to V3, it does not rigorously prove that the resulting networks Delhi-Network-V1, Delhi-Network-V2 and Delhi-Network-V3 will be improved accordingly. The question is, could we find a network metric to measure the uncertainty reduction of the three networks directly and quantitatively?

Our first choice would be entropy. The concept of entropy comes from thermodynamics,
quantum mechanics, statistical mechanics, and then communication theory and ergodic theory [103, 115, 134]. Entropy is inherently suitable for measuring randomness and uncertainty. In the past decade, a number of different entropy metrics have been studied in the context of complex networks [5, 24, 29, 45]. In [5, 20], the concept of network entropy is introduced for micro-canonical network ensemble and the canonical ensemble. “In classical statistical mechanics, the micro-canonical ensemble is formed by configurations having constant energy $E$, while the canonical ensemble is formed by configurations having constant average energy $<E>$” [5]. For example, the random graph $G(N, L)$ is a micro-canonical network ensemble, which is formed by networks of $N$ nodes with a constant total number of links $L$. The Gibbs entropy per node, in network of $N$ nodes, denoted $\Sigma$, was introduced for micro-canonical network ensembles following a set of statistical constraints:

$$\Sigma = \frac{1}{N} \log \mathcal{N},$$

where $\mathcal{N}$ indicates the total number of networks in the corresponding network ensemble. [5] give an example of calculating the Gibbs entropy for a Chung-Lu network ensemble with a given degree sequence $k_i$ as follows:

$$N\Sigma = \log[(2L - 1)!!] - \sum_i \log(k_i!) - \frac{1}{4} \left(\sum_i k_i^2\right)^2$$

Similarly, the Shannon entropy of a network ensemble was defined in [5] for canonical network ensembles given a number of constraints, mimicking classic Shannon entropy.

However, these network entropies cannot be applied to our realistic networks directly. Unlike canonical or micro-canonical networks, there is not a mathematical formula to describe the constraints of a realistic social contact network. That is why we propose our simulation-based model in this paper. This makes the calculation of network entropy of Delhi networks either infeasible or computationally intractable (given the massive scale of an urban network). The network entropy in the literature does not answer our question about measuring uncertainty reduction.

Therefore, instead of measuring network entropy of realistic network ensembles directly, our method is to measure its uncertainty relative to a reference network ensemble. The reference ensemble could be a micro-canonical network ensemble or a canonical ensemble for which we can calculate entropy directly. A way to quantify the relative uncertainty of a random network is to conduct local edge switches and calculate the number of switches required for it to resemble a reference network [54].

An edge switch randomly selects two edges $(A,B), (C,D) \in E(G)$ and replaces them with $(A,C), (B,D)$ provided that neither of them existed in the first place, as illustrated in Figure 6.12. The edge switch is also called the switch Markov chain [41].

We select the Chung-Lu random graph model [3] to generate the reference network ensemble. A Chung-Lu model generates a random network with a specified degree sequence. Given
a social contact network $S$, we generate a random Chung-Lu network $R$ using $S$’s degree sequence, and take $R$ as the reference network for $S$. To compare the two networks, we run epidemic simulations on them and compare their epidemic size. We do so for Delhi-Network-V1, Delhi-Network-V2 and Delhi-Network-V3 respectively, and plot the comparison results in figure 6.13. The results show that simulated epidemic dynamics do not change in a Chung-Lu network with edge switches. This applies to all three reference networks: Chung-Lu-V1, Chung-Lu-V2, and Chung-Lu-V3. On the other hand, we observe significant changes in epidemic dynamics for the three Delhi networks as the edges are switched. In addition to that, the edge switches make the Delhi networks become closer and closer to their reference networks. This observation suggests that the space of synthetic social contact networks forms a much smaller subset of the whole space of graphs than the space generated by the Chung-Lu model with the same degree distribution. Each edge switch increases the randomness of the Delhi networks, eventually destroying their social structure. In this sense, the number of switches a social network takes to match its reference Chung-Lu network indicates a metric of certainty (away from randomness).

Since the Delhi networks use three different reference networks, it is not completely clear that we can compare their uncertainty based on figure 6.13. Therefore, we normalize the epidemic size on those Delhi networks against their reference networks. The normalized results are shown in figure 6.14. Clearly, Delhi-Network-V1 catches up to its reference network with the smallest percentage of edge switches and contains higher uncertainty than Delhi-Network-V1.

Figure 6.13: Epidemic dynamics under various level of edge switches. Epidemic dynamics in all three Delhi networks are very sensitive to edge switches, while they do not change in the referenced Chung-Lu networks with edge switches.
Huadong Xia Chapter 6. Network Improvement

Figure 6.14: Normalized epidemic dynamics under various level of edge switches. Delhi-Network-V1 catches up the epidemic size in its reference network (normalized as 1) with least fraction of edges, and it has a initial value closest to the reference network as well. Delhi-Network-V2 and Delhi-Network-V3 do not show significant difference with regard to the change speed of epidemic dynamics.

V1 and Delhi-Network-V2. We do not see a significant difference between Delhi-Network-V2 and Delhi-Network-V3, which is reasonable as the improvements from V2 to V3 are about accuracy instead of uncertainty reduction.

6.3.4 Comparison in Network Dynamics and Effective Policy

In conducting an epidemic dynamics analysis, traditional approaches result in analytical solutions, such as differential equations, or deduce the results from network structures. However, as our analysis shows, the network structure is irregular and massive; it is unlikely we could retrieve meaningful results purely from the analytical solutions. Therefore, we resort to distributed agent based systems [16,21] to conduct large scale simulations.

Our simulation aims to provide insights and partial answers to the two questions proposed earlier, whether the three networks differ and whether they contain adequate details.

We run exactly the same simulations with Delhi-Network-V1, Delhi-Network-V2 and Delhi-Network-V3, calibrated with the same $R_0$ value. We listed the results with $R_0$ 1.35 in Figures 6.15, 6.16. The results with other $R_0$ values are omitted, because they are very similar to the results of $R_0$ 1.35 therefore won’t change our conclusion below.

The vulnerability distribution in Figure 6.15 reveals a clear difference between the three networks. The distribution of Delhi-Network-V2 and Delhi-Network-V3 is relatively flat. On the other hand, Delhi-Network-V1 has a sharp bimodal curve, dividing all nodes roughly into two sets of completely different vulnerability. We point out that the bimodal distribution is consistent with our observation of the network structure. The degree and clustering
Figure 6.15: Vulnerability histogram in Delhi-Network-V1, Delhi-Network-V2 and Delhi-Network-V3. We simulate the diffusion of H1N1 and assume $R_0=1.35$. The histogram shows very different structural properties for the three networks.

coefficient distributions of Delhi-Network-V1 reveal a similar bimodal pattern, suggesting the huge variance among nodes of Delhi-Network-V1. It seems a large portion of nodes have a low degree (lower than 10) and a high clustering coefficient (higher than 0.9), both of which contribute to lower vulnerability; the other portion of nodes have a higher degree (higher than 50) and a lower clustering coefficient (lower than 0.4), both of which contribute to higher vulnerability. Vulnerability is a result of both the local as well as the global network structure. Given that local structural effects might cancel out each other in Delhi-Network-V1 and Delhi-Network-V2, a larger portion of shortest paths in Delhi-Network-V1 indicate it is collectively more vulnerable than Delhi-Network-V2. The several structural measures of Delhi-Network-V3 happen to sit between Delhi-Network-V1 and Delhi-Network-V2, indicating a medium vulnerability among the three networks. The attack rate difference in the three networks verifies this hypothesis. As shown in the base case of Figure 6.16, the overall attack rate ordering in the base case is V1, V3, and then V2.

Delhi-Network-V1, Delhi-Network-V2 and Delhi-Network-V3 differ greatly in epidemic dynamics despite the $R_0$ calibration. The results are shown in Figure 6.16, where we use the attack rate to characterize epidemic dynamics. Let’s focus on the base case in the figure. Delhi-Network-V1 has a higher overall attack rate than the other two networks. If we decompose the epidemics into sub-populations, the difference is even clearer. Delhi-Network-V1 has the least vulnerable senior group, and Delhi-Network-V2 has the least vulnerable kids subpopulation of the three networks.
We point out that epidemic dynamics are closely related to the detailed interaction structure of the subpopulations. For example, preschool kids in V2 stay at home (section 6.3.1) and thus belong to those high clustering low degree nodes in the network (Figure 6.8), contributing to their low vulnerability. Kids in V1/V3, on the other hand, are put in day care and have strong connections with a large number of peers. These facts lead to different mixing patterns (section 6.3.1) and different volumes of connections in the three networks. Eventually these detailed structures impact the epidemic dynamics of the subpopulations.

The different subpopulation behaviors in the three networks have strong implications for the effective public health policy. In Figure 6.16, we conduct four different vaccination policies in the three networks, which vaccinate four different subpopulations respectively: a randomly selected subpopulation, kids, students and adults. Using the policy of vaccinating kids as an example, we see that vaccinating kids is least effective in Delhi-Network-V2. Clearly, this is due to the low social activities of the kids in V2. Recalling the entire analysis across the chapter, we can find a chain of analysis from data input (kids’ activity survey), to the detailed structure of the synthetic population (age-group mixing pattern), to the network structure, to the epidemic dynamics on the network, and finally discover the correct (by which we mean most effective) policy.

6.3.5 Summary and Observations

In Table 6.7, we list the metrics computed when comparing the three networks. The results begin to shed light on the structure of the networks. Clearly, the metrics studied here are not exhaustive; an open question to consider is to identify what one would term as critical metrics.

Through the analysis chain from detailed structure to network dynamics to effective policies,
Table 6.7: Measurements used in characterizing the static structure and dynamics.

<table>
<thead>
<tr>
<th>Measurements</th>
<th>Description</th>
<th>Findings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Synthetic System</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HH structure</td>
<td>the combination of individual demographics in a household.</td>
<td>The incorporated household samples decrease the uncertainty in Delhi-Network-V2 (quantified with entropy).</td>
</tr>
<tr>
<td>location layout</td>
<td>the collective distribution pattern of locations over the city, as well as the proximity of locations to infrastructures.</td>
<td>The layout of the locations in Delhi-Network-V2/Delhi-Network-V3 is associated with roads, which is not in Delhi-Network-V1.</td>
</tr>
<tr>
<td>duration of activities</td>
<td>the length of each continuous activity.</td>
<td>The duration in Delhi-Network-V2/Delhi-Network-V3 is highly clustered to a small range, while that of Delhi-Network-V1 is scattered.</td>
</tr>
<tr>
<td>number of daily activities</td>
<td>average number of daily activities for the whole population.</td>
<td>The schedules for people in Delhi-Network-V1 are busier than those in Delhi-Network-V2/Delhi-Network-V3 on average.</td>
</tr>
<tr>
<td>travel distance</td>
<td>the sum of the travel distances of all trips made by an individual on a typical day.</td>
<td>Delhi-Network-V1 has more people with long distance travel.</td>
</tr>
<tr>
<td>radius of gyration</td>
<td>the root mean square distance of an object’s parts from its center of gravity</td>
<td>Delhi-Network-V2/Delhi-Network-V3 contains more medium radius of gyration, while Delhi-Network-V1 contains a higher portion of long and short radius.</td>
</tr>
<tr>
<td>( G_{PL} ) temporal degree of a location</td>
<td>the number of visits of a location during each hour of the day</td>
<td>The variation of temporal degree in two networks is very different.</td>
</tr>
<tr>
<td>degree of people-location graphs</td>
<td>the number of visits to each location</td>
<td>The degree distribution in Delhi-Network-V1 is not well formed due to the flaw in modeling. That of Delhi-Network-V2 and Delhi-Network-V3 follows power-law.</td>
</tr>
<tr>
<td>( G_P ) degree clustering coefficient contact duration shortest path</td>
<td>basic network metrics</td>
<td>The two networks differ significantly in those structural properties. It is possible to infer the relation between modeling procedure to some structural difference.</td>
</tr>
<tr>
<td>Dynamics</td>
<td></td>
<td></td>
</tr>
<tr>
<td>baseline</td>
<td>simulation results without interventions</td>
<td>The two networks have very different dynamics.</td>
</tr>
<tr>
<td>interventions</td>
<td>simulations with strategies: vaccine, antiviral, school closure, work closure</td>
<td>The efficacy of intervention policy might be different in the three networks, meaning the details in Delhi-Network-V1 are not adequate.</td>
</tr>
</tbody>
</table>
we show that detailed structure affects epidemics within a population. This validates our modeling effort in collecting metrics at different levels. It also indicates that it is possible to use detailed structures of a network to guide the design of public health policy.

6.4 Conclusion

Social contact networks play a critical role in understanding epidemic dynamics within a population. However, construction of such a network remains a challenging problem. Social contact networks for a populous urban area often contains millions of nodes and billions of edges. The scale and complexity make it impossible to obtain such networks by simply measuring social interactions. As a result, such networks are often synthesized using a diverse set of real-world data. The synthesis methods can be viewed as a complex stochastic process that outputs one realization of such a network. Two questions naturally arise during the modeling process. First, given two such urban scale networks, how does one capture the similarities and differences between them? Second, when are details modeled in a given social network deemed adequate from the standpoint of the questions at hand?

To study the above questions, we generate and compare three versions of social contact networks for NCT area of India based on different data sources and methodology, referred to as Delhi-Network-V1, Delhi-Network-V2 and Delhi-Network-V3. In answering the questions, we introduce a number of metrics to describe the networks. Some metrics are well studied in the literature while others are new and capture specific aspects of the networks. Our analysis reveals interesting differences between the networks, even though the networks are supposed to capture the interactions in the same urban area. Our analysis also reveals the details necessary for particular class of control policies.
Chapter 7

Comparison and Validation of Synthetic Social Contact Networks across Geographical Regions

HUAĐONG XIA, JIANGZHUO CHEN, MADHAV MARATHE AND SAMARTH SWARUP

Abstract

We describe the synthesis of a detailed social contact network of Delhi, India, for urban-scale epidemiological simulations. The network synthesis is done by combining information from multiple data sources, since social contact information (the set of people a person comes into contact with during a day) cannot be obtained through direct surveys. We compare this network to a previously generated social contact network for the city of Los Angeles, USA, on various structural and dynamical metrics. These networks are comprised of millions of nodes and hundreds of millions of edges.

Through the comparison between the two cities, we show important similarities and differences between urban regions in different parts of the world. Epidemic simulations highlight policy-relevant differences in outcomes. The comparison also serves as a validation of the Delhi social contact network generation methodology, because the LA network generation methodology is a mature technology that has been used in many studies and refined through multiple iterations, and the Delhi network shows sufficiently similar behavior even though it has been constructed using different data sources. The differences we observe are expected and explainable in terms of the differing socio-economic and demographic conditions in the two cities.
7.1 Introduction

This paper describes the construction of large-scale, data-driven synthetic social contact networks for epidemic simulations. In a social contact network, the nodes represent people, and edges represent interactions between people who come into close enough contact with each other to transmit an infection. This is the network over which an epidemic propagates.

This network has a few important properties. It is a time-varying network, since edges only exist while people are in contact. It is also a weighted network, since duration of contact affects probability of disease transmission. Third, perhaps most importantly, it cannot be obtained through direct surveys, because most people do not know all the people they come into contact with during a day.

Our approach, therefore, is to construct an approximation to the social contact network by combining data about demographics, activity patterns, and activity locations. The idea here is that if we build a model of where people are over the course of the day, then we can induce edges between any two people who are at the same location at the same time.

In the present work, we describe the construction of synthetic social contact networks for the cities of Los Angeles, USA, and Delhi, India. These are large populations, and their social contact networks consist of millions of nodes and hundreds of millions of edges. Furthermore the networks are quite irregular and are not well-described by simple network models.

A crucial question, then, is how to assess the quality of the constructed networks. This is broadly a question of validation. However, validation of such complex models is not simply a matter of comparing epi-curves against real data, since it is very easy to match a sequence of numbers of infections by tweaking any of a large number of parameters. Since no model is ever completely correct, validation in practice is a process of building trust in a model through repeated and varied use.

Our approach here, therefore, is to do a detailed comparison of structural and dynamical metrics on the two networks to determine the differences between them. This serves as a kind of cross-validation, since the two networks are generated from different data sources. Additionally, the process for generating the Los Angeles network follows relatively mature technology using high-quality data sources, and has been used in multiple prior studies that have gone through rigorous peer review [15, 53]. By comparing the Delhi network with the
Table 7.2: Demographics, location and activity data used in construction of Delhi-Network-V2.

<table>
<thead>
<tr>
<th>Data</th>
<th>Statistics</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>13.85M population</td>
<td>India Census 2001 [72]</td>
</tr>
<tr>
<td></td>
<td>2.67M households</td>
<td>Household Microdata: India Human Devel. Survey 2005 [46]</td>
</tr>
<tr>
<td>Locations</td>
<td>2.64M non-home locations</td>
<td>MapMyIndia Dataset [87]</td>
</tr>
<tr>
<td>Activity</td>
<td>samples: 14,428 respondents from 3,505 households</td>
<td>2001 Thane Household Travel Survey [102]</td>
</tr>
<tr>
<td></td>
<td>-</td>
<td>2000-07 school attendance statistics from the UNESCO Institute of Statistics (UIS) [128]</td>
</tr>
<tr>
<td></td>
<td>37 samples</td>
<td>India residential area activity survey by us</td>
</tr>
<tr>
<td>Resultant Network</td>
<td>13.85M nodes, 207M edges</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 7.3: Demographics, location and activity data used in construction of the Los Angeles network.

<table>
<thead>
<tr>
<th>Data</th>
<th>Statistics</th>
<th>Sources</th>
</tr>
</thead>
<tbody>
<tr>
<td>Demographics</td>
<td>16.23M population, 5.41M households</td>
<td>US Census Bureau [130]</td>
</tr>
<tr>
<td>Locations</td>
<td>0.67M non-home locations</td>
<td>Dun &amp; Bradstreet [1]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NCES [2]</td>
</tr>
<tr>
<td></td>
<td></td>
<td>NAVTEQ [101]</td>
</tr>
<tr>
<td>Activity</td>
<td>-</td>
<td>National Household Travel Survey [109]</td>
</tr>
<tr>
<td>Resultant Network</td>
<td>16.23M nodes, 459M edges</td>
<td>-</td>
</tr>
</tbody>
</table>

Los Angeles network, we can build trust in the Delhi model if the differences between it and the Los Angeles network are explainable in terms of cultural and demographic differences between the two cities.

We describe a number of metrics for comparing the two networks. Those metrics can be divided into four different levels, from node level, to the level of the network. Please refer to Chapter 1 (Table 1.1) for the list of the metrics. We have utilized them in Chapter 5 to compare different versions of Delhi networks.

We proceed with this paper as follows. First we describe the methodology for generating the two networks. This is followed by a detailed comparison of demographics, activity patterns, network structural measures, network dynamical measures, and effects of interventions.
7.2 Network Generation: Data and Methodology

In this section, we first review the data sets used, and then introduce the methodology for constructing the networks for the two cities.

7.2.1 Data

We use Delhi and NCT interchangeably to refer to the National Capital Territory (NCT) of India. It is the capital of India, including New Delhi and several adjacent urban areas. NCT and Los Angeles are two of the most populous urban areas in the world, having populations of 13.85M and 16.23M respectively. Details regarding the two cities are summarized in Table 7.1.

Multiple data sources are used to construct the two contact networks. The data we use to construct the networks are listed in the “Sources” column of Table 7.2 for Delhi-Network-V2 and Table 7.3 for the Los Angeles network.

There are several differences between the two data sets. First, subjective surveys regarding individual behavior in residential areas are used to synthesize Delhi-Network-V2 but not for the Los Angeles network. This is because about 40% of the population in Delhi do not travel on a daily basis and have activities mainly in their home and neighborhood. Second, some data sets used to synthesize the Los Angeles network are more accurate than the ones used to construct Delhi-Network-V2, for example the activity survey data. In Delhi-Network-V2, no activity survey is available for the Delhi population. Therefore, a travel survey for another Indian city, Thane, is utilized.

As shown in Tables 7.2 and 7.3, both networks utilize data from diverse sources; the input data are often noisy and unstructured. All these features impose challenges in network synthesis and analysis.

7.2.2 Methods

In this section, we provide an overview of the two sets of methods to generate synthetic populations and networks. The methods are based on our earlier work [15], and consists of the following broad steps: (i) synthesize a baseline population with a detailed individual structure and the same aggregate statistical properties of the real population; (ii) assign each individual a reasonable activity schedule based on mobility survey; (iii) create locations in the region where synthetic people can take their activities. In the following, we describe the two methods respectively with emphasis on the differences.
The Los Angeles network

In prior work [15], we presented a method to generate a synthetic population and contact network for cities in the United States, including Los Angeles. We briefly describe the procedure in the following.

First, a synthetic population is generated from census data based on disaggregate details and aggregate statistics. The US Census divides the country into Public Use Micro Areas (PUMAs). Within each PUMA, a set of household samples representing 5% of all households is collected [46]. The synthetic population captures details of household structure by utilizing the 5% Public Use Micro Sample (PUMS) for each PUMA modeled. Assuming the surveyed household samples are representative in PUMAs, any household in the real population can be estimated with a carefully selected sample in terms of its household size and household members. We then replicate the selected samples to create all Los Angeles households. The family members in those synthetic households naturally compose the LA population.

During the procedure, the sample selection is critical. Our selection of the samples follows the joint distributions of household level demographic variables such as householder’s age, household income, and household size. The marginal distribution of each variable is provided in the census data, but such a joint probability table is usually unavailable, so we apply the iterative proportional fitting (IPF) procedure to estimate them [18]. IPF is an iterative algorithm for estimating cell probabilities in a contingency table such that the marginal totals remain fixed.

Next, we generate an activity sequence for each person. An activity sequence is a set of major activities a person conducts on a typical day. Each activity includes an activity type, a start time, a duration, and a location. To retrieve the most representative activity sequence for each individual, we use data from the US National Household Travel Survey (NHTS) [109]. The survey captures the interdependence of people’s activities, especially adults in the same household across all surveyed households in the United States. Those activity sequences are associated with the sample demographics, and they are assigned to synthetic individuals.
according to their demographics.

We also collect land use information to model the locations where people conduct their activities. Each household is associated with a home locations. Each activity for any individual is also associated with a location. When assigning a daytime location, distance from home to the location is considered. A daytime location with distance $d$ is selected randomly with probability following an exponential distribution: $f(x; \lambda) = \lambda e^{-\lambda x}$ where $\frac{1}{\lambda}$ is the mean distance. Such a distance model is called a gravity model.

Finally we synthesize the contact network. The activity sequences and the associated locations define a labeled person-location (PL) bipartite graph $G_{PL}$, where people and locations are the nodes and there is an edge between person $P$ and location $L$ if $P$ visits $L$; the time of the visit is the label associated with the edge. The person-person contact network $G_P$ is constructed from the PL graph as follows. We assume each activity location is divided into sublocations. The sublocation model is a way of defining interactions among persons who visit the same location at the same time and are in spatial proximity. Each person is assigned to a sublocation (within the activity location) randomly. There is an edge between two people if they are in the same sublocation at the same time.

\textit{Delhi-Network-V2}

The method for constructing \textit{Delhi-Network-V2} is generally consistent with the method for constructing the Los Angeles network. The differences are described below.

In \textit{Delhi-Network-V2}, the activity sequences are not directly collected from the activity survey. Instead, a travel survey for residents in Thane\textsuperscript{1} is utilized. The activities of the residents are inferred based on their travel types and frequencies. This is different from the Los Angeles network where a nation-wide household activity survey is available.

The other major difference between the two methods is the creation of the residential network. People who do not have a formal job and stay at home are assumed to interact only with their family in the Los Angeles network. However, social contacts do exist among those people when they walk around their residential area, even though the interactions might not be as intensive and extensive as those in school and work places. In Delhi, about 40% of the population stay around home and do not have a regular daily travel schedule. Therefore their activities around residential areas cannot be neglected as for Los Angeles. We conducted a survey in Delhi and several other cities nearby, collecting data regarding these “at-home people”. Since these people claim they do not travel, we assume they are in contact only with those people within their own locality. The survey gives us the typical number and duration of contacts for people in different demographic groups. With this information, we then use a configuration model to randomly generate new contacts for each person in the

\textsuperscript{1}Since both Thane and Delhi are Indian cities and they share a lot in terms of cultural and economic features, we expect behaviors of residents in the two cities to be similar.
community. Those contacts form a contact network which we call the *residential network*. We then incorporate it into *Delhi-Network-V2*.

### 7.3 Comparison of the Two Cities

We focus on two basic questions: (*i*) how to compare two synthetic networks that represent two different regions and (*ii*) how to validate the construction of such large scale networks.

We compare the networks at various levels. We study the demographic attributes (labeling structure) as well as static structural properties and dynamical outcomes. We emphasize here that, different from traditional approaches like ERGM models, realistic social contact networks are irregular, unstructured and dynamically changing, therefore difficult to measure or describe with any single metric. “The heterogeneity of the natural world ensures that any model that strives for realistic representation must contain large quantity of empirical input” [112].

#### 7.3.1 Population-level Comparison

**Population**

The population is created using Census 2001 data in both cities. Figures 7.1a, 7.1b, and 7.1c compare various age groups and household structure of the two populations.

We see significant differences between the two populations. Delhi has a much younger population but larger household size in general. The distribution of household structure is shown in Figure 7.1c, where we see that the proportion of single-person, couple-without-children, and single-parent-with-children families is much higher in the Los Angeles network, which we believe are reflective of cultural differences between the two regions, and may impact social contact structure in the two cities.

**Activity Statistics**

Synthetic activities are extracted from survey data of the real populations, which capture region specific details pertaining to the two cities.

Activity patterns in the two cities are quite different. Here, we pick two metrics of the activity statistics to show the resulting differences in the two synthetic networks.

The first metric is the temporal degree of typical locations. Real social networks are dynamic, even simple measures such as degree distributions are temporal functions. In Figure 7.2, we select three types of locations: work, home, and school. For each type of location, we
randomly select five locations, then plot their temporal degree. The two networks show some similarity visually, e.g., the peak degree of work locations or schools are around noon, while for home locations, it is in the morning and at night, which is very consistent with intuitive expectations. Nevertheless, we observe some major differences between the two networks. The degree value changes sharply in *Delhi-Network-V2*, but remains stable and changes slowly in the Los Angeles network. This shows that people in *Delhi-Network-V2* are involved in an activity within a period roughly the same as others conducting the same activity, while for the Los Angeles network variance for activity time is larger. *Delhi-Network-V2* reflects clear region-specific features for school activities, where primary school children attend classes in the morning and middle school students attend classes in the morning and middle school students attend classes in the afternoon. Therefore we see a clear shift at noon. This is very different from what happens in Los Angeles and other cities in the world.

The second metric we calculate for activity statistics is the average number of activities a person conducts over a typical day. For the Los Angeles network, the number is 5.4 and for *Delhi-Network-V2* it is 2.1. That means the activity schedule in the Los Angeles network is busier, and each individual on average will meet more people. Please note that the difference comes mainly from the input data rather than the models.

**Mobility Pattern**

People’s mobility pattern is a highly relevant factor for epidemic spread. Demographics, activity patterns, and geographic locations in a city together shape mobility patterns of people. To measure the differences in mobility pattern introduced by the two methods, we propose two metrics: the daily travel distance and the radius of gyration. The daily travel distance is the summation of all trip lengths of a person in a typical day. Assume a person visited $N$ places on a day (repetitive visits to a same place are counted multiple places), then the radius of gyration of the person is defined as:

$$R_g = \sqrt{\frac{1}{N} \sum_{k=1}^{N} (r_k - r_{mean})^2}$$

where $r_{mean}$ is the mean position of the $N$ visited places. The travel distance depicts the length of the movement, and the radius of gyration depicts the area covered by the person’s movement. From Figure 7.3, we see that the distribution of the two metrics for the two networks both reveal a heavy tail trend. This is consistent with the findings in the literature. However, the two networks differ in the actual values. The Los Angeles network contains more longer distance travel, and the proportion of long gyration radius is higher than that of *Delhi-Network-V2*. The former is consistent with our observation in section 7.3.1 that people in the Los Angeles network have a busier schedule and travel more frequently.
Figure 7.2: Temporal degree of important POIs. Each figure plots the temporal degree varying over 24 hours for 5 randomly selected locations, where $x$-axis shows the time in hours, and the $y$-axis shows the number of people at that time $x$. The first column (7.2a, 7.2c and 7.2e) depicts three types of locations in *Delhi-Network-V2*, and the second column depicts for the Los Angeles network.
Figure 7.3: Distribution of travel distance and radius of gyration. This plot is for people in Delhi and Los Angeles within a 24-hour-period. The travel distance tells the length of the travel, and the radius of gyration tells the area covered by the travel.

7.3.2 Network Structure Comparison

Person-location networks $G_{PL}$

The degree distribution of the people-location graph is plotted in Figure 7.4, wherein both networks reveal a power law like degree distribution. This large-scale structure is documented frequently in the literature [53].

Figure 7.4: Degree distribution for people-location networks.

Person-person networks $G_{PP}$

The person-person networks $G_{PP}$ are significantly different for both global features such as shortest path distribution or local structure such as degree distribution, edge weight or clustering coefficient (Figure 7.5). Some differences seem counter-intuitive, e.g., we might
expect that Delhi should have higher edge density, given its higher population density than Los Angeles, but the degree in Delhi-Network-V2 is lower in general. We argue that the social contact density is different from the population density. Actually, these structural differences can be traced back to the different activity patterns of the two populations. People in Los Angeles conduct more activities on average and each activity takes shorter time compared to people in Delhi people. Further, a significant proportion of people in Delhi stay close to home for the entire day. Such activity characteristics introduce higher average degree in $G_{PP}$ of Los Angeles and lower clustering coefficients. Consequently we also see smaller shortest paths for Los Angeles. All these structural features imply epidemics may spread faster and easier in the Los Angeles network.

Graphlets are another important feature that have been extensively studied in different network applications [62, 79]. We consider labeled graphlets to understand the micro structure within the two networks. Graphlet label refers to the properties of its nodes or edges, such as demographics and activities. In Figure 7.6, we count 3 and 4-node graphlets with various node labels in the two networks. The corresponding visualization of those labeled graphlets is shown in Figure 4.11. Each graphlet here represents an isomorphism class in terms of rotations.

Since the Los Angeles network contains a larger number of edges and nodes, it is not surprising that it contains a higher count of most labeled graphlets. Exceptions are g3-0-0, g3-1-0, g3-1-7, g4-2-0 and g4-5-0, whose counts in Delhi-Network-V2 exceed the counts in the Los Angeles network. All these exceptions either contain only “child” nodes, or “child” and
“adult” nodes. The child subpopulations in both networks are comparable (Figure 7.1a), so the exceptions are not incurred due to the size of child populations. Actually the Los Angeles network still contains a higher number for the graphlets that contain both children and other people (seniors and teenagers). We infer that children in Delhi-Network-V2 do not mix extensively with people from other age groups. Consequently, the subpopulation of Delhi tends to stay “safer” during epidemic spreading, which is verified in simulations below.

The Los Angeles network peaks at graphlets g3-0-5, g3-1-5, g3-1-9, g4-0-1, g4-0-2, g4-0-3, g4-2-2 and g4-5-1. All these graphlets contain all or partial nodes with ‘adult’ label. However, adult nodes do not interact with other adult nodes in half of those graphlets, depicting a subpopulation with low clustering, which is consistent with Figure 7.5.

7.3.3 Comparison of Epidemic Dynamics

In conducting an epidemic dynamics analysis, traditional approaches result in analytic solutions, such as differential equations or deduce the results from the network structures. However, as our analysis shows, the network structure is irregular and massive; it is unlikely we could retrieve meaningful results purely from the analytical solutions. Therefore, we resort to distributed agent based tools [16,21] to conduct large scale simulations.

We run exactly the same simulations with the Los Angeles network and Delhi-Network-V2, calibrated with the same $R_0$ value. We listed the results with $R_0 = 1.35$ in Figures 7.7, 7.8 to compare against Delhi-Network-V2. The results with other $R_0$ values are omitted, because they are very similar to the case of $R_0 = 1.35$ here.

The vulnerability distribution in Figure 7.7 reveals a clear difference between the two networks. The vulnerability of a node is the probability that it will get infected. It is a result of both the local as well as the global network structure. Given that local structural effects...
might cancel out each other in the two networks, a larger portion of shortest paths in the Los Angeles network indicate the Los Angeles network is more vulnerable than Delhi-Network-V2 collectively. The attack rate difference in the two networks, as shown in the Figure 7.8, verify this hypothesis. On the other hand, the right plot in Figure 7.7 shows the reduction percentage of attack rate after vaccinating top 10% most vulnerable nodes in each city. We compare the results under different vaccination compliance. Here compliance means the probability a person will take vaccines after he/she is asked to do so. We see that for all non-zero compliance values, the reduction effect in Delhi is more significant than that in Los Angeles. The trend is clearer as the compliance increases. This reveals that the top 10% most vulnerable nodes in Delhi have a larger impact to neighborhood in disease diffusion and therefore vaccinating them achieves better results.

The Los Angeles network and Delhi-Network-V2 differ a lot in epidemic dynamics despite the $R_0$ calibration, as shown in Figure 7.8, 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 policies, the Los Angeles network has a much higher attack rate, higher peak and earlier outbreak dates than Delhi-Network-V2, conforming to our observation earlier.

Further analysis shows that the internal interactions between subpopulations are very different in the two networks (Figure 7.9). For example, in the Los Angeles network, preschool children are a little bit more vulnerable than average (the red line), but in Delhi-Network-V2, they are the most resistant subpopulation. The difference comes from the different structural role the subpopulation plays in the two networks as we show in the graphlet analysis. Further analysis shows that most preschool children in Delhi-Network-V2 stay at home and thus correspond to high clustering low degree nodes, contributing to their low vulnerability. In addition, the school children are the most vulnerable subpopulation in both networks. These structural features may help us in designing effective intervention strategies.
Figure 7.8: Epidemics and policy efficacy in the Los Angeles network and Delhi-Network-V2. We simulate the diffusion of H1N1 and assume $R_0=1.35$. The two networks show significantly different epidemic dynamics, despite the calibration of $R_0$ to 1.35 for the two networks.

We design four different intervention strategies, including two pharmaceutical interventions, Antiviral and Vaccinating Kids (VacKids), and two non-pharmaceutical interventions, School Closure and Work Closure. The simulation results are shown in Figure 7.8.

In general, the pharmaceutical interventions are more effective in reducing attack rate, but non-pharmaceutical interventions help to better alleviate the intensity of the outbreak, meaning that school closure and work closure reduce the peak value and postpone the peak day more significantly. School closure is more effective than work closure since the school age group is the most vulnerable subpopulation in both networks (Figure 7.9).

### 7.4 Validation

In addition to the work presented here, we have made extensive efforts to validate our models; see [19] for an in-depth discussion. This includes: (i) data validation: matching diverse measured data sources, such as properties reported in the census, traffic, and people’s activities, (ii) functional validation: ensuring the synthetic network is based on accepted social theories and data integration techniques, and (iii) structural validation: we use approaches from statistical physics and complex systems to show several emergent properties are consistent with observed system-level phenomena.

**Data Validation.** The population mobility and social network synthesis modules in our methods guarantee (i) the synthetic population is statistically identical to the original census data, (ii) the locations are as reported in commercial location databases (e.g., the Dun & Bradstreet database), the activities and trips undertaken by individuals is based on NHTS survey data collected by transportation authorities, (iii) population dynamics produced by our models are matched including activities and population densities in the city, number of people in various locations at different time of the day, etc.
**Functional Validation.** Whenever possible, our models are developed based on well accepted social, economic and behavioral theories. This is particularly important when developing behavioral representations and social interactions. These theories are used as representations of collective knowledge and agreement by experts on basic elements of any model used to represent individual or collective behavior. Specific theories we have used include: (i) Homophily theory [90] used to develop the Delhi residential network; (ii) Proximity theory [28] used for the assignment of locations for individuals and activities, which turns into a variant of gravity model.

**Structural Validation.** We show (i) the structural properties of urban population dynamics, such as temporal location degree, were measured and found to be consistent and accurate; (ii) structural properties of the social networks are consistent with reported measurements in [52,53] in terms of power law degree distribution of locations; (iii) the distance traveled by individuals has been found to follow a heavy tail distribution in our study, consistent with the distribution of displacements of individuals, which is reported by Gonzalez et al. [62] to be approximated by a truncated power law distribution. In addition, the two cities’ networks in this paper are constructed independently and based on different data sets. Therefore, the similarity between the two networks may serve as a cross validation of each other. Note that all of the above properties are not explicitly added to the model as input but rather a model outcome and in this sense “emergent”.

![Figure 7.9: Subpopulation epicurves of Los Angeles and Delhi. Epidemic curves show subpopulation epidemics in Delhi-Network-V2 (left) and the Los Angeles network (right) when $R_0 = 1.35$. Both populations are 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 7.8, where there is no intervention. The red curve shows the fraction of people in the whole population infected on each day in the same base case.](image)
7.5 Conclusion

Social contact networks play a critical role in epidemic dynamics. The network for a populous urban area may contain millions of nodes and hundred of millions of edges. It is impossible to obtain such networks by simply measuring social interactions. As a result, we are developing techniques for synthesizing these networks using a diverse set of real-world data.

To cross-validate the methodology, we generate and compare two social contact networks for the NCT area of India and Los Angeles, USA, based on different data sources and methodology. We introduced a number of metrics to describe the networks. Some metrics are well studied in the literature while others are new and capture the specific aspects of the networks. Our analysis reveals interesting differences and similarities between the networks.
Chapter 8

Concluding Remarks

Here we have studied and demonstrated how to retrieve the social contact network structure for an urban area. The scale and complexity make it impossible to obtain such networks by simply measuring social interactions. As a result, such networks are often synthesized using a diverse set of real-world data. The synthesis methods can be viewed as a complex stochastic process that outputs one realization of such a network.

Different network synthesis methods may generate networks with significantly different quality and structure for the same urban population. This leads us to study the problems of network analysis, comparison and evaluation. We introduce a number of metrics to analyze and measure the networks, revealing interesting differences between them, even though the networks are supposed to capture interactions in the same urban area. Our analysis also reveals the details necessary for a particular class of control policies, which we use as the adequacy criteria in modeling improvement.

An immediate application of a synthesized social contact network is to study the epidemic dynamics of a population. We use a high performance agent-based modeling environment to study how an influenza-like illness would spread over the Delhi network. We also analyze well understood pharmaceutical and non-pharmaceutical containment strategies to control a pandemic outbreak. These results help us better understand the collective behavior of a massive complex system like a populous urban area, and also provides feedback to guide synthetic network modeling.

Our methodology is entirely general and can be applied to other areas of network science. Our models can be easily modified to study other reaction diffusion processes, including emotion infection, the spread of rumors, financial crisis diffusion, and many others.

Nevertheless, our methods are limited in several aspects. First, we have not studied long distance travel here. We assume all people travel within the city, yet there are inter-city and inter-country travelers as well. Barabasi et al. [62,124] and Vespignani group [39] have studied the human mobility pattern of longer distances and we should incorporate their
Second, we have not studied adequately how individuals move within a building location. We have used simple random graph models to abstract this. However, the author’s other work [140, 141] has suggested the detailed contact structure within office buildings or classrooms may have a significant impact on network dynamics. Modeling of the network at a finer level is expected, and we can make use of the mature models of school networks [116, 120, 140, 141], hospital networks [73], and others.

Third, we have assumed people’s activities are conducted on normative days; representations that have weekly, monthly and yearly patterns are not studied here. Mao [86] has suggested these patterns cannot be neglected. In addition, important gatherings, including parades, flash mobs, and sporting events, will have an unusual influence on network structure and should be studied in the future.

Lastly, we have not investigated much in terms of use of social media data, as well as data that is collected by wireless devices [17, 62, 120]. Work in [131] indicates that online social media augments offline network capital and we should make use of that in our models.
Bibliography


