Jpn. J. Infect. Dis., 58, 2005

# Network Based Models of Infectious Disease Spread

Stephen Eubank*

*Virginia Bioinformatics Institute at Virginia Tech, Virginia 24061, USA*

**SUMMARY**: It has recently become possible to simulate directly dynamics on very large networks. This paper describes a model of epidemiology on a social network, the Epidemiological Simulation System (EpiSims), and offers general speculation on analyzing disease dynamics on networks. We describe the process of building a realistic social network, describe several different definitions of the network, each useful for certain purposes. Finally, we raise some important questions about structural properties of networks and how they influence dynamics.

## Introduction

It has recently become possible to simulate directly dynamics on very large networks (1). These simulations create new possibilities for mathematical modeling of epidemiology (2-4). The Modeling Infectious Disease Agent Study (MIDAS) established by the National Institute for General Medical Sciences, part of the U.S. National Institutes of Health, is currently composed of three research teams and a bioinformatics and computational support team. Each of the research teams has developed social network based models for epidemiology. This paper describes one of the models, the Epidemiological Simulation System (EpiSims) (5), and offers general speculation on analyzing disease dynamics on networks.

Section I describes the process of building a realistic social network. Some of this process is common to all the MIDAS models. Section II demonstrates several different definitions of the network, each useful for certain purposes. Section III introduces some important questions – specific to these models – about structural properties of networks and how they influence dynamics.

## Estimating a social network

Epidemiological models based on contacts among individuals complement more traditional models based on an assumption of homogeneous mixing among large subpopulations, often referred to as SIR (for Susceptible-Infected-Removed) models. If we represent people's contacts as a graph in which vertices represent people and edges represent contacts, an SIR model would resemble a collection of complete graphs: ones in which every pair of vertices is connected by an edge. Graphs representing more realistic contact patterns have far fewer edges and are much less regular.

One approach to building more realistic contact patterns is to select edges to remove from the complete graph. This is effectively the approach taken by structured population SIR models, which allow heterogeneous mixing among large, but homogeneously mixed, subpopulations. However, there are an enormous number of ways to remove edges, and it is difficult to pursue this approach far enough to create realistic mixing.

An alternative approach is to start with a set of vertices (i.e., people) and slowly build up their contacts. This is the approach we have taken. Using census data, land use data, activity (or time use) surveys, and transportation networks, we have created a detailed representation of a specific urban area. There are several steps to this process: population synthesis, activity assignment, location choice, and travel time estimation. Each of these is briefly described in this section.

The first step is to create a synthetic population, typically using census data for the U.S. The data consist of marginal distributions for a large set of demographic variables such as age, gender, and income in small areas, and joint distributions of the demographic variables for small samples in less well-defined areas. The data are presented in these two forms to protect individual privacy. Through a statistical fitting technique, it is possible to create a population with the correct marginal and joint distributions and assign each synthetic household to an appropriate home location. For example, both the number of people in a household and the ages of those people agree with the census data. Indeed a "census" of the synthetic population would be statistically indistinguishable from the original census. Each person in the synthetic population is assigned an identifier that allows retrieval of associated demographics at any stage in the analysis.

The next step is to assign a set of activities to each person in every household. This step determines the fidelity of the model, i.e. how well it captures the behavior of people on a typical day. Some models use only a few activities for each person, perhaps only home and either work or school. EpiSims relies on an extremely realistic set of activities produced by the TRansportation ANalysis and SIMulation System (TRANSIMS) (6). How the level of fidelity affects the simulated outcomes is an area of active research.

From activity surveys, we extract a set of templates along with the demographics of households associated with each template. We choose an activity template for each synthetic household by matching its demographics with one of those in the survey data. The templates may need to be adjusted in the case of a partial match, for example, if the number of children in the synthetic household is different from that in the survey household.

Our estimate of the social network is complete when every activity for each person has been assigned a location, arrival and departure time. Constraints imposed by the transportation system along with land use data and observations of travel time distributions by purpose of trip can be combined to provide good estimates for activity locations. In order to combine all these data, we must estimate the travel time between many locations by time of day given an initial assignment of activities to locations. This estimate is produced using a micro-simulation of the traffic induced by the specified demand. These estimated travel times are fed back into location choice and activity assignment algorithms in a loop. The feedback process converges into an equilibrium in which each

*Corresponding author: E-mail: eubank@vt.edu

person's activity location assignments and routes are feasible and cannot be improved in the context of choices made by all the other people.

This method of assigning activity locations is time-consuming, but it has several advantages over attempts to "observe" location choices directly. First, the model parameters are fit to more fundamental aspects of human behavior. That is, we effectively capture how and why people choose to move about as opposed simply to describing where they go. The difference is crucial when we try to model hypothetical situations. For example, no amount of observation of behavior on a typical day will allow generalizing to behavior in a crisis unless we use the observations to understand why people make the observed choices. A second advantage is that this method produces detailed information about individual transportation choices as well as location of activities. We estimate not only the origin and destination of each trip, but also the vehicle (be it car, bus, or train) used, and who the other passengers were. For a disease spread by casual, brief contact this is likely to be important.

Depending on the resolution chosen for activity locations, we may need to further partition the groups of people who meet in a location, with what we refer to as a "sublocation" model. In our model of Portland, for instance, a single activity location might be a large office building or university. We have developed an ad-hoc set of rules for partitioning these large locations. The rules are activity-specific. For example, people in a work location may visit the same workgroup every day, whereas shoppers might pick a different store at random each time they visit a shopping location. Note that, although our population is synthetic, the locations they visit are real.

In principle, this process could be repeated with different activity survey data to develop different typical days – weekends instead of weekdays or holidays instead of school days. We have not yet implemented this for EpiSims. Instead, we repeat the same typical day over and over in each simulation. It is also important to allow synthetic people to change their behavior in response to a disease outbreak in general or to becoming ill themselves. They might seek medical treatment or over-the-counter medications, withdraw to the home, or attempt to flee the city. Each of these actions would change their contacts and hence the possible path of an outbreak.

## Varieties of social networks

The structure that best represents our social network is a time-dependent, bipartite, labeled graph. We adopt a minimum time resolution, say one minute, and create a separate graph for each different time. Because EpiSims is nearly periodic in the absence of an outbreak, we might consider 1,440 graphs representing one day, one for each minute of the day. By definition, a bipartite graph can be partitioned into two subsets of vertices such that every edge goes from an element of one subset to an element of the other. In this case, one subset of vertices represents people and the other represents locations. In the most general case, we might consider a multipartite graph, with sets of vertices representing different concepts, perhaps related to a person's role in society. Locations may include such places as transit vehicles. "People" vertices are labeled with demographic attributes; "location" vertices, with geographic and land use attributes. An edge between two vertices represents that the corresponding person is currently in the corresponding location. Each edge is labeled with the reason for the visit. The edge sets change from one time to another. An alternative, more compact representation uses a single graph and includes the set of arrival and departure times in the label for each edge. On the time scales we are interested in (days, and weeks) we ignore changes in the vertex sets (due to births, deaths, migration, and construction).

Unfortunately, such graphs are too large to analyze easily. Even for a relatively small urban area such as Portland, there are 1,440 graphs, each with 1.6 million people vertices and 160,000 location vertices, and tens of millions of edges. It is easier to work with several projected versions of the graphs, in particular the static person-person graph and the static location-location graph. A static graph is produced from a time dependent set of graphs by including in the static graph every edge present in any of the time dependent graphs. Alternatively, one drops arrival and departure times from the labels and retains only total amount of time spent in each location. This projection onto static graphs discards distinctions between concurrent and sequential contacts that may be important for some diseases.

Two further projections are natural for bipartite graphs. The person-person graph includes an edge between two people if and only if they were present in the same location at the same time. The edge is undirected, and is labeled with the duration of contact between the people. The location-location graph includes an edge between two locations if and only if a person traveled from one location directly to another. The edge is directed, since there is a clear origin and destination for the movement.

The static person-person graph is what is generally intended by the term "social network" and our discussion below focuses on this form of the network, but the other forms are often useful. The location-location graph could be used to model how contamination is spread by humans as disease vectors. The full bipartite graph determines how a response targeted at locations affects people. It is important to note that the structural properties of the different graphs (described in the next section) can vary significantly among the different projections.

The social network summarizes how people move and come into contact with each other in ways that are important for studying the epidemiology of diseases spread through casual contact. An additional transformation completes the static picture for a particular infectious disease: replace the duration of contact labeling each edge with a probability of transmission in each direction, given that one of the people is infectious. This probability depends on the demographics of infectious and susceptible as well as the activity they are engaged in, and of course the natural history of the disease. For example, the probability of transmission from student to teacher during a given time interval may be very different from the probability of transmission from a worker to a co-worker in the same time interval.

## Characterizing networks for epidemiology

Simulation is provably the most efficient way to determine the detailed dynamics of systems of interacting discrete entities in an irregular network. Furthermore, simulations can naturally incorporate intervention strategies such as delivering treatment or prophylaxis and increasing social distance by closing schools or workplaces or limiting public gatherings. However, it is possible to develop some intuition by examin-

ing the static person-person network directly (7,8). The goals of this analysis are fourfold:

1. determine vertices and edges that contribute most to the propagation of disease;
2. determine structural properties of social networks that are most relevant to disease propagation;
3. measure those properties on the instances we have constructed;
4. construct constrained random graphs similar in the properties deemed relevant to disease propagation.

With this information we can create new random graphs that resemble social networks for the purposes of epidemiology. Such random graphs would be helpful for populating continental scale models or conducting sensitivity studies. We can also design effective response strategies that make efficient use of limited resources.

The problem of probabilistic transmission across a network is closely related to the problems of random walkers and percolation on a network. A random walker steps from one vertex to another by choosing an edge at random. The problem is to determine the distribution of vertices a random walker will reach after some number of steps. This problem has been studied in depth on regular networks such as lattices and on Erdös-Rényi random graphs, but the version needed for epidemiology is slightly different. Instead of a single walker taking random edges through network, an outbreak is better modeled by a set of random walkers who can potentially be created and destroyed at each step along the way. Furthermore, the social networks the walker must traverse are not well modeled by Erdös-Rényi graphs or lattices. Likewise, the percolation problem, which is to determine whether a path across a network exists as more and more edges or vertices are removed, has been thoroughly studied, but once again only for special network structures.

In this section, we describe some characterizations of network structure that may provide insight into the dynamics of disease outbreaks on social networks. The analysis here is restricted to the person-person static network, ignoring the weights along the edges. This is an extreme over-simplification of the problem, but even so yields a rich set of questions. Simulation naturally takes into account the full time-dependent weighted network but is less amenable to analysis.

Some of the measures we will discuss are distributions of a pointwise, or vertex-specific, statistic over all vertices. It is often convenient to define the pointwise statistic in terms of a local network centered on an index vertex. We organize the local network as shown in Figure 1a, by defining sets of vertices $S_k(v_0)$ at a graph distance $k$ from the vertex $v_0$.

Three local structural properties often discussed in this context are degree distribution, clustering coefficient, and assortativity. The degree of a vertex is the number of edges connected to it (in a graph with directed edges, vertices can have separate in and out degrees). The degree distribution is simply the frequency of occurrence for each degree over all the vertices. A graph whose degree distribution obeys a power law, that is, one for which the frequency of occurrence of degree $d$ is proportional to $d^{-\alpha}$, is said to be scale free, because the variance in degree diverges as the number of vertices increases for $\alpha <= 2$ (9). The clustering coefficient at each vertex is the ratio of the number of edges between its neighbors to the total possible number of edges. If the vertex's degree is $d$, then the possible number of undirected edges between its neighbors is $d\ (d-1)/2$. If there are actually $n$ edges, the clustering coefficient at that vertex is $2n/d(d-1)$. It describes the fraction of a person's contacts who come into contact with each other. Assortativity is the correlation between values of vertex labels at each end of an edge. For example, assortativity by degree is the correlation over all the edges of the degrees of the vertices connected by the edge. If high degree vertices are typically connected to other high degree vertices, the assortativity by degree is high. Figure 2a shows an example of two graphs with identical degree distributions but different assortative mixing by degree.

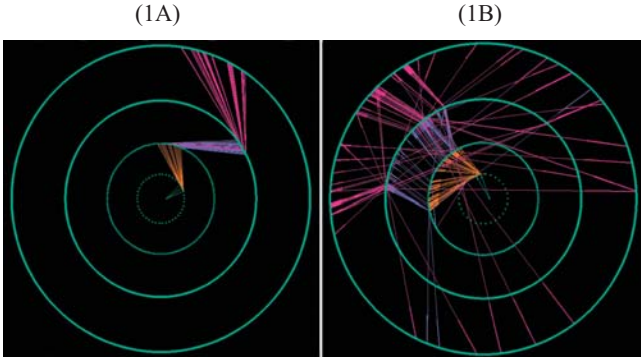


Fig. 1. Neighborhoods at increasing distance from a vertex $v_0$. Each vertex represents a specific synthetic person. Edges represent transmission from one person to another person and are colored to denote the transmission's generation. A) and B) show two different simulated outbreaks originating at $v_0$.

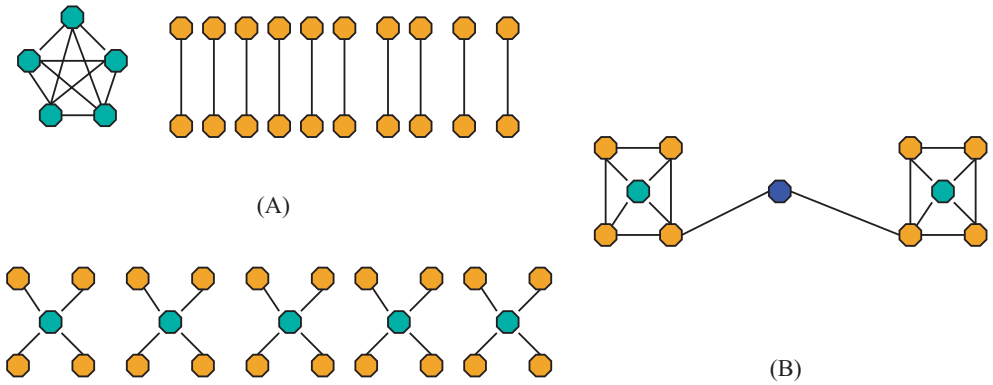


Fig. 2. A) top and bottom show two graphs with the same degree distribution (degree is 4 for each of the 5 green vertices and 1 for the 20 others) but different assortative mixing by degree. B) The blue vertex has a smaller degree than any other, but a high betweenness, because it is on the only path between two clusters.

The graphs we have constructed for Portland do not fall neatly into the scale free category. In fact, the degree distributions of the different projections described in the previous section are dramatically different. Conclusions drawn on the assumption that the people-people graph is scale free do not hold for our estimated social networks. To demonstrate this, we have studied the behavior of the network as people with high degree are removed. Contrary to some expectations, it does not become disconnected until a large fraction of people has been removed. Clustering and assortativity by degree are much higher than for an Erdös-Rényi random graph with a similar number of edges and vertices.

We turn now to global measures of structure. There are again several well-known candidates: diameter, shortest path distribution, expansion, and betweenness. The diameter of a graph is the maximum length of the shortest path between any pair of vertices. If the diameter grows slowly with the number of vertices, the graph is said to represent a small world, because there exists a short path between any pair of vertices (10). The diameter is an extreme statistic for the distribution of shortest path lengths. The distribution itself may hold more structural information than its maximum value. Expansion measures how rapidly neighborhoods grow. More formally, the vertex expansion is the minimum over all subsets (of fewer than half the vertices) of the ratio of the number of neighboring vertices to the number of vertices in the subset. Betweenness is best thought of as the number of shortest paths traversing a given vertex. Its value is high for a person who connects two otherwise non-interacting groups as in Figure 2b.

The diameter of our estimated social network is 6. This indicates that it is indeed a small world network, although since there is only one instance there is no way to determine how diameter scales with size. Our estimates for expansion indicate that social networks expand very rapidly, corresponding to potential extremely rapid spread of disease. Furthermore, they are robust against being broken into many disconnected pieces by any obvious strategy for deleting vertices or edges.

Unfortunately, because these measures are designed to be sensitive to global structure, they are very costly to compute, especially on such large graphs. In addition, betweenness is very sensitive to deletions of individual vertices or edges from the graph. These computational difficulties have led us to consider alternative measures. Defining a new measure presents an opportunity to incorporate two epidemiologically important aspects into the analysis of static graphs: the temporal interpretation of graph distance and the importance of distance from the initially infected people.

For an infectious disease, transmission involves a time delay during which the newly infected person incubates a disease and becomes infectious. There may be an additional delay as long as the length of the infectious period before transmission occurs. The speed with which an outbreak spreads through a population, and the speed with which response measures must be put in place, depend on both the number of people infected by each case and the generation time. Although the initial cases of an outbreak may be distributed randomly throughout the population, as the outbreak develops, infected people will represent a biased subset, chosen according to the mixing properties defined by the network. After $k$ generations of transmission, the set of people most likely to have been infected is related to the mixing properties of paths of length $k$.

Consider an outbreak beginning with a single infected person. The number of people exposed in the first generation will be the number of contacts, or degree, of that person. In the second generation, however, many contacts will be shared among several people, and several people who are exposed may have been contacts of the index case. In terms of the neighborhood sets defined above, the disease may either follow edges from $S_1$ to $S_2$ or from $S_1$ to $S_1$. These dynamics are related to clustering. High clustering means that the disease spreads outwards to ($S_k$ with larger $k$) from the index case more slowly, but that the probability of eventually reaching any fixed distance from the index case is increased. How these two dynamics are balanced depends on the disease and specifically on the probability of transmission. For very large transmission probabilities, the neighborhood $S_k$ will be entirely infected in generation $k$, hence the expansion of these subsets is crucial and clustering is irrelevant. For very low transmission probabilities, both clustering and expansion are important in determining how fast the disease spreads outward from the index case. We seek statistics that will generalize the structural properties discussed above from strictly local or global properties to intermediate scale properties.

We thus propose meso-scale versions of the most useful path based statistics described above: betweenness and expansion. A vertex's $k$-betweeness is the betweenness calculated using all paths of length $k$, i.e. the number of self-avoiding paths of length $k$ (differing by at least one vertex) that pass through a given vertex. We extend the definition to include all paths rather than just shortest paths because we are interested in any person who might be infected after exactly $k$ generations. We restrict to self-avoiding paths because for many diseases, an infected person becomes immune to another infection. Note that degree is the 1-betweeness. We define vertex $k$-expansion at a vertex $v_0$ as the number of vertices in $S_{k+1}$ divided by the number of vertices in $S_k$. To study the relation between clustering and expansion, we also define the relative edge $k$-expansion as the number of edges from $S_k$ to $S_{k+1}$ divided by the number of edges from $S_k$ that remain in $S_k$ or end in $S_{k-1}$.

It seems likely that distributions of these meso-scale statistics for a few values of $k$ will be most appropriate for characterizing social networks for epidemiology. As for global measures, it will be very hard to estimate these properties for a real person. There remain two reasons to consider such essentially unobservable properties. One is that we will use these structural properties only to compare social networks and thus need never estimate the property for any real person. The second is that we may be able to correlate the properties in our estimated networks with some more readily measurable property.

## ACKNOWLEGMENTS

We wish to thank Dr. Suzuki, Dr. Yamamoto, and the National Institute of Infectious Diseases for providing us the opportunity to meet with other interesting modelers.

This work was supported in part by a MIDAS grant from NIGMS.

## REFERENCES

1. Barrett, C., Eubank, S. and Marathe, M. (2005): "Modeling and simulation of large biological, information, and socio-technical systems: an interaction-based approach",

to appear in interactive computing: a new Paradigm. *In* D. Goldin, S. Smolka and P. Wegner (eds.). Springer Verlag.
2. Longini, I. M., Halloran, M. E., Nizam, A. and Y. Yang, Y. (2004): Containing pandemic influenza with antiviral agents. Am. J. Epidemiol., 159, 623-633.
3. Longini, I. M., Halloran, M. E., Nizam, A. and Y. Yang, Y. (2002): Containing bioterrorist smallpox. Science, 298, 1428-1432.
4. Ferguson, N. M., et al. (2003): Planning for smallpox outbreaks. Nature, 425, 681-685.
5. Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M., Srinivasan, A., Toroczkai, Z. and Wang, N. (2004): Modelling disease outbreaks in realistic urban social networks. Nature, 429, 180-184.
6. Barret, C. L., et al. (1999): Transportation Analysis Simulation System. Los Alamos Unclassified reports LAUR-99-1658, 99-2574 - 99-2579.
7. Barrett, C., Eubank, S., Marathe, M., Mortveit, H., Srinivasan, A. and Wang, N. (2004): Structural and algorithmic aspects of massive social networks. p. 718-727. Proceedings of the 15th Annual ACM-SIAM Symposium on Discrete Algorithms, New Orleans.
8. Eubank, S., Anil Kumar, V. S., Marathe, M., Srinivasan, A. and Wang, N.: Structure of Social Contact Networks and Their Impact on Epidemics, to appear in AMS-DIMACS Special Volume on Epidemiology.
9. Albert, R. and Barabasi, A.-L. (2002): Statistical mechanics of complex networks. Rev. Mod. Phys., 74, 47-97.
10. Watts, D. and Strongatz, S. (1998): Collective dynamics of small-world networks. Nature, 393, 440-442.