10-23-2014

Synthesizing an Agent-Based Heterogeneous Population Model for Epidemic Surveillance

Madiha Sahar
Purdue University, Main Campus, msahar@purdue.edu

Nadra Guizani
Purdue University - Main Campus, nguizani@purdue.edu

S Besalamah
Umm Al-Qurra University, SA.

M.N. Ayyaz
Univeristy of Engineering & Technology, Lahore, Pakistan

M. Ahmad
Institute of Public Health

See next page for additional authors

Authors
Madiha Sahar, Nadra Guizani, S Besalamah, M.N. Ayyaz, M. Ahmad, T. Mustafa, and Arif Ghafoor

Synthesizing an Agent-Based Heterogeneous Population Model for Epidemic Surveillance

Madiha Sahar [1], Nadra Guizani[1], S. Basalamah,[2] M.N. Ayyaz[3], M. Ahmad,[4] T. Mustafa[5] and A. Ghafoor[1]

***Abstract*—In this paper we propose a probabilistic approach to synthesize an agent-based heterogeneous population interaction model to study the spatio-temporal dynamics of an air-born epidemic, such as influenza, in a metropolitan area. The methodology is generic in nature and can generate a baseline population for cities for which detailed population summary tables are not available. The joint probabilities of population demographics are estimated using the International Public Use Microsimulation Data (IPUMS) sample data set. Agents, are assigned various activities based on several characteristics. The agent-based model for the city of Lahore, Pakistan is synthesized and a rule based disease spread model of influenza is simulated. The simulation results are visualized to analyze the spatio-temporal dynamics of the epidemic. The results show that the proposed model can be used by officials and medical experts to simulate an outbreak.**

***Index Terms*—Demographic Domain Knowledge, Agent based graph, Spatio-temporal dynamics, Rule based simulation.**

I. INTRODUCTION

Increased global connectivity enables migration of microbes, leading to drastic changes in the pattern of global health and disease. To protect human lives as well as strengthen the national security, a nation in collaboration with its international partners must be prepared for and be able to respond quickly to localized and global events of such diseases. Coordinated global surveillance is necessary for early detection and taking appropriate actions to protect the population from such pandemics.

In essence, peoples' health around the world are more closely linked than ever before. Epidemics of novel re-emerging infectious diseases can quickly spread globally through various avenues. Greater movement of people and products can increase exposure to potential health risks originating outside a country. An epidemic that begins in a single community can quickly evolve into a multinational health crisis that can spark major disruptions to travel and trade. Estimates of deaths from the 1918 Spanish influenza pandemic was estimated to have reached 21 million to a 100 million. Similarly, the 1968 Hong Kong pandemic was estimated to have killed between 1 million to 4 million people [1]. With the recent spread of the H5N1 influenza virus, the World Health Organization (WHO) described the world's status as in Phase 3 (Pandemic Alert Period). Thousands of human cases were confirmed with H5N1 and hundreds of deaths [2]. The recent hit of the Ebola virus in West Africa confirms that pandemics are still a threat even in the 21st century.

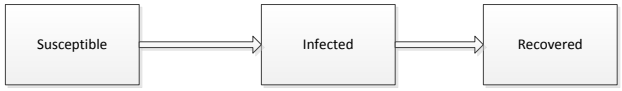

Fig. 1: State Diagram of the SIR Model

Addressing these transnational risks requires advance preparation, and extensive collaboration with the global community. In the specific case of a pandemic, it is estimated that 30% of the US population could become ill, this includes first responders, health care professionals, and policy makers [3].

We propose a methodology to synthesize baseline population by computing demographic attributes of persons/agents for socio-economic groups in a city. The joint distributions for each socio-economic group is calculated from the International Public Use Microsimulation Data (IPUMS). We describe the development of spatio-temporal agent model based on the set of assumptions using a demographic domain knowledge (DDK) about the interactions and activities performed by the general population in a city. The DDK includes the percentage distribution of work types, work and education activity locations, and the rules of assignment of various activities.

The synthesized spatio-temporal activity based agent model can be abstracted as a temporal agent graph, which is used to analyze the spatio-temporal spread of epidemic through a rule based simulation. The rules govern the spread of disease among humans. The simulation results are visualized to observe the spatio-temporal spread of the disease for a given initial triggering point of the epidemic. The visualization provides an insight of the epidemic and assists in making effective decision measures to prevent epidemic spread [4].

A. Disease Transmission Process and Model

The infection time of an individual is divided into two parts: latent period and infectious period. In the latent period of infection, an infectious agent is transmitted to an individual and the infected person cannot transmit infection to others. The infectious time is further divided into two time periods: incubation period and infected period.

Direct or indirect human contact plays an important role in an epidemic outbreak and its duration. Direct contact requires an infected and a susceptible person to be in close proximity such as being inside a house. Indirect contact is when an infected and a susceptible are not interacting directly and the

[1]School of Electrical & Computer Engineering, Purdue University, West Lafayette, IN, USA e-mail: (ghafoor@purdue.edu).
[2]Umm Al-Qura University, Saudi Arabia.
[3]Dept. EE, Univeristy of Engineering & Technology, Lahore, Pakistan.
[4]Institute of Public Health, Lahore.
[5]Community Medicine Dept., Fatimah Jinnah Medical College, Lahore.

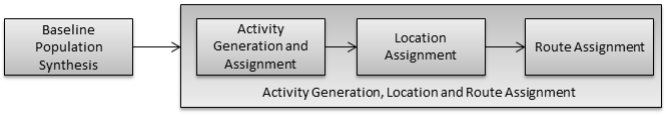

Fig. 2: Synthesis of Spatio-Temporal Agent Model

environment acts as a medium to transmit the disease such as at market places, hospitals, etc [7].

In a basic model a person can play three types of roles in disease dynamics. The individuals can be classified as susceptible, infected, or recovered. Susceptible persons are those who can get infected due to some infectious agents such as interacting with pathogens present in air, water, or soil. The infected individual interacts with susceptible individuals and transmits the infectious agent through air, water or other means. A recovered person never gets infected again [5].

There are three commonly known disease spread models: deterministic models, stochastic models and agent based models. In a simple deterministic model, a homogeneous, uniform mixing, closed population is assumed and every person at any time is either susceptible, infected or recovered [6]. In a stochastic model, infection is transmitted if a susceptible person is in close proximity of an infected person. The spatial dimension of a population is not considered in these models.

Agent based models allow us to incorporate heterogeneity aspects to the traditional epidemic models. The models require data to be collected form every person, which is difficult to achieve. One solution is to synthesize an agent population. The population synthesis process is highly data dependent and often has the shortcomings of unavailabe data requiring several assumptions to be made [8]. Synthesizing an agent population to study the disease spread requires a connected network. This contact network plays a key role in the transmission of disease among the population. The network is built by understanding the types of activities performed by the individuals, places of activities, time required to perform an activity and estimated number of persons in an activity.

Microsimulation is a process of simulating individual behavior to study and predict the effect of various complex and dynamic processes. In order to run the process, a micro dataset is created such that it represents the desirable characteristics of agents. Once synthesized, the behavior of the agents is simulated using certain mathematical models. To date most of the microsimulation models synthesized are application specific and are based on a conventional approach proposed in [9]. The challenge in synthesizing a microsimulation dataset is to generate a baseline population which is used to construct the agent model. This process requires sample data from real social networks, the demographic details, the education and work and information about other activities to create the baseline agent population [3]. Iterative Proportional Fitting (IPF) is a well established and commonly used approach to synthesize agent population data [9]. This approach is used to synthesize population in travel demand systems [11], [12]. There are some limitations to IPF when it is used to synthesize the baseline populatio widely discussed in various research such as [10], [13].

TABLE I: IPUMS Sample Data

Persons	Urban	Pernum	Age	Edupk
4	2	1	19	320
4	2	2	18	310
4	2	3	15	310
4	2	4	13	230
3	2	1	27	320
3	2	2	25	220
3	2	3	3	000
2	2	1	23	310
2	2	2	17	230

II. POPULATION SYNTHESIS FOR A SPATIO-TEMPORAL AGENT MODEL

We propose a methodology for synthesizing a generic agent based model for epidemiology analysis and predictions about disease spread in large metropolitan cities. The model incorporates human interaction and is generic in terms of its applicability to any metropolitan city. The overall methodology of an agent model synthesis is divided in five components. Figure 2 shows these processes.

As mentioned before, constructing an individual based population to study epidemic spread is a highly data dependent process. IPUMS sample data and DDK are used to synthesize a population's demographic attributes, social interactions and activity behaviors which represent the actual population of a city. All the individuals or agents for the synthesized population must be consistent with the percentage of sample data available for the city under consideration.

The first challenge is to identify demographic attributes which are important to analyze the disease spread in the city. The census data, and the IPUMS sample data are collected according to the selected demographic attributes. Using the above mentioned datasets, the baseline population is synthesized. A city can be divided into small geographic units such as union councils (UCs) to create a geo-centered synthesis of heterogeneous population. The geographic data (Geo-referenced shapefile) if available provides continuous space for the agents to perform various activities including transportation. Using this file, the spatial patterns of an agent's activities are observed.

A. Input Datasets

Below is the description of data sets needed to synthesize the agent model including IPUMS, and Census data. For the proposed methodology, we select age, gender, the type of area of residence, and the education level from the available demographic attributes in IPUMS.

The first data set used in the synthesis of the agent model is the IPUMS. Generally, IPUMS is available for a small percentage of households. The methodology proposed in [14] is used to use this data for generating micro data records for all the households upto 100%. IPUMS for the city of Lahore is available by the University of Minnesota, it represents a 2% sample of the census data collected in 1973 for the whole city [15]. Every individual has several attributes such as household size, age, and gender.

In the sample data, age of an individual can have a value between 0 and 99. For this research, we divide the age attribute

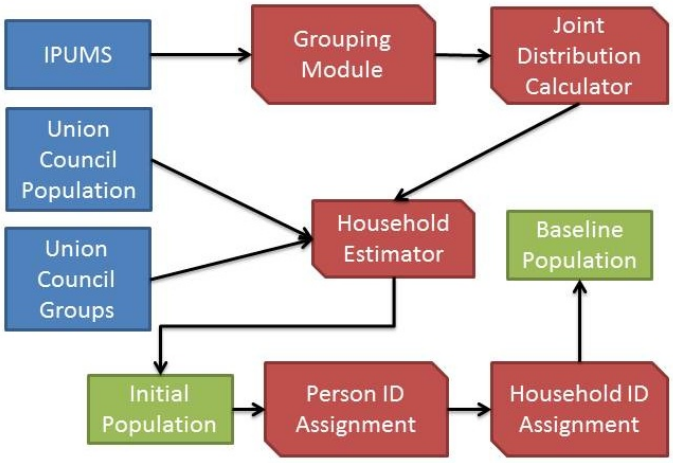

Fig. 3: Process of Baseline Population Synthesis

into five groups. Each of the first four groups corresponds to a duration of 15 years. The fifth age group represents all the individuals who are of age 60 years or above.

There are ten levels of education reported in IPUMS data. Education and age of individuals are combined to provide a higher level of grouping. This aggregation considerably reduces the complexity to compute the joint distributions without losing important details. We consider four levels of education: Basic, Intermediate, Higher and no education.

Table I represents a sample of this data. The first column (Persons) represents the family size of each individual considered in the sample. The second column (Urban) represents the residential area of the city which has either a value of 1 for urban or 2 for rural. The third column (Pernum) is the identifier attached to individuals living in one family. For example, all individuals with Persons value = 4, i.e., having (Pernum) from 1 to 4 represents one household. Columns for Age and Education levels, are grouped as mentioned in Section II-B.

The limitations of the IPUMS sample data is that it provides population samples for only two types of residential areas and without any information on their geographical locations. According to the census data, "inner" city areas are also identified, but not classified in IPUMS data. To address this discrepancy, we assume that the population in the inner city (IC) has a population distribution which is mixed of both urban and rural areas. For the inner city areas, the joint distributions are computed by taking an average of both urban and rural populations. *We use our knowledge of the city* to associate UCs to each of the three population groups.

B. Baseline Population Synthesis Module

Module 1 in Figure 2 is where the joint distributions for demographic attributes are calculated and used to synthesize the baseline population. This module takes the IPUMS sample data, the grouping of UCs and information about the percentage population for each UCs as input datasets. The output generated by this module is the percentage distribution of the overall population over the whole city.

Every household in the baseline population is assumed to be either a measure of a household or does not have any family household. Households with family size less than 15 are assumed to be family households. A household with a size above fifteen is assumed to be a *group quarter* [11]. Every household is assumed to have at least one person living in it.

The data flow diagram to synthesize the baseline population is displayed in Figure 3. The first module, called Grouping and Joint Distribution Calculation Module performs a two step process. It groups the population based on their age and education level. Subsequently, the probability distribution for each group of age, education, gender, and family size is computed. We consider a population in groups of age, gender, and education and compute the joint probabilities in terms of size of every group. The next module uses the UC population data and UC group data to compute the probability of households of every size in every socioeconomic group of population. Using the household size probabilities, the probability of houses with each family size in every UC is computed. Each household is then populated with persons using the joint distributions of demographic attributes extracted in Grouping and Joint Distribution Calculation Module. The output of this module is a file with number of persons in every group of population. The Person Identifier Assignment Module generates an agent model and each agent is assigned a unique identifier along with demographic attributes. Household Identifier Assignment Module assigns household ID to the population to complete the process of synthesis of baseline population.

C. Case Study: Baseline Population Synthesis for the City of Lahore

We now illustrate the process of population synthesis for the city of Lahore. For this purpose, we assume four control variables which are age *(A)*, gender *(G)*, education *(E)*, and family size *(F)*. These are random variables and can assume various values as described below. We compute the number of households in every socio-economic group. The process of computation is an existing approach extending the approach proposed in [14].

For our case, the possible values for A, G, E and F are $\{1, 2, 3, 4, 5\}$, $\{1, 2\}$, $\{E0, E1, E2, E3\}$, and $\{1, 2, 3 \ldots 25\}$, respectively. Let U be a random variable representing ID of a union council for the city of Lahore. The possible values for U are $\{1, 2, 3 \ldots, 151\}$. Let S be the random variable that represents the socio-economic status of the union councils. The possible values for S are $\{1, 2, 3\}$. Let n_f = Number of persons with family size f, N = Total population, n_s = Number of persons in socio-economic group s. The probability of a person being in group s is then computed as follows: $Prob(S = s) = n_s/N$. And the probability of having a household of family size $F = f$ is calculated for group $S = s$ as follows: $Prob(House\ with\ F=f \mid S=s) = n_f/N$.

Given a population size, we can identify the number of persons in every UC with the family size f. We define matrix M representing the joint percentage distribution of the overall population with values represented in A, G and E matrices. The total number of elements in M is equal to the product of all possible values in $|A| \times |E| \times |G|$. Each member m_{ijk} of matrix M is a triplet where: $i \in [1,2,3,4,5]$; $j \in [1,2,3,4]$; $k \in [1,2]$, e.g., *[Age 1, Education 1, Male]* represents a tuple.

For the IPUMS sample data, the conditional probability distribution of each group represented as element of M is calculated as follows: n_{ijk} represents the total population with attributes i, j, k. n_{fs} represents the total population in socioeconomic group s with family size f. *Prob(tuple takes value i, j, k| F=f, S=s)* = $n_{ijk} / \sum f \sum s\ n_{fs}$

This joint probability distribution indicates the number of persons for each of the socio-economic groups. The joint probability distribution gives us a number of persons in every UC that falls under these groups and is displayed.

The computed number of individuals in every group may not be a whole number. Note, the rounding of these numbers can give error by adding extra persons or having less persons. To fix this issue, extra persons are deleted or added to the population. Every person in the group is then assigned a unique identifier and represents an agent. In addition, each person is assigned his/her own demographic information. One person is written in each row of the output data file, known as population file. Each person is assigned its own demographic attributes and a household identification number.

The population of the city is divided into three groups. One group of in areas which can be highly dense with a low education level average. The other type of areas can be characterized as having a low population density, but are highly educated. These two groups represent, the extreme cases and must be analyzed separately since the spread of disease in these areas can have different spatio-temporal patterns. The major part of the city can have mixed population which can be represented by the attribute values falling in the middle of the two extreme groups. The probabilities of the first two groups can be extracted from the sample data and the probabilities of the third group are assumed to be the average of the above mentioned two extreme groups.

III. Demoghrahic Domain Knowledge for Activity Generation

The overall process of activity assignment requires the following datasets: **Baseline Population** discussed in section II; **GIS Dataset** which includes the shapefile of the city of Lahore providing the geo-coordinates of the city; **Demographic Domain Knowledge Dataset (DDK)** which provides the various statistics, assumptions, and rules about the activities of the population based on socio-economic and demographic knowledge. The overall DDK is categorized in 3 parts: Percentage Distributions of Work Types, Activity (Work, Education, Transportation, Household) Centered Geographic Location Data, and Rules of Assignment of Work, Education, and Route Activity. Which are briefly described with application to the city of Lahore. The following is thier description:

A. Percentage Distributions of Work Types (DDK Part I)

Census data provides aggregated summary tables that represent the percentage of employed and unemployed persons in a city, the number of working persons per household and the average number of employed persons per household. *Using the knowledge about the city of Lahore*, an estimate about the employment percentages and work percentages for each group of population in the city is made. Table II shows the employment percentage distribution of the city assumed for the agent model.

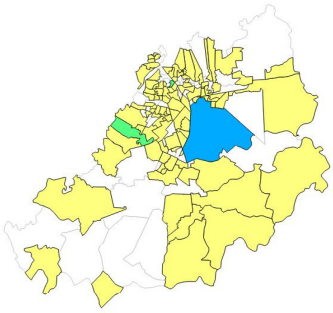

Fig. 4: Blue: high work density Areas, Yellow: average work density areas, Green: low work density areas

TABLE II: Employment and Unemployment Percentages

SocioEcon Stat	Urban	Urban	Rural	Rural	IC	IC
Gender	M	F	M	F	M	F
Employed	80	40	60	10	80	25
UnEmployed	20	60	40	90	20	75

Again, *using the knowledge about the city of Lahore* we assume that the high income households are more likely to be populated in urban residential area, whereas the low income households reside mostly in rural areas. The reason for this division is the difference in the levels of education attained by the individuals in each residential area. Table III shows percentages of various types of work assumed for persons of every residential areas which is used for work distribution in synthesizing the agent model. These percentages are assumed for all the three residential areas which were also provided through the census data.

B. Activity Centered GIS Location Data (DDK Part II)

Geo-referenced GIS Layers or shapefiles are the third required input dataset. The proposed methodology requires the location coordinates for assigning locations to activities performed by the agents pertaining to residential areas, work areas, educational institutes, and transportation routes.

The shapefile for Lahore provides us with boundaries of UCs in every town of the city. Figure 4 shows the number of UCs on the map of the city. The color shows the classification of UCs based on their population density. In essence, the synthetic population is distributed over various areas of the city. The database files, for the shapefile, provides information about the population per UC and the size of its area.

In synthesizing our model, we assume that the residential areas do not overlap with any other activity location such as work, schools, parks, etc. To ensure a uniform distribution of

TABLE III: City's Work Type Percentage Distributions

	Urban	Urban	Rural	Rural	Inner City	Inner City
W1	.23	.13	.12	.1	.12	.16
W2	.12	.29	.10	.12	.32	.29
...	...	...	...	...	...	...
W10	.09	.01	.1	0	.2	.01
Total	1	1	1	1	1	1

TABLE IV: Work Assignment Rules Based on City's DDK

Work Activity Type	Education	Age Group
W1	E3	3, 4
W2	E3	2,3,4
W3	E2	2,3,4
...	...	...
W8	E1	2,3,4
W9	E0	2,3,4
W10	E1, E2	2,3,4

residential places in the city of Lahore, we divide the area in small grids. Each grid represents a block and population is assumed to be uniformly distributed in every grid.

All the different activities performed at one place are assigned the same activity location type. Activity locations are identified on the map of Lahore *using the existing knowledge* of the city to extract coordinates of locations. Accordingly, six different types of work locations are identified. Likewise, we select educational institutes (schools, colleges, universities) from the shapefile of Lahore. We assume that there are five colleges and five schools in every UC, and ten universities in the whole city.

The proposed model assumes that there is only one mode of transportation available which is public transportation. We identify that there are 9 main bus routes that cover the whole city. Each bus route has multiple stop points which are used the in route assignment component.

C. Activity Assignment Rules Based on Demographic Attributes (DDK Part III)

Three types of knowledge based rules for assigning work, education, and route to agents is performed by Activity Generation module described in Figure 2. The existing knowledge about the population of Lahore indicates that the type of work performed by an individual depends mainly on age, gender, education and experience. For example, a managerial position usually requires an individual to have a college degree and several years of experience. This is how Table IV is produced.

Using DDK I, we can stipulate the percentage distribution of work for different socio-economic areas. People living in higher income areas are more likely to be more educated than the ones living in low income areas. Also, the people in high income areas are less likely to be unemployed as compared to other areas of the city. Also, the percentage of women working in elementary (low income) occupations is higher in low income areas. Work activity types can then be assigned based on these assumptions. Work type is represented as a vector of ten elements. *W = [W1, W2, W3 ... W10]*. Similarly, education activity type can be represented as the following vector *E = [E0, E1, E2, E3]*.

Route assignment module assumes that the only mode of transportation available to the general population is public transport. We estimate the map of bus routes using the shapefile of the city. Transportation is an activity performed by every individual. During weekdays, only working population of the city is assumed to use transportation. Further assume each individual uses a single bus route for commuting that is closest to the location of the household.

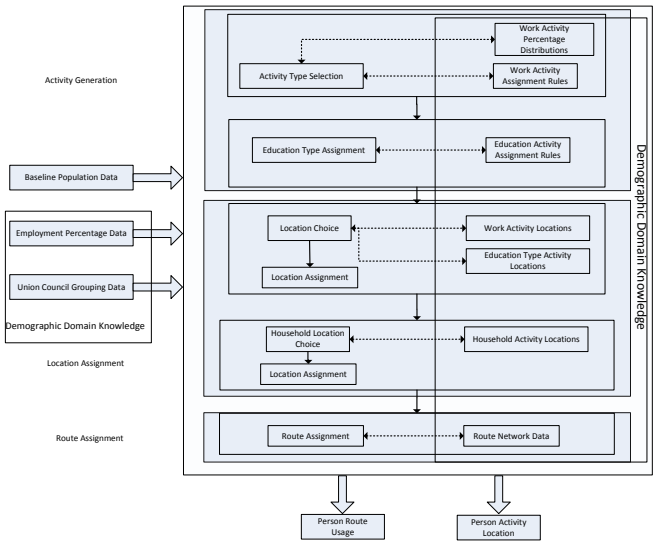

Fig. 5: Conceptual Diagram of Activity Assignment, Location Assignment, Route Assignment Modules

IV. Synthesis of Agent Model

Based on the baseline population data, we now present a three step process for synthesizing an agent model. The process entails assigning spatial and temporal activities to agents through the DDK III. Step one is activity generations of household. The process of assigning activities to the agents is implemented by the last three components shown in Figure 2. In this section we discuss three components of the system which consult the DDK as discussed in Section III. These modules are labeled as Activity Generation, Location Assignment, and Route Assignment. Figure 5 shows a more detailed diagram of the activity generation and location assignment process.

After work and education activity types are assigned to persons, location choice module selects a location according to the activity type assigned to a person. If the activity type location is empty, the location is assigned to the person. For a household, each entry is unique. Routes are assigned to the population depending upon their work, education, and household activity location. In the following sections, we explain the agent synthesis process in detail.

A. Activity Generation

Our assumption for the activity generation module is that every individual lives in a house and there are no homeless people in the city. Table II shows the employment percentages of working and non working population in every socio-economic group. The employed persons are selected following the percentages from Table II. These employment percentages for both genders in each socio-economic group provides us with information of how many people are working in every UC. Assignment of the work type to each working individual depends on their age, gender, and education. We calculate the number of persons working in every work type by using the percentages from Table III for each socio-economic group. Every person is equally likely to be selected as employed and

unemployed by this module. Also, all the individuals satisfying the requirement for every work type also follow a uniform distribution of selection.

The other type of non household activity is education. Using the baseline population synthesized in the previous section provides us with the information of the highest education level attained by every individual. There are three levels of education, basic, medium, and higher. Individuals who are assigned E1 (basic), E2 (middle) or E3 (higher) if declared students will go to schools, colleges, and universities respectively. The second assumption is that there are no students in the city with age 30 years or older. The third assumption is that every person in the city is either working or is a student.

Algorithm 1 Work Assignment Algorithm

Require: Employment Percentage Data, Work Type Percentage, Baseline Population
Ensure: Total Population
for each tuple of Employment Precentage Data **do**
  Insert into Work Table
  From Baseline Population
  With Employment Percentage Criteria
for each tuple in Work Type Percentage **do**
  for each tuple in Work Table **do**
    Update Work Type in Work Table
    With Work Type Percentage Criteria
for each tuple in Baseline Population **do**
  if Baseline Population $\neq$ Work Table **then**
    Insert tuple into Work table
    Update tuple Work Type = NW
    if Education $>$ E0 **then**
      Update Work Type in Work Table
      With Baseline Education

Algorithm 2 Location Assignment Algorithm

Require: Work Activity Location, Education Type Activity Location, Work Table
Ensure: Total Population
for each tuple of Work Activity Location **do**
  if Work Type count $>$ 1 **then**
    AgentCount = WorkType Count*100
  else AgentCount = 100
  while AgentCount $>$ 1 **do**
    Update Activity Location
    From Work Table
    Where Work Table.WorkType= Work Activity Locationi.WorkType
for each tuple GIS Household Coordinate File **do**
  if Work Table Household XY = NULL **then**
    Update Work Table Household XY
    With GIS XY Coordinate
    For All Work Table Same Household ID

B. Location Assignment

The location assignment module performs two tasks. The first task is to assign work locations and educational institutes to individuals. The second task is the assignment of household locations. In this process, we spatially spread households over the whole city. The rules of these two assignment tasks are different and are explained in Algorithm 2.

Household assignment is similar to the work activity location assignment. The only difference is that each household is assigned just one point of x-y coordinate. Depending on the size of the population, an approximate number of coordinate points can be extracted from the GIS file of the city of Lahore if needed.

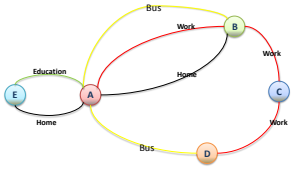

Fig. 6: An Agent Graph Model

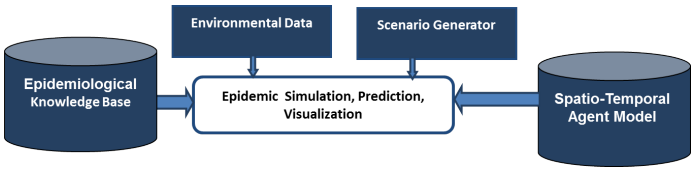

Fig. 7: Simulation Model of Disease Spread

Routes are assigned to agents following the rules explained in Section III. During weekdays, the transportation activity is work dependent. We use the Euclidean distance between the individual's house location and coordinates of the bus route which is closest to the house. Multiple buses are assigned on each route. Individuals are assigned a bus number which is randomly selected. This is a basic model of transportation used to understand the impact of transportation on disease spread.

C. Agent Model and Implementation

In this synthesized model, an agent can be represented as an entity with its demographic attributes such as age, gender, etc. Similarly, work, education, and route activities can be represented as entities. The model can also be represented as an abstract temporal graph. In this abstraction, agents constitute vertices and activities are represented as edges with temporal attributes between each agent. An example is represented in Figure 6. Edges can be static or dynamic depending upon the type of activity. The static edges correspond to activity associated with people living or working in the same location. Agents riding a bus have dynamic connections since their interaction with agents change daily. This representation will be the backbone of how disease transmission is implemented.

V. Simulation and Visualization

The spatio-temporal agent model synthesized subsequently is used to simulate an epidemic spread. Several computations can be extracted from the simulation such as the number of infected agents at different levels of granuality. The epidemic spread is visualized to assist in preplanning exercises that can predict further possible epidemic spread The simulation allows us to initially select infected agents on the map of a city which will then be run through different disease dynamics (rule sets) based on several demographic attributes.

TABLE V: Rule Set

Rule Set	Contact Duration	Effective Contact	Recovery Rate
1	4+ agents	30 mins	5 days
2	2+ agents	30 mins	5 days

Infection is transfered through similar activities performed by the agents. As mentioned in previous sections, these activities are represented as edges of the network and agents represent vertices where agents are conected when performing the same activity at the same location. Visuallization by coloring UCs where color intensity is based on the number of infectious people in the area.

The simulation system is composed of several modules which is portrayed in Figure 7. The Epidimological Knowledge Base is developed by health officials and through data provided by hospitals and domain experts. The scenario generator is the parametric inputs that define the initial infection group described more in the coming sections. The spatio-temporal agent module contains the database of relational schemas that will be stored and used as the basis of keeping track of the spread. The spatio-temporal agent model synthesized in the previous section is presented in the simulation as a relational database table. Therefore, the simulation of the infection is exectuted through a database environment.

A. Scenario Generator

To trigger an infection spread, several parameters need to be input into the simulation process. These inputs are ususally insetrted by a user to the system, and the parameters are specific to that users' need to understand certain aspects of the area's demographic or geographical importance to the population. These parameters are defined under the intial infection state.

To intiate the spread of infection, there must be an inital infection state. How this intial infection manifests itself depends on several parameters, including demographic and geographical characteristics such as gender, age, and location distribution. Using the schema shown in Figure 8, a random range queries on these parameters will than give a subset of agents that will be characterisized as the the intital infection agents. An example of an intial infecton can have the descriptors of only a certian gender and age group can be infected. In this specific case study, the density of the UC was taken into account to see how that could effect the overall infection rate. The intial infectioin characteristics can be chosen at the household granularity. For instance, one could choose to infect a specific number of agents within one household (this also can have the added description of age and gender).

B. Epidemiological Knowledge Base

We run a rule based disease spread simulation on the agent network synthesized in previous sections. The simulation uses two sets of rules pertaining to epidemic spread that are listed in Table V. These are preliminary test sets for this particular case study. Depending on the the area or city that is being simulated the rule sets will change accordingly.

This knowledge base is inserted into the simulation model as a series of SQL queries. Once the population relational data table is split into susceptible agents and infected agents, each susceptible agent is then taken and the queries are executed to find if there are any eligible agents (infected individuals) that follow the epidemiological knowledge base (rule set). In addition to the epidemiological data, the enviormental data can be introduced into the rule sets of infection as the enviorment can have bearing on the disease. Enviromental data can include temperature, humidity, pH, rainfall, etc.

Algorithm 3 Simulation Process Algorithm

```
Require: Main Agent Table, Activity Table, Rule Table
Ensure: Updated MAin Agent Table
    for each tuple <t> in Main Agent Table do
        if t.SIR = I then
            Add t to ITable
        else
            if t.SIR = S then Add t to STable
    for each tuple <s> in STable do
        Count tuples Into C
        From ITable
        Where Activity = s.Activity
        if C ≥ Rule Table then
            Update s.SIR to I and s tuple to Itable
            From Main Agent Table
```

C. Simulation Process

The simulation is an API running the SIR model through a JAVA program that uses SQLite to process the queries. The schema presented in Figure 8 is considered the main relational table that contains all the agents information including the SIR state they are in and what day an agent became in that state.

Algorithm 3 shows the simulation proces of infection. This algorithm is repeated from day to day. This is done for each activity (home, transportation, work) in series. If an agent is eligible for infection (example of a eligibility is shown in Table V), the SIR and IDAY columns are updated to the infectious state. An agent only becomes infected a day after it has become eligible for infection which is recorded in column "Day". After all the activities are processed, the agents are processed for the recovery state.

D. Case Study for the city of Lahore

In this section, we discuss the results of epidemic spread simulation. We run the simulation for Rule Set 1 and 2 with the same initial infected individuals. The simulation follows SEIR model of disease spread [3]. Exposed individuals can stay in this state for a day and disease cannot be transmited. After the day agents are in the state I. In our simulation, we assume the duration of state I to be 5 days. After the infected period, individuals are in R state. In this simulation, we assume dynamic contact between individuals during transportation and static contact is assumed at work activity locations.

For Lahore, two different UC densities were infected with the same number of intially infected agents. One experimental run was intilized in a highly dense populated UC. While the other was intilized in a sparsley dense populated UC.

On day 1, epidemic is triggered by introducing initial infected agents in the population of the city. For this example visualization, we select 147 agents to be initially infected

Gender	Age	Education	Work Type	Route	Bus Num	House ID	House X	House Y	Work X	Work Y	SIR	Day	Activity

Fig. 8: Schema for the Main Table in the Simulation Process

agents working in the inner city at 12:00 PM. The agents are selected from various locations. Figure 9 displays initialily infected agents. For case 1 simulation, on day 3 there are 425 cases of new infections observed in the city. All of the infected individuals got infected at work locations. For case 2 simulation, new infected individuals count 352 and points of transmission of infection are work locations.

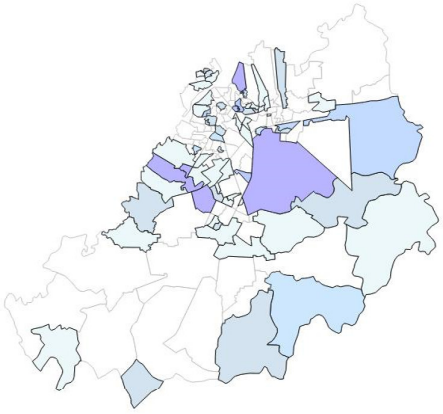

Fig. 9: Day 1 Initial Infected Residential Location

For case 1 simulation, on day 3 there are 425 cases of new infections observed i n the city. All of the infected individuals got infected at work locations. For case 2 simulation, new infected individuals count 352 and points of transmission of infection are work locations. Infected agents for both cases are from all three types of areas of the city. Figure 10 shows the areas of infection incidence on the city map.

On day 4, new infections observed in case 1 and case 2 simulations are 90 and 16, respectively. All the incident infections of both simulations occur in households. We observe that for case 1, the incidence of infection occurs at both household locations and transportation location whereas infection

Fig. 10: Day 3: Case 1 (Left), Case 2 (Right)

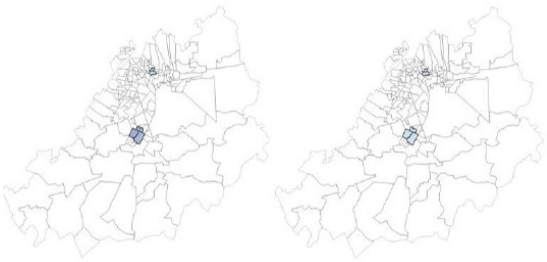

Fig. 11: Day 4: Case 1 (Left), Case 2 (Right)

TABLE VI: Table: Day 5 Disease Dynamics

Activity	Case 1 Infected	Case 2 Infected
Transportation	193	0
Household	196	47
Work	0	0

incidence of case 2 occurs only at household locations. The infection spreads in inner and rural areas of the city.

The transmission of infection during the transportation activity is responsible for spread of infection across the population of a city. We observe that most of the infection transmission occurs during the transportation activity on day 5. Most of the infections are observed in the inner city and rural areas. Disease dynamics of incident infection of day 6, 7, and 8 are displayed in Table VII. No new infection is observed after day 5 of case 2 simulation.

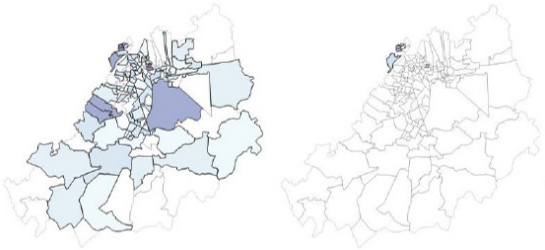

Fig. 12: Day 5: Case 1 (Left), Case 2 (Right)

TABLE VII: Day 6, Day 7, Day 8 Disease Dynamics

Activity	Day	Case 1 Infected
Household	6	17
Transportation	6	56
Work	6	33
Household	7	110
Transportation	7	158
Work	7	806
Household	8	1
Transport & Work	8	0

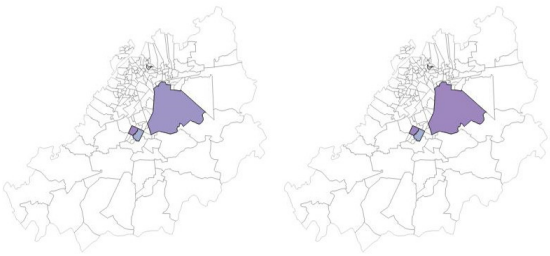

Fig. 13: Day 6 (Left), Day 7 (Right)

VI. CONCLUSION

In this paper we have presented a two step methodology to synthesize an activity based agent model. The first step is to synthesize a baseline population of agents using joint distributions of demographic attributes tabulated by IPUMS sample data. The second step is to assign activity types to the agents as discussed in section 4. Using this DDK and baseline population, agents are assigned activity type and an activity location. A rule based simulation of a disease spread is run on the synthesized activity based agent graph. This visualization provides a preplanning exercise to predict an upcoming epidemic and for decision making and other visual analytics. A case study for the city of Lahore is presented to examplify the proposed methodology.

REFERENCES

[1] C. Viboud et al., "Multinational Impact of the 1968 Hong Kong Influenza Pandemic: Evidence for a Smoldering Pandemic," *J. Infect. Dis.*, vol. 192, pp. 233–248, 2005.

[2] E. C. Claas et al., "Human Influenza a H5N1 Virus Related to a Highly Pathogenic Avian Influenza Virus," *The Lancet*, vol. 351, no. 9101, pp. 472–477, 1998.

[3] A. J. Heppenstall, A. T. Crooks, and L. M. See, *Agent-based Models of Geographical Systems*. Springer, 2012.

[4] C. L. Barrett et al., "Estimating the Impact of Public and Private Strategies for Controlling an Epidemic: A Multi-agent Approach.," in *IAAI*, 2009.

[5] J. Satsuma et al., "Extending the SIR Epidemic Model," *Phys. A*, vol. 336, no. 3, pp. 369–375, 2004.

[6] A. R. McLean, "Infectious Disease Modeling," in *Infectious Diseases*, pp. 99–115, Springer, 2013.

[7] S. Riley, "Large-scale Spatial-transmission Models of Infectious Disease," *Science*, vol. 316, no. 5829, pp. 1298–1301, 2007.

[8] D. Ballas et al., "Exploring Microsimulation Methodologies for the Estimation of Household Attributes," in *GeoComp, Virginia, USA*, 1999.

[9] R. J. Beckman, K. A. Baggerly, and M. D. McKay, "Creating synthetic baseline populations," *Transportation Research Part A: Policy and Practice*, pp. 415–429, 1996.

[10] J. Y. Guo and C. R. Bhat, "Population synthesis for microsimulating travel behavior," *Journal of the Transportation Research Board*, vol. 2014, no. 1, pp. 92–101, 2007.

[11] J. L. Bowman, "A comparison of population synthesizers used in microsimulation models of activity and travel demand," *http://jbowman.net/papers/2004.Bowman.Comparison_of_PopSyns.pdf*.

[12] N. Jonnalagadda et al., "Development of microsimulation activity-based model for san francisco: destination and mode choice models," *Journal of the Transportation Research Board*, vol. 1777, no. 1, pp. 25–35, 2001.

[13] T. Arentze, H. Timmermans, and F. Hofman, "Creating synthetic household populations: problems and approach," *Journal of the Transportation Research Board*, vol. 2014, no. 1, pp. 85–91, 2007.

[14] N. Wongchavalidkul and M. Piantanakulchai, "Estimating synthetic baseline population distribution when only partial marginal information is available," in *Proceedings of the Eastern Asia Society for Transportation Studies*, vol. 7, 2009.

[15] S. Ruggles et al., "Integrated public use microdata series (ipums): Version 5.0 [machine-readable database]," *University of Minnesota, Minneapolis, available at http://usa. ipums. org/usa*, 2010.

Acknowledgments: this work has been partially supported by the National Science Foundation, and the US Defense Threat Reduction Agency, under Grant IIS-0964639, and HDTRA 1-10-1-0083, respectively.