Understanding Complex Systems: Population Interactions Resulting in Disease Transmission

By Sara Y. Del Valle and James P. Smith

Keeping our nation safe from natural, accidental, and intentional disasters requires extensive planning. Integral components of the country that need to be protected include the population, economy, and infrastructure. Failures in any of these three highly interconnected systems can have catastrophic effects. Because humans are central to all of these components, protecting the population from contagious diseases and devastating events is crucial.

With this goal in mind, mathematicians and computer scientists at Los Alamos National Laboratory created a synthetic population of more than 280 million individuals to represent the entire population of the United States (Stroud 2007). We use this population to model the movement of people through a representative timeframe to understand how people interact and spread transmittable diseases, such as influenza. To represent social interactions, we connect people as they cross paths at the same place and time. For example, George teaches at a school from 8 to 9 AM, travels by bus from 9–10 AM, and eats at a restaurant from 10 AM to noon. Similarly, the driver of George's bus works from 8 AM to noon, with passengers including George from 9–10 AM, and a business executive from 10–11 AM. For any instant in time, we can look at which people share locations. George shares locations with another teacher at the school at 8:30 AM, and with the bus driver from 9–10 AM. The bus driver then shares the bus with the business executive, who goes to his office and works with other people. Combining these social interactions creates a time-dependent, person-to-person social network (See Figure 1). Studying that network can reveal how a disease that requires physical proximity for transmission could be spread.

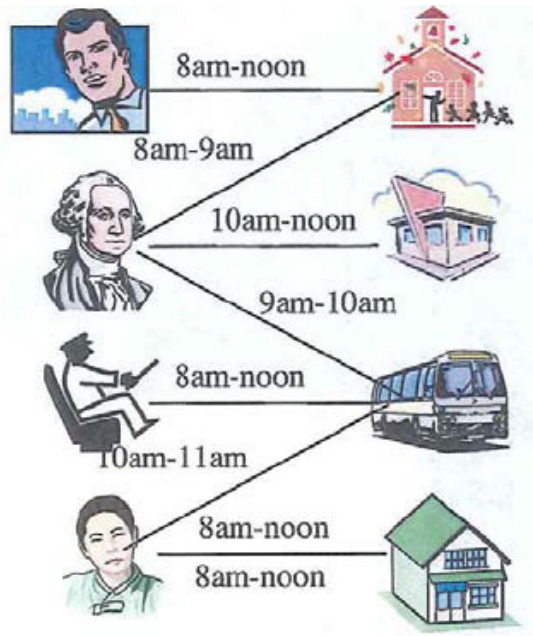

Figure 1. *Schematic representation of social interactions.*

Generally, a college student is unlikely to be at a senior citizens center and elementary school children are unlikely to be at work in an office. To generate the complex interactions within the population, we need a realistic representation of individual demographics. We use U.S. census data for this purpose. For privacy reasons, the Census Bureau does not release its full records; the only way to obtain a realistic population is to create a synthetic one that identically matches all the available information. The Census Bureau does release a 5% sample of its complete records attached to small study areas, along with marginal distributions (e.g., the total value of one or more demographic characteristics) of demographics by block groups. Using a statistical procedure called iterative proportional fitting (see Appendix), we create a synthetic population from the census sample data that matches the United States population. The result is a set of households and individuals, geographically distributed with correct demographics, that is statistically indistinguishable from the real population. In each block group, the synthetic population matches the actual population in several statistical measures: number of residents, number of households, ages of the household's residents, household size and membership distribution, household income distribution, number of workers, and number of vehicles.

In the next step, we generate a set of activities, such as shopping, going to school, working, and college, that fits people within specific demographics and home locations. We assign the appropriate activities to every person in the synthetic population. For data, we turn to large metropolitan planning offices, most of which regularly perform detailed surveys of small samples of their populations, keeping track of every movement of all members of a household in the course of one or more days, and recording respondent demographics. Because the surveys track these events by time of day, we also know how long each activity lasts. We study the surveys first to determine which demographic fields optimally classify people by some important characteristic—like total time spent per day in each of their activities. For each household in the synthetic population, we use a procedure called "binary tree matching" on the chosen demographics to assign appropriate activity patterns based on the survey data (see Figure 2).

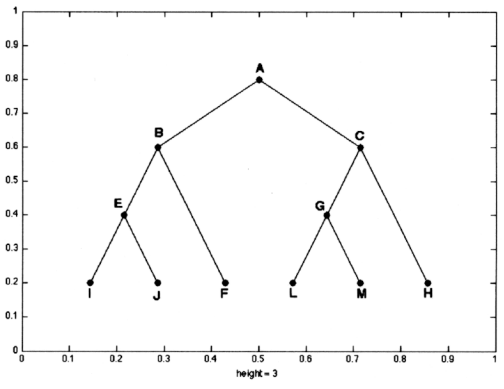

Figure 2. *Binary tree for the assignment of households in a survey. For example, the population could be split into households with income above (B) or below (C) $50,000/yr, then split again into those with (E & G) or without children (F & H), and again into those with children above (I & L) or below (J & M) the age of 5 years. Finally, a set of survey households is randomly selected from each of the final categories: F, H, I, J, L, M.*

Given a reasonable set of activities assigned to the entire metropolitan population, each of these activities must be located in the city. For this we use a simple gravity model (see Appendix), which assigns a location to each non-household activity of each individual, based on distance from residence, work, or school. For these assignments, the synthetic individuals carry out their activities near where they are; for example, workers will go to lunch near their work location.

The model then partitions each location into sub-locations: individual classroom mixing groups within a school, shops within a shopping center, work groups within a business location. The number of sub-locations is computed by dividing the location's peak occupancy by

the appropriate mixing-group size. The mean work-group size varies by a standard industry classification code (SIC). We used information from studies conducted by Yee (1999) and Michaels (2003) to estimate the mean work-group size by SIC code. The mean work-group size was computed as the average from the two data sources (normalizing the worker density data), and ranges from 3.1 for transportation workers to 25.4 for health service workers. The average number of workers over all types of workgroups is 15.3.

The social contact network emerges from the simulation as individuals move through their daily activities and move into and out of contact at sub-locations (Del Valle 2007). Once we generate the complex network of interactions in the population, we study the effects on the spread of disease. Disease transmission occurs only between individuals who are in the same sub-location at the same time. The probability that a susceptible individual becomes infected during an activity is computed by accumulating transmission probabilities per unit time of contact with each infectious co-occupant of the sub-location. The transmissivity, T, (e.g., disease transmission) between pairs of individuals is multiplied by a susceptibility multiplier (e.g., how susceptible a person is to a disease) and an infectiousness multiplier (e.g., how infectious a person is). The infectiousness multiplier depends on the disease stage, the treatment history, and the age of the infectious person. The susceptibility multiplier depends on the treatment history or age of the susceptible individual. If susceptible person j has a susceptibility multiplier S_j, and infectious person i has an infectiousness multiplier I_i, then the probability that susceptible individual j will be infected during the activity is computed as

$$P_j = 1 - e^{-\sum_i TS_j I_i t_{ij}},$$

where t_{ij} is the time during which susceptible person j was in the same sub-location as infectious person i; the sum extends over all infectious individuals who co-occupied the sub-location with individual j. The model computes the co-occupation times for all overlapping pairs of individuals as they enter and leave sub-locations.

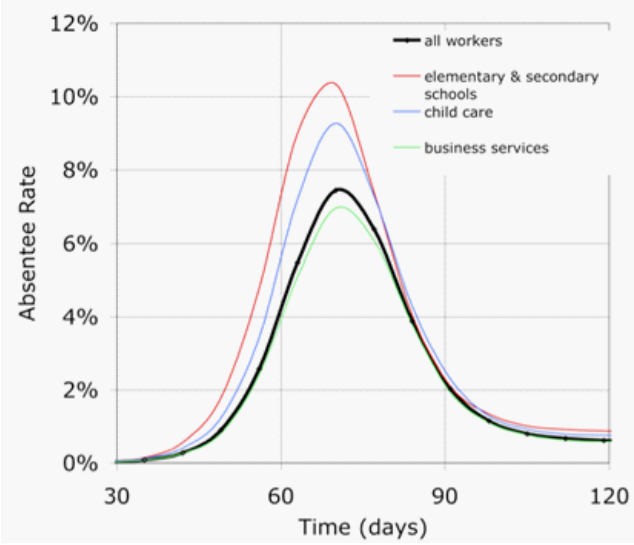

Figure 3. *Absentee rates for different industries from a hypothetical H5N1 pandemic influenza outbreak. Notice that schoolteachers could have higher absentee rates when compared to the average and to other industries.*

Disease impacts by different demographic characteristics (e.g., age), activities, industry classification, including workforce reductions (see Figure 3), and geospatial differentiation can be extracted for use in planning mitigation activities (Mniszewski 2008). Our simulations can provide policy makers with information regarding where disease spreads and who should be targeted for treatment strategies such as vaccination.

The results from our simulations (see Figure 3 and Del Valle 2007) are consistent with previous studies that have shown that disease transmission is typically higher in school-age children due to their highly connected social networks. Furthermore, our simulations can be used to estimate the potential impact of disease spread and workforce absenteeism in workplaces. We are confident that our simulated population captures realistic social interactions and can be used as a tool for understanding disease transmission.

References

Stroud PD, Del Valle SY, Sydoriak SJ, Riese J, Mniszewski S. Spatial Dynamics of Pandemic Influenza in a Massive Artificial Society. Journal of Artificial Societies and Social Simulation 2007;10(4),9.

Yee D and Bradford J. Employment Density Study, Canadian METRO Council Technical Report, April 6, 1999.

Michaels J (2003) Commercial Buildings Energy Consumption Survey, www.eia.doe.gov/emeu/cbecs/cbecs2003/detailed_tables_2003/detailed_tables_2003.html.

Del Valle SY, Hyman JM, Hethcote HW, Eugank SG. Mixing patterns between age groups in social networks. Social Networks 2007;29:539-554.

Mniszewski SM, Del Valle SY, Stroud PD, Riese JM, Sydoriak SJ. Pandemic simulation of antivirals + school closures: buying time until strain-specific vaccine is available. Computational & Mathematical Organization Theory 2008;14:209-221.

Appendix

Iterative Proportional Fitting (IPF)

Iterative proportional fitting is a statistical procedure that allows us to create a synthetic population that matches the marginal distributions (e.g., the total value of one or more demographic characteristics) at the census block-group level while retaining the demographic correlation structure of the samples in the large areas (Bishop 1975). It essentially creates the joint distribution matching all

the marginal distributions by taking samples from the partial set of full records. The result is a set of households and individuals geographically distributed with correct demographics, statistically indistinguishable from the real population.

Gravity Model

The model uses a two-stage gravity algorithm to assign a location to each non-household activity of each individual. The gravity model is widely used in traffic analysis and has been described in detail (Voorhees, 1956; FHWA, 1978; Martin and McGuckin, 1998). Each individual has an anchor activity: For workers, the anchor is their work activity; for students, their school activity; otherwise, it is the individual's residence. The first stage of the gravity algorithm assigns a location to the anchor activity of each worker and student. The model implementation of the gravity algorithm, in effect, sets the probability that worker i works at location j to be proportional to

$$p_j\left(i,\beta,N_j,\gamma,d_{i,j}\right) \propto \frac{e^{\gamma N_j} e^{-\beta d_{ij}}}{d_{ij}},$$

where d_{ij} is the travel distance in meters from the residence of worker i to work location j, and N_j is the number of workers employed at location j. The coefficient values are fit to ensure that the number of workers assigned to each work location matches Dun & Bradstreet business location data, and that the distribution of commuting distances from home to work matches the distribution extracted from the National Household Transportation Survey data. The second-stage gravity model that assigns locations to non-anchor activities follows a similar formulation, except that the distance is replaced by the sum of the distance from the anchor activity to the non-anchor activity plus the distance from the non-anchor activity to the place of residence.

Bishop YMM, Fienberg SE, Holland PW (1975) Discrete multivariate analysis. Theory and practice. Cambridge: MIT Press.

Voorhees AM (1956) A General Theory of Traffic Movement. 1955 Proceedings, Institute of Traffic Engineers, New Haven, CT.

FHWA (Federal Highway Administration) (1978) Quick-response Urban Travel Estimation Techniques and Transferable Parameters, NCHRP Report 187. Available from nationalacademies.org/trb/bookstore.

Martin WA, McGuckin NA (1998) Travel estimation techniques for urban planning, NCHRP Report 365. Available from nationalacademies.org/trb/bookstore.

Sara Del Valle is a scientist at Los Alamos National Laboratory. James P. Smith is a group leader at Los Alamos National Laboratory.