Statistical inference to advance network models in epidemiology

Penn State Department of Statistics Technical Report #11-01

David Welch[1,*,+], Shweta Bansal[2,*], David R. Hunter[1]

[1]Department of Statistics, 326 Thomas Building,
The Pennsylvania State University,
University Park, PA 16802.

[2]Center for Infectious Disease Dynamics, 208 Mueller Lab,
The Pennsylvania State University,
University Park, PA 16802.

*These authors contributed equally.

+ Corresponding author.

January 23, 2011

Abstract

Contact networks are playing an increasingly important role in the study of epidemiology. Most of the existing work in this area has focused on considering the effect of underlying network structure on epidemic dynamics by using tools from probability theory and computer simulation. This work has provided much insight on the role that heterogeneity in host contact patterns plays on infectious disease dynamics. Despite the important understanding afforded by the probability and simulation paradigm, this approach does not directly address important questions about the structure of contact networks such as what is the best network model for a particular mode of disease transmission, how parameter values of a given model should be estimated, or how precisely the data allow us to estimate these parameter values. We argue that these questions are best answered within a statistical framework and discuss the role of statistical inference in estimating contact networks from epidemiological data.

1 Introduction

It is not surprising that networks have been used to model the spread of infectious disease for over 50 years (Bailey, 1957; Dietz, 1967; Frisch and Hammersley, 1963). A network represents individuals in a host population as nodes and the interactions among them that may lead to the

transmission of disease as edges. The related ideas that a disease can spread from one individual to another via contact and that an individual can only have a limited number of contacts naturally lend themselves to this abstraction. With the increasing availability of data, computational power, and methodological advancement in the last two decades, the approaches of network theory have been increasingly sought for epidemiological modeling of human (Eubank et al., 2004; Meyers et al., 2005; Bansal et al., 2006), livestock (Kao et al., 2008; Kiss et al., 2006), and wildlife (Craft et al., 2009; Perkins et al., 2009; Hamede et al., 2009) disease systems.

The network-based studies to date have largely focused on the impact of network structure on disease dynamics and the effect of control strategies. Network structure has largely been determined by collecting host data to inform probabilistic models of host interactions, which are then used to generate simulated networks over which disease spread can be studied. It is not clear that this method leads to accurate models for prediction, or that the collected data and constructed models are always of relevance to the disease of interest. An alternative strategy is to statistically infer contact network models using all available host and disease data, including time series of incidence, known transmission chains or genetic pathogen sequences. This statistical approach, which has received relatively little attention, involves three stages (Gelman et al., 2004, chapter 1): specifying a probability model; fitting the model to observed data by using likelihood techniques to estimate model parameters and associated error; and evaluating the model fit, typically by simulation from the fitted model. While this alternative still allows the generation of networks that may be used to study disease spread, it alone provides additional insight about the level of information provided by the data regarding the particular choice of the generation method itself. In other words, while a wholly simulation-based approach is valuable and may inform the development of scientific hypotheses, only the statistical inferential approach allows us to test these hypotheses using data. As such, we argue that a statistical inferential framework has great scientific value for epidemiological network modeling.

The remainder of the paper is organized as follows. Section 2 defines a few crucial yet difficult terms like contact network and transmission network. In Section 3, we look at the role that explicit statistical models and inference can play in studying contact networks. Section 4 reviews how direct network data are gathered, provides an example of how such data may be used in the statistical framework and discusses how other forms of data (specifically, epidemiological and genetic data) can be integrated into this framework. Section 5 discusses future directions for this work. We emphasize that this paper focuses on the role of statistical inference in estimating contact networks from epidemiological data. As such, we touch on a wide range of well-studied topics in epidemiology and network theory without treating them fully. Topics not covered include network sampling strategies (Morris, 2004), missing data (Burt, 1987; Kossinets, 2006; Handcock

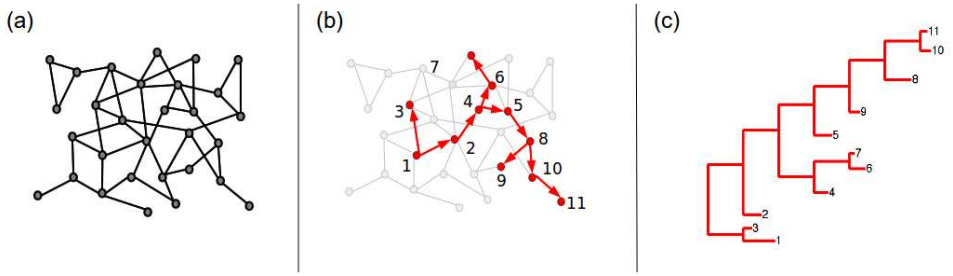

Figure 1: (a) A network of contacts (undirected), (b) the (directed) transmission network that could result from an epidemic and (c) the phylogenetic tree corresponding to the transmission network.

and Gile, 2010), spatial models (Ferguson et al., 2001; Chis Ster et al., 2009; Jewell et al., 2009) and the statistical estimation of epidemic parameters (in the absence of network parameters) from data (Bolker, 2008; Chowell et al., 2009; Bailey, 1975; Gibson, 1997; Streftaris and Gibson, 2004a,b; O'Neill, 2002).

2 Contact networks and transmission trees

The dominant models of mathematical epidemiology for the last century have been differential equation-based and entail the implicit assumption that all individuals (or groups of similar individuals) are equally likely to contact each other. These homogeneous mixing models have been used with a history of success in studying many aspects of disease dynamics (Anderson and May, 1991), yet they often fail to accurately capture population-level dynamics shaped by individual-level heterogeneities (Hethcote and Yorke, 1984; Meyers et al., 2005; Bansal et al., 2007). Network-based models provide an elegant alternative to homogeneous-mixing models by intuitively capturing diversity in the underlying patterns of interaction in a population.

We do not consider here models that extend homogeneous mixing models by including population structure such as age stratification or spatial heterogeneity without specifying an explicit contact network. While there is a considerable body of work fitting these non-network models and there are analogous network models allowing for these structural or spatial features, explicit network models present different statistical challenges from those encountered with non-network spatial or structured population models.

2.1 Defining contact networks

A *contact network* is a network (or graph) in which nodes (or vertices) represent individual hosts and the edges (or ties) connecting pairs of nodes represent potentially disease-causing contacts. This definition is intentionally general; the notion of a *contact* varies with the host, pathogen and transmission route in question. Where transmission of one infection might require sexual intercourse between two individuals, another might require nothing more than physical proximity. We assume here that the drivers of transmission for a particular disease are well-understood, an important consideration since this determines whether a contact can be well-defined. Examples of studied contact networks range from the spread of syphilis among a heterosexual human population (Patrick et al., 2002), to the spread of influenza in an urban area (Bansal et al., 2006), to the the spread of canine distemper virus among lions in the Serengeti (Craft et al., 2009). These different modes of transmission necessitate very different types of contact networks. Conversely, diseases that share a mode of transmission in the same host population, and thus the same notion of contact, would be expected to share a common contact network.

We emphasize that a contact between two hosts, one infected and one susceptible, does not imply that infection is transmitted. Thus, a contact network necessarily includes any edges that result in a transmission, but may also include edges that do not. Furthermore, we assume that the contact network is blind to the disease status of any of its individuals; here, we do not consider the complication that behavior may be influenced by disease status. Thus, the concept of a contact, in the context of some disease, has meaning whether or not the individuals involved are an infected/susceptible pair. Contact networks can also be defined at scales other than that of an individual. For example, in the context of a livestock disease, a network node might logically be defined as an entire farm (e.g., Kao, 2002), while large-scale disease dynamics might be studied with a network of cities as nodes, connected by the movement of individuals between them (e.g., Colizza et al., 2006).

Perhaps the most obvious objection to our notion of contact network is that the network is static; that is, an edge between two nodes is either always present or absent, and that temporal features such as duration, order or frequency are not accounted for. In contrast, in most human and animal diseases, contacts between individuals are fleeting, constantly formed and broken. Ideally, a model of the contact network would explicitly account for this evolution through time within a dynamic network model such as those discussed in Snijders and Doreian (2010); Bansal et al. (2010) and references therein. Yet a static network is a natural place to begin model specification, and generalization to dynamic networks may turn out to be straightforward in some cases; when it is not, it may require additional theory or more nuanced data, as discussed by Krivitsky (2009). We focus on static models for the rest of this work.

2.2 Defining transmission networks

A contact network describes the set of contacts via which infection is possible, but does not give information on contacts that lead to successful transmission of disease. When a pathogen is introduced into a population and an outbreak occurs, the pathogen follows a path on the contact network as it spreads from node to node, traversing some edges and not others. This path is directed, since each transmission must occur from an infected individual to an uninfected one. This path is itself a network, called a *transmission network*, and is defined on the same set of nodes as the contact network but for which a directed edge occurs from A to B if and only if A and B share an edge in the contact network and A transmits disease to B. Thus, the transmission network is a subgraph of the contact network. Furthermore, for many diseases it can be assumed that every individual may be infected only once and by exactly one other individual within a single epidemic, so the transmission network is a tree, i.e., a network without cycles. We illustrate the distinction between a contact network and a transmission network in Figure 1.

3 Statistical approaches to modeling networks

Models are used widely in the study of networks and are not the sole preserve of a statistical approach. All studies involving repeated stochastic simulation of networks must choose some model for these networks. Sometimes this choice is explicit, while other times it is implicit. For instance, a "random network" model (which is more accurately called a uniform random network model), in which each possible network is equally likely, is an explicit choice: It attaches a closed-form probability to each possible network. On the other hand, while a network-simulation method based on preferential attachment as in Barabasi and Albert (1999) imposes a probability model on the set of all possible networks, this model is implicitly defined through the stochastic rules for constructing the network edge by edge rather than explicitly specified. Similarly, large agent-based simulations such as those of Eubank et al. (2004); Barrett et al. (2008); Ferguson et al. (2005) impose implicit probability models on the space of possible networks, but the complexity of these models often prevents explicit description, let alone statistical inference.

The distinction between the simulation or probability paradigm and the statistical paradigm is summarized by the description of statistics as "probability in reverse." To wit, where probability or simulation studies start with parameters and a model and describe how data will behave, statistical inference starts with data and a model and describes what can be said about the parameters. The relationship between the two approaches is further illustrated in Figure 2.

While complete knowledge of a specific contact network associated with a particular population might be interesting, we stress that discovery of this single network is not our aim. Rather, we seek

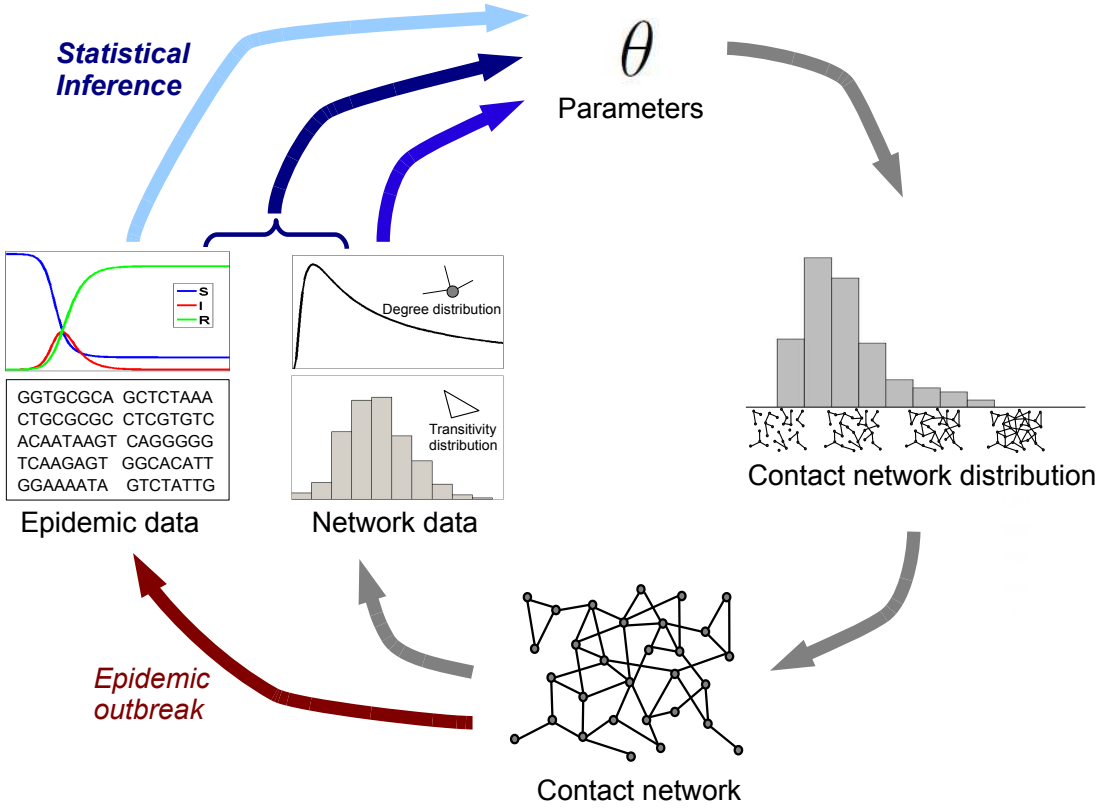

Figure 2: θ is a set of parameters that determine a distribution over all possible contact networks. A particular contact network is drawn from this distribution. Network data, such as degrees of sampled nodes, may be observed directly. An epidemic on the network results in epidemic data, indicated by the red arrow. Examples are prevalence time series and pathogen genetic sequences. This forward process is described by a probability model and may be simulated. Statistical inference (blue arrows) seeks to take the data and reverse the process to estimate θ. Inference may be based on network data (mid-blue arrow, see Section 4.2), or epidemic data (light blue, Section 4.3). A combined approach (dark blue) that uses all available data to estimate network parameters, as discussed in Section 4.4.

a simplified description — a model — of a stochastic process that could realistically have resulted in this network. Indeed, for many scientific purposes, knowing a model for a contact network is more useful than knowing the specific contact network itself, since the former is useful to study general behavior across an ensemble of similar contact networks, whereas the latter is not. Furthermore, the goal is not merely to estimate a single model for simulating networks but also to know how precise the model estimate is given the data; such information is completely absent even in the case of a perfectly observed true contact network. Of course, perfect data on a network may be exploited in statistical inference; an enormous literature in social networks describes exactly this process. (See Box 1 and articles such as Robins et al. (2006), Goldenberg et al. (2009), and the references therein.)

In the statistical paradigm, we also attempt to quantify the uncertainty inherent in using incomplete data to choose a model. Though either paradigm can be used to simulate contact networks on which to study the spread of disease, only the former leads to a measure of the information contained in the data about the specificity with which we may claim to understand how contact networks are generated. Furthermore, in statistics, the parameters we associate with features of the model allow us to learn about various aspects of the network-generating mechanism. That is, the models are descriptive rather than merely predictive. While it is possible to reduce data through the use of descriptive summary statistics, it is only through a model and its associated parameter estimates that we may summarize the random behavior of a complex system such as a contact network. And while it is likely that certain problems will forever remain beyond the scope of inferential methods (e.g., large agent-based simulation models), there have been recent advances in statistical methodology that show some promise. We discuss these advances in Section 4.

4 Inferring contact networks from data

4.1 Data collection for contact networks

Because contact networks are defined in terms of individuals in a population and encode every contact among the individuals that is relevant to the infection in question, collecting data to determine complete contact networks can be a herculean task. In addition, issues of privacy, misreporting and sampling biases can complicate the task (Keeling and Eames, 2005). Despite these drawbacks, some techniques are used to gather data about the host population and contact network in the absence of disease.

Methods for sampling contact network data in the absence of disease can be broadly classified as either direct or indirect. We outline some of the main techniques below.

Direct techniques, primarily used for humans only, such as diary-based studies or device-based

studies focus on collecting explicit information on contact behavior among individuals that is relevant to the disease in question. In diary-based studies, a population sample is selected to record contacts as they occur (Keeling and Eames, 2005). Definitions of contact vary in these studies: in the context of infection that can spread by respiratory droplets, an in-person two-way conversation of more than two words has been used (Mossong et al., 2008; Wallinga et al., 2006; Edmunds et al., 1997; Read et al., 2008), while sexual intercourse has been used in the context of sexually-transmitted diseases (Ghani and Garnett, 1998, 2000). Device-based studies rely on the use of electronic recording devices, usually to measure proximity among individuals, as a proxy for contact for the spread of airborne or respiratory droplet infections. RFID tags (Barrat et al., 2008), animal ear tags (Kao et al., 2008), and cellular phones (Gonzalez et al., 2008) have all been employed in these studies.

Diary-based and device-based studies allow for detailed data collection on individuals, but they may be limited in their spatial and temporal scope. In addition, the epidemiological relevance of the measured contact must be carefully understood in these studies. Other sources of host data are less direct and involve the use of data collected on general human or animal behavior. Examples of this include transportation data (Eubank et al., 2004; Colizza et al., 2006), census data on population age and household size distributions, and data on school attendance, employment or hospital occupancy (Meyers et al., 2005; Eubank et al., 2004; Halloran et al., 2002), social structure or mating behaviors (Craft et al., 2009; Hamede et al., 2009). These independent data can guide us about individual behavior and allow us to consider disease spread at larger spatio-temporal scales, but may not alone be specific enough for reconstructing individual-level contact networks.

4.2 Inference about contact networks from contact data

As illustrated in Figure 2, network data of the form described above, can be fit to a specified probability model to attain model parameters and estimates of model fit. In Box 1, we provide a simple example of inference of a network model from observed contact data. The statistical inferential procedure we have employed in this example gives not only an estimate of the parameter β_0 and a sense of how precise this estimate is, but also a natural way to obtain new, randomly generated networks and to check the model itself. The former is accomplished because we now have an explicit probability model on the space of possible networks, whereas the latter may be accomplished by simulating repeatedly from the model and then using descriptive statistics to compare the simulated networks with the original network to determine whether the observed network appears unusual relative to the population of networks implied by the fitted model. For details on this model-checking procedure, see Hunter et al. (2008). This inherent model-checking capability is a major advantage of the statistical approach, as is the fact that formulating a model

forces one to be explicit about the assumptions made about the random network-generating process.

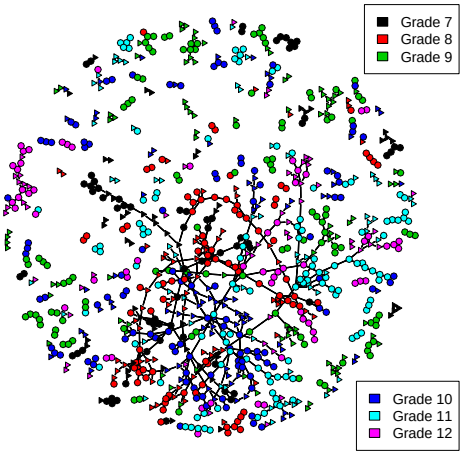

Figure 3: A network of mutual friendship relationships among 1461 junior-high- and high-school students in the southern United States. Isolates (nodes of degree zero) are not shown here. Nodes are colored according to grade level and shaped according to sex (triangles are male and circles are female).

4.3 Inference about contact networks from disease data

Given the tenets of the statistical approach, it is natural to ask what can be said about the parameters governing an underlying contact network by observing the progress of an epidemic. From disease data (e.g., event times) something can be said about who was likely to have infected whom, which is the information needed to construct the transmission network, and in turn infer something about the contact network. Finally, given a contact network, an estimate of the parameters of the network model can be made. The relationship between each of these levels of information can be codified in a likelihood. However, much of the technical difficultly of this problem lies in the fact that to calculate the likelihood of the data for any given parameter values, it is necessary to include, as auxiliary variables, the transmission network and the contact network itself.

One paper that addresses the network problem is Britton and O'Neill (2002), which demonstrates that the parameters of a network model can be estimated given infection and/or recovery times of individual hosts. The paper describes a Bernoulli, or Gilbert-Erdős-Rényi, model for the contact network in which any pair of nodes is connected with probability, p, independently of

Suppose that we observe the network shown in Figure 3 and we wish to fit a class of models to it. Our example network describes the social relationships between students in grades 7 through 12 at a large school community in the southern United States. Specifically, each edge represents a self-reported mutual friendship between two individuals. Though it is not an epidemiological contact network strictly speaking, we consider social relationships here to be a proxy for disease-causing contact for the purposes of illustration. Here we illustrate a method of statistical network fitting that is well-known in the social networks literature and that could be applied to the study of contact networks in cases where they might be observed. Recall that in such cases, the specific observed contact network is not of as much interest as a useful model that might realistically have given rise to the specific network.

This network is based on data collected in the AddHealth study of Resnick et al. (1997). For reasons of confidentiality, the network depicted is not the original network but rather a version simulated from a statistical model based on the real network. (See the documentation on the faux.magnolia.high network in the ergm package (Handcock et al., 2010) for the R computing environment (R Development Core Team, 2009) for more details on this network dataset.) We illustrate the use of exponential-family random graph models (ERGMs) to fit this network dataset (Robins et al., 2006). The model we fit, which is discussed along with other models for the same dataset in Section 5 of Goodreau et al. (2008), assumes a model in which each possible pair of nodes, say i and j, has an edge (a mutual friendship) with probability p_{ij}, where

$$\begin{aligned}\log\left(\frac{p_{ij}}{1-p_{ij}}\right) &= \beta_1 + \beta_2 I\{i \text{ and } j \text{ are in the same grade}\} \\ &\quad + \beta_3 I\{i \text{ and } j \text{ are of the same race}\} \\ &\quad + \beta_4 I\{i \text{ and } j \text{ are of the same sex}\}.\end{aligned}$$

In the expression above, $I\{\}$ is an indicator function, i.e., it equals 1 if the argument is true and 0 otherwise. In this model, each edge is independent of all others, so that the probability that a random network (denoted Y) equals a given fixed network denoted by y may be expressed as

$$P(Y = y) = \kappa \exp\{\beta_1 E(y) + \beta_2 G(y) + \beta_3 R(y) + \beta_4 S(y)\}, \quad (1)$$

where $E(y)$ is the total number of edges in y and $G(y)$, $R(y)$, and $S(y)$ are the number of edges in y between nodes of the same grade, race, and sex, respectively. In Equation (1), the κ is a normalizing constant (a function of the β_i but not y) that ensures that Equation (1) defines a legitimate probability distribution.

We may estimate the parameters in model (1) using the method of maximum likelihood, whereby we search for the values of $\beta_1, \ldots, \beta_4$ that maximize equation (1) given y is the observed network data. These maximizers, which we denote by the vector $\hat{\beta}$, serve as the maximum likelihood estimators of the true parameters, which we denote by the vector β_0 and which is typically unknown. In our example, plugging in the observed values $E(y) = 974$, $G(y) = 820$, $R(y) = 787$, and $S(y) = 689$, we may obtain the following estimates (with use of the ergm package for R):

```
                 Estimate Std. Error  p-value
edges           -10.01277    0.11526   <1e-04
nodematch.Grade   3.23105    0.08788   <1e-04
nodematch.Race    1.19646    0.08147   <1e-04
nodematch.Sex     0.88438    0.07057   <1e-04
```

From an epidemiological point of view, these results tell us, for example, that the grade, race, and sex of an individual are all important in determining his or her patterns of contacts on average; in particular, we estimate that individuals are from $\exp\{0.88\} = 2.4$ (in the case of sex) to $\exp\{3.23\} = 25.3$ (in the case of grade) times more likely to associate with other individuals of the same category. This information might help us to formulate an intervention strategy if an individual were to become infected. Note that approximate standard error estimates are also calculated along with the β estimates themselves, which allows for (say) standard hypothesis tests and confidence intervals.

Box 1: An example of inference from an observed network

all other node pairs (Gilbert, 1959). An SIR epidemic process with exponential waiting times is assumed over the contact network with transmission rate β and recovery rate γ. The three parameters p, β and γ are estimated for various small data sets by the authors using Markov chain Monte Carlo (MCMC) methods. We emphasize that no direct observation of any part of the contact network is utilized in the estimation process employed. Their results indicate that all these parameters can be simultaneously estimated, although there exists a strong correlation between the network parameter p and the transmission parameter β in some parts of the parameter space. This correlation is such that the product $p\beta$ can always be estimated but the individual factors may be unidentifiable, making it difficult to distinguish between a mildly transmissible disease (low β) on a highly connected network (high p) from a highly transmissible infection (high β) on a sparse network (low p).

Others have built on the model and methodology of Britton and O'Neill (2002). Neal and Roberts (2005) focus on the statistical methodology of the problem, introducing potential computational efficiencies to the MCMC method. Ray and Marzouk (2008) extend the graph model so that sub-populations may have varying degrees of contact with each other, extend the epidemic model to include a latent period (making it a susceptible-exposed-infected-recovered, or SEIR, model), and use gamma-distributed latent and infectious periods. From a modeling perspective, these are both useful extensions, but the authors have some difficulty fitting this extended model to data. A more successful approach has recently been demonstrated by Groendyke et al. (2010), who also investigate the identifiability of the contact parameter p and transmission parameter β and show that for a significant portion of the (realistic) parameter space these parameters are indeed identifiable.

In fitting a compartmental model with stochastic dynamics (Ball et al., 1997; Anderson and May, 1991, chapters 3 and 4), Demiris and O'Neill (2005) impute a type of transmission network in order to aid the likelihood calculation for their model. The model they consider is of a population divided into households where individuals in the same household transmit disease to one another at a higher rate than individuals in different households. The imputed network is a directed network containing all contacts made by an individual during its infectious period. This network is clearly more general than a transmission network—it contains many possible transmission networks—but is not quite as general as a contact network in the sense that we are using that term. It could perhaps be thought of as an emergent network of the disease as described by Keeling and Eames (2005), where the underlying population is panmictic yet when contacts are recorded over a short period of time we see a relatively sparse network structure emerge. The network itself is not the object of interest in this problem, but instead is a nuisance parameter which must be dealt with so that the contact rates within and between groups can be calculated.

The papers discussed here all suffer to some extent from a paucity of data relative to the number of unknown variables in the respective models.Britton and O'Neill (2002) and related papers rely on a limited number of event times (it is highly unusual to have all infection and recovery times) while Demiris and O'Neill (2005) is based on having final outcome data for an epidemic (which is, again, unusual to have in a complete form in practice). At best in the Britton and O'Neill (2002) case, there are $2N$ data points required (i.e., the infection and recovery times for all N hosts in a population), yet the number of variables is $3 + N(N-1)/2$ (i.e., the model parameters, β, γ and p, and $N(N-1)/2$ possible network edges that are considered latent variables here). Even when the presence or absence of a particular edge in the network is not of interest, it is necessary to include these extra variables so that the likelihood can be calculated. An obvious solution to this rapid growth in the number of unknowns is to narrow down the space of plausible networks by bringing additional data to bear on the problem. We discuss this next with particular reference to genetic and host data.

4.4 Integrating multiple sources of data to infer networks

Genetic sequences taken from pathogens are potentially informative about the transmission network, especially where the pathogen is a rapidly evolving RNA virus (Pybus and Rambaut, 2009). Although epidemiological data may provide information on who was infected, when, and how long, it cannot provide positive information on who acquired infection from whom. Comparison of pathogen sequences taken from different hosts makes it possible to infer the most likely infector for a given infectee, providing an additional constraint on the space of possible transmission trees. Typically, genetic sequences are treated within the framework of phylogenetic analyses.

There exist two clear analogies between epidemiological models and phylogenetic models. First, if genetic samples are taken from pathogens, the related phylogenetic tree contains much of the same information as the associated transmission network for those pathogens; see Figure 1. The phylogenetic tree shows the time of the most recent common ancestor for any subset of sampled sequences. The extra information that a transmission tree displays is the direction of transmission at each ancestral node in the phylogenetic tree. Studies such as Cottam et al. (2008), which combines epidemiological and genetic data to infer transmission trees from the 2001 UK foot and mouth disease outbreak, and Lewis et al. (2008), which uses a Bayesian approach to reconstruct transmission networks of HIV patients from London, show that genetic data can greatly aid in the process of reconstructing transmission networks.

Second, basic population genetic models assume a panmictic population but, as in epidemiology, the panmictic assumption has been extended to a structured population with a constant rate of contact/migration among the sub-populations (Donnelly and Tavare, 1995). Recently, more com-

plex models of population structure have emerged in population genetics, some with a distinctly network flavor (Lemey et al., 2009). Applications of phylogenetics to epidemiology have informed estimates of the prevalence, rates of spread, and time of origination of various epidemics and provided useful information about the evolution of pathogens (Goodreau, 2006; Holmes and Grenfell, 2009). Similarly, the introduction of genetic data to network models should improve the estimation of transmission networks and, therefore, of contact networks.

Data sources directly relating to the host population and individuals, discussed in Section 4.1, could also be used to inform these models along with epidemiological data (discussed in Section 4.3). Census data, location data and host covariates —age, occupation, socio-economic status— should not be treated separately from epidemiological data but should play an important role in defining and constraining the underlying contact network model. This could be achieved via direct inclusion of these data in the estimation process, or, within a Bayesian framework, by constructing informative priors that take these data into account. Although limited in its size, one study that successfully makes use of host data, epidemiological data, and molecular data is that of Spada et al. (2004). The authors use a minimum spanning tree approach on molecular data, combined with information about the contact patterns of the hosts (e.g. colocation of patients within hospital wards), to reconstruct a transmission tree for a hepatitis C outbreak.

5 Discussion

Epidemiology has much to gain from the use of networks, from a deeper understanding of the impact of heterogeneities on the ecology and evolution of host-pathogen systems, to the exploitation of network processes for the design of efficient intervention strategies. Any of these goals requires the use of accurate network models which should be informed by as much of the data as possible. In this article, we have argued that most of the existing work in network epidemiology has focused heavily on the probabilistic question of what influence specific network structure may have on disease dynamics. This work is valuable in many respects including identifying forms of heterogeneity that may aid or impede epidemics, designing epidemic models that capture observed dynamics, and assessing the impact of control strategies. We argue further that to gain a full understanding of the structure of contact networks and of the network characteristics which influence the outcome of an epidemic, a statistical approach should be taken when possible. This involves starting with all available data—those from one or more observed epidemics and those relating to the contact process—and then fitting as much of the data as possible to a model to produce parameter estimates and, crucially, estimates of the associated errors.

In Section 3, we explained how diverse data could be treated within a statistical framework. In Section 4.1, we discussed a wide range of data sources regarding host behaviors that may generate contact networks, and demonstrated statistical inference on such data with an example. The challenge of introducing epidemic data into this framework has been broached by some preliminary studies as discussed in Section 4.3. While this work is promising, the existing models are relatively simplistic and do not take full advantage of all available data. This can be explained partially by the fact that dealing with even one type of data is difficult, can be computationally challenging and may require certain simplifying assumptions that are less appropriate for other forms of data. While it is reasonable that novel methods might only work with a particular type of data, to become more broadly useful, there is a need for models that incorporate as much available information as possible.

A larger problem is that we typically require complete data to make estimates of network parameters. This is partly due to the structure of network models; when nodes interact in a heterogeneous manner, it may be necessary to model the behaviors of each individual node and, thus, the entire network. Under current methods, the properties of any nodes that are not observed but may have played a role in the spread of the infection need to be imputed. This rapidly leads to an explosion in the number of unknown variables and hobbles current methods. Thus, network models that work with incomplete data (Handcock and Gile, 2010) and that do not require the unobserved data to be imputed need to be further developed in the context of contact networks for infectious disease epidemics.

If these technical problems can be overcome, we would expect to see more accurate models informed by data leading to a deeper understanding of host-pathogen systems and epidemiology generally. Indeed, informing network models with epidemic data presents intriguing possibilities in those cases beyond epidemiology where we may be interested in the contact network itself, or a larger social network in which the contact network is embedded. For instance, we may wish to study the social structure of a particular population, and we merely use disease transmission data as a means for learning about this structure. While human populations and their social networks can often be studied directly, the same does not apply to populations of other species. In these cases, important information regarding levels of contact among widely dispersed populations is of utmost interest to researchers and could be further informed by the studying the contact networks associated with transmissible pathogens present within the populations of interest. The epidemic here can be viewed as a probe which passes through the population of interest and can then be studied to examine features of the population itself.

Acknowledgments

The authors are grateful to two anonymous reviewers and several colleagues who offered helpful comments and suggestions on early drafts of this manuscript: Michael Schweinberger, Matt Ferrari, Mary Poss, and Nicole Carnegie. This work is supported by NIH grant R01-GM083603-01; and by the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security, and the Fogarty International Center, National Institutes of Health.

References

Anderson, R. and May, R. (1991). *Infectious diseases of humans.* Oxford University Press, London.

Bailey, N. (1957). *The mathematical theory of epidemics*. Griffin, London.

Bailey, N. (1975). *The Mathematical Theory of Infectious Diseases and its Applications*. Griffin, London.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Ann. App. Prob*, 71:46–49.

Bansal, S., Grenfell, B., and Meyers, L. (2007). When individual behavior matters. *J. R. Soc. Interface*, 4(16).

Bansal, S., Pourbohloul, B., and Meyers, L. (2006). Comparative analysis of influenza vaccination programs. *PLoS Medicine*, 3.

Bansal, S., Read, J., Pourbohloul, B., and Meyers, L. (2010). The dynamic nature of contact networks in infectious disease epidemiology. *Journal of Biological Dynamics*.

Barabasi, A. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, 286:509–512.

Barrat, A., Cattuto, C., Colizza, V., Pinton, J.-F., den Broeck, W. V., and Vespignani, A. (2008). High resolution dynamical mapping of social interactions with active rfid. *arXiv.org*, 0811.4170.

Barrett, C. L., Bisset, K. R., Eubank, S. G., Feng, X., and Marathe, M. V. (2008). Episimdemics: an efficient algorithm for simulating the spread of infectious disease over large realistic social networks. In *SC '08: Proceedings of the 2008 ACM/IEEE conference on Supercomputing*, pages 1–12, Piscataway, NJ, USA. IEEE Press.

Bolker, B. (2008). *Ecological Models and Data in R.* Princeton Press.

Britton, T. and O'Neill, P. (2002). Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure. *Scandinavian Journal of Statistics*, 29(3):375–390.

Burt, R. (1987). A note on missing network data in the general social survey. *Social Networks*, 9:63–73.

Chis Ster, I., Singh, B. K., and Ferguson, N. M. (2009). Epidemiological inference for partially observed epidemics: The example of the 2001 foot and mouth epidemic in great britain. *Epidemics*, 1(1):21 – 34.

Chowell, G., Hayman, J., Bettencourt, L., and Castillo-Chavez, C., editors (2009). *Mathematical and Statistical Estimation Approaches in Epidemiology*. Springer.

Colizza, V., Barrat, A., Barthelemy, M., and Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS*, 103:2015–22.

Cottam, E. M., Thebaud, G., Wadsworth, J., Gloster, J., Mansley, L., Paton, D. J., King, D. P., and Haydon, D. T. (2008). Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proc B*, 275:887–95.

Craft, M., Volz, E., Packer, C., and Meyers, L. (2009). Distinguishing epidemic waves from disease spillover in a wildlife population. *Proc R Soc B*, 276 (1663):1777–1785.

Demiris, N. and O'Neill, P. D. (2005). Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Statistical Society Series B*, 67(5):731–745.

Dietz, K. (1967). Epdiemics and rumours: a ruvey. *J R Stat Soc A*, 130:505–28.

Donnelly, P. and Tavare, S. (1995). Coalescents and genealogical structure under neutrality. *Annual Review of Genetics*, 29(1):401–421.

Edmunds, W., O'Callaghan, C., and Nokes, D. (1997). Who mixes with whom? a method to determine the contact patterns of adults that may lead to the spread of airborne infection. *Proc B*, 264:949–57.

Eubank, S., Guclu, H., Kumar, V., Marathe, M., Srinivasan, A., Toroczkai, Z., and Wang, N. (2004). Modeling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184.

Ferguson, N., Cummings, D., Cauchemez, S., Fraser, C., and Riley, S. (2005). Strategies for containing an emerging influenza pandemic in southeast asia. *Nature*, 437.

Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001). Transmission intensity and impact of control policies on the foot and mouth epidemic in great britain. *Nature*, 413:452–548.

Frisch, H. and Hammersley, J. M. (1963). Percolation processes and related topics. *J. SIAM*, 11(894918).

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC, second edition.

Ghani, A. C. and Garnett, G. P. (1998). Measuring sexual partner networks for transmission of sexually transmitted diseases. *Journal of the Royal Statistical Society A*, 161:227–238.

Ghani, A. C. and Garnett, G. P. (2000). Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks. *Journal of the Royal Statistical Society ASexually Transmitted Diseases*, 27:579–587.

Gibson, G. (1997). Markov chain monte carlo methods for fitting spatiotemporal stochastic models in plant epidemiology. *Appl. Statist.*, 46:215–233.

Gilbert, E. N. (1959). Random graphs. *The Annals of Mathematical Statistics*, pages 1141–1144.

Goldenberg, A., Zheng, A., Fienberg, S., and Airoldi, E. (2009). A survey of statistical network models. *Foundations and Trends in Machine Learning, to appear*.

Gonzalez, M., Hidalgo, C., and Barabsi., A.-L. (2008). Understanding individual human mobility patterns. *Nature*, 453:479–82.

Goodreau, S. M. (2006). Assessing the effects of human mixing patterns on HIV-1 interhost phylogenetics through social network simulation. *Genetics*, 172:2033–2045.

Goodreau, S. M., Handcock, M. S., Hunter, D. R., Butts, C. T., and Morris, M. (2008). A statnet tutorial. *Journal of Statistical Software*, 24(9):1.

Groendyke, C., Welch, D., and Hunter, D. R. (2010). Bayesian inference for contact networks given epidemic data. Technical Report 10-02, Pennsylvania State University Department of Statistics.

Halloran, M., Longini, I., Nizam, A., and Yang, Y. (2002). Containing bioterrorist smallpox. *Science*, 298:1428–1432.

Hamede, R., Bashford, J., McCallum, H., and Jones, M. (2009). Contact networks in a wild tasmanian devil (sarcophilus harrisii) population: using social network analysis to reveal seasonal variability in social behaviour and its implications for transmission of devil facial tumour disease. *Ecology Letters*, 12 (11):1147–57.

Handcock, M. and Gile, K. (2010). Modeling social networks from sampled data. *Annals of Applied Statistics*, 4(1):5–25.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Morris, M., and Krivitsky, P. (2010). *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*. Seattle, WA. Version 2.2-4. Project home page at http://statnet.org.

Hethcote, H. and Yorke, J. (1984). Gonorrhea transmission dynamics and control. *Springer Lecture Notes in Biomathematics*, 56.

Holmes, E. and Grenfell, B. (2009). Discovering the phylodynamics of rna viruses. *PLoS Comput Biol*, 5(10).

Hunter, D., Goodreau, S., and Handcock, M. (2008). Goodness of fit of social network models. *Journal of the American Statistical Association*, 103(481):248–258.

Jewell, C., Kypraios, T., Christley, R., and Roberts, G. (2009). A novel approach to real-time risk prediction for emerging infectious diseases: A case study in avian influenza h5n1. *Preventive Veterinary Medicine*, 91(1):19 – 28. Special Issue: GisVet 2007.

Kao, R. (2002). The role of mathematical modelling in the control of the 2001 FMD epidemic in the UK. *Trends in Microbiology*, 10(6):279–86.

Kao, R., Danon, L., Green, D., and Kiss, I. (2008). Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. *Proceedings of the Royal Society B*, 273 (1597).

Keeling, M. and Eames, K. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, 2(4):295.

Kiss, I. Z., Green, D. M., and Kao, R. R. (2006). The network of sheep movements within great britain: network properties and their implications for infectious disease spread. *Journal of the Royal Society Interface*, 3(10):669–77.

Kossinets, G. (2006). Effects of missing data in social networks. *Social Networks*, 28:247–268.

Krivitsky, P. N. (2009). *Statistical Models for Social Network Data and Processes*. PhD thesis, University of Washington. unpublished.

Lemey, P., Rambaut, A., Drummond, A., and Suchard, M. (2009). Bayesian phylogeography finds its roots. *PLoS Comput Biol*, 5(9).

Lewis, F., Hughes, G. J., Rambaut, A., Pozniak, A., and Brown, A. J. L. (2008). Episodic sexual transmission of hiv revealed by molecular phylodynamics. *PLoS Medicine*, 5 (3).

Meyers, L., Pourbohloul, B., Newman, M., Skowronski, D., and Brunham, R. (2005). Network theory and sars: predicting outbreak diversity. *J. Theo. Biol*, 232:71–81.

Morris, M., editor (2004). *Network Epidemiology : A Handbook for Survey Design and Data Collection*. Oxford University Press.

Mossong, J., Hens, N., Jit, M., Beutels, P., and Auranen, K. e. a. (2008). Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Medicine*, 5:3.

Neal, P. and Roberts, G. (2005). A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing*, 15(4):315–327.

O'Neill, P. (2002). A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical Biosciences*, 180:103–114.

Patrick, D. M., Rekart, M. L., Jolly, A., Mak, S., Tyndall, M., Maginley, J., Wong, E., Wong, T., Jones, H., Montgomery, C., and Brunham, R. C. (2002). Heterosexual outbreak of infectious syphilis: epidemiological and ethnographic analysis and implications for control. *Sexually Transmitted Infections*, 78(suppl 1):i164–i169.

Perkins, S., Cagnacci, F., Stradiotto, A., Arnoldi, D., and Hudson, P. (2009). A comparison of social networks derived from ecological data: implications for inferring infectious disease dynamics. *Journal of Animal Ecology*, 78:1015–22.

Pybus, O. G. and Rambaut, A. (2009). Evolutionary analysis of the dynamics of viral infectious disease. *Nat Rev Genet*, 10(8):540–550.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Ray, J. and Marzouk, Y. M. (2008). A Bayesian method for inferring transmission chains in a partially observed epidemic. In *Proceedings of the Joint Statistical Meeting*.

Read, J., Eames, K., and Edmunds, W. (2008). Dynamic social networks and the implications for the spread of infectious disease. *J R Soc Interface*, 5(26):1001–1007.

Resnick, M. D., Bearman, P. S., Blum, R. W., Bauman, K. E., Harris, K. M., Jones, J., Tabor, J., Beuhring, T., Sieving, R. E., Shew, M., Ireland, M., Bearinger, L. H., and Udry, J. R. (1997).

Protecting adolescents from harm. findings from the national longitudinal study on adolescent health. *Journal of the American Medical Association*, 278(10):823–832.

Robins, G., Pattison, P., Kalish, Y., and Lusher, D. (2006). An introduction to exponential random graph (p*) models for social networks. *Social Networks*.

Snijders, T. A. and Doreian, P. (2010). Introduction to the special issue on network dynamics. *Social Networks*, 32(1):1 – 3. Dynamics of Social Networks.

Spada, E., Sagliocca, L., Sourdis, J., Garbuglia, A. R., Poggi, V., Fusco, C. D., and Mele, A. (2004). Use of the minimum spanning tree model for molecular epidemiological investigation of a nosocomial outbreak of hepatitis c virus infection. *Journal of Clin Mic*, pages 4230–4236.

Streftaris, G. and Gibson, G. (2004a). Bayesian analysis of experimental epidemics of foot-and-mouth disease. *Proc. R. Soc. Lond. B*, 271:1111–1117.

Streftaris, G. and Gibson, G. (2004b). Bayesian inference for stochastic epidemics in closed populations. *Statistical Modeling*, 4:63–75.

Wallinga, J., Teunis, P., and Kretzschmar, M. (2006). Using data on social contacts of estimate age-specific transmission parameters for respiratory-spread pathogens. *Am J Epidemiol*, 164:936–44.