The Dissertation Committee for Eamon Brendin O'Dea
certifies that this is the approved version of the following dissertation:

Analyses of Infectious Disease Data with Attention to Heterogeneity

Committee:

Claus Wilke, Supervisor

Lauren Ancel Meyers, Co-Supervisor

Mark Kirkpatrick

James Bull

Carlos Carvalho

Analyses of Infectious Disease Data with Attention to Heterogeneity

by

Eamon Brendin O'Dea, B.S.

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

August 2013

To my family

Acknowledgments

I wish to especially thank my advisers Lauren Meyers and Claus Wilke. Their efforts and cooperation provided me with the large number of research assistantships (funded by both the NSF and NIH) and larger number of guiding directions that made this work possible. Next, I must thank Kim Pepin and Ben Lopman for making the analysis of the norovirus data possible and Richard Rothenberg and Stephen Muth for making the analysis of the HIV data possible. These collaborators provided valuable feedback as well as access to data. Finally, I would like to thank all the members of my committee and of my labs. Their encouragement and feedback has had a larger influence than they may realize.

Analyses of Infectious Disease Data with Attention to Heterogeneity

Eamon Brendin O'Dea, Ph.D.
The University of Texas at Austin, 2013

Supervisors: Claus Wilke
Lauren Ancel Meyers

This work comprises three projects that extend previous models to include features of practical significance for the statistical analysis of infectious disease data. In the first, we find from a simulation study how the degree of heterogeneity in the number contacts that individuals have affects the relationship between estimates of a pathogen's effective population size based on coalescent theory and the true prevalence and incidence of that pathogen. In the second, we find that aggregating data from many small outbreaks allows the parameters of stochastic epidemic models to be consistently estimated with a generalized linear model. Application of this method to a set of 77 small norovirus outbreaks reveals interesting differences in the transmission parameters between hospital and nursing-home outbreaks. In the third project, we gain insight into HIV contact networks in the United States by fitting data from a number of surveys to a simple stochastic model of a dynamic network.

Table of Contents

List of Tables

List of Figures

Chapter 1

Contact heterogeneity and phylodynamics: How contact networks shape parasite evolutionary trees

Introduction

Epidemiology is a data-driven field, and it is currently being infused at an increasing rate with molecular sequence data. This new and growing data source has led to a call for multi-level models of the relationship between sequence data and infectious disease dynamics [39, 48], dubbed phylodynamic models.

By allowing for additional data to be used and integrated, phylodynamic modeling may lead to improvements in the accuracy and quality of the surveillance of infectious diseases. For example, the number of norovirus outbreaks reported increased in 2002. It was not clear, however, whether the higher reported numbers were a sign of more outbreaks or more frequent reporting of outbreaks. Case-reporting bias does not affect molecular data, however. So coalescent-analysis of molecular data [93] provided a valuable and largely independent line of evidence that the increase in outbreaks was real. Of course, coalescent analysis will have its own biases, and here we examine those that result from host heterogeneity in contact.

To model heterogeneity in contact, we represent individuals in a population as nodes, and we represent the potential for two hosts to infect each

other as an edge that links two nodes. Researchers call the resulting networks contact networks. Contact-network structure necessarily affects the genealogy of any replicating infectious agent that is spreading through a host population. In this work we use the term parasite to refer to all such infectious agents, including bacteria and viruses. The genealogy of these parasites must fit inside the tree of infections that forms as the parasite spreads from host to host, and this tree of infections must fit inside the host population's contact network. While more elaborate elements of contact-network structure may be important, we here focus simply on variation in the number of edges coming out of nodes, which corresponds to heterogeneity in contact rates.

Contact heterogeneity has often not been discussed as a possible bias in coalescent analyses [e.g., 33, 81, 94]. Researchers performing coalescent analyses have considered contact heterogeneity in a variety of other ways. Hughes et al. [50] linked it to the phylogenetic clustering of sequence isolates. Biek et al. [12] mentioned that it may haved contributed to changes in an estimation of R_0 (the expected number of new cases a single case produces in a susceptible population). Nakano et al. [74] discussed how iatrogenic transmission may have been an important type of transmission in the spread of hepatitis C. Bennett et al. [11] pointed out that population-size estimates from coalescent analyses are really ratios of population size to reproductive variance. But researchers have rarely quantitatively considered how contact heterogeneity might be directly influencing the results of their coalescent analyses. Volz et al. [107] did account for contact heterogeneity in their coalescent model with a saturation parameter, but this application does not provide a general illustration of how contact-network structure can affect genealogies.

Our primary goal here is to assess how contact heterogeneity affects the

accuracy of coalescent-based estimates of population dynamics for epidemic models. First, we build contact networks with different levels of heterogeneity. Then, we simulate the spread of parasites through the networks, generating epidemic dynamics and a genealogy of the parasite with each simulation. Then, we use the BEAST software package [23] to produce Bayesian skyride reconstructions of parasite population dynamics based on the simulated genealogies. By reconstruction, we mean an estimated trajectory of population size. A skyride reconstruction is such a trajectory estimated by adding a smoother to the non-parametric skyline method of estimation [71]. We use the framework of Volz et al. [107] to predict the skyride reconstructions based on the simulated epidemic dynamics. We explain how the contact network structure affects the epidemic dynamics that, in turn, affect the predicted reconstructions. The close agreement between the predicted skyrides and the skyride reconstructions validates this explanation. We also examine how much of the simulated genealogy the skyride reconstruction requires as input in order to produce a reconstruction that agrees with the theoretical prediction.

Materials and Methods

We simulated infectious disease progression on networks. The nodes of the networks represented hosts and had states of susceptible, infectious, or recovered. The edges of the network determined the set of possible transmission events; infectious hosts transmitted infection across edges shared with susceptible hosts until the infectious hosts recovered. The number of nodes in the network was kept at 10,000 and the mean degree (degree is the number of edges coming out of a node) was kept at 4. The networks were built to be either regular, meaning that all nodes have the same degree, or with degree

distributions sampled from Poisson, Exponential, or Pareto distributions. The minimum degree in the Pareto networks was 1. The regular networks served as models with zero heterogeneity, Poisson networks as models with heterogeneity similar to a Poisson process, exponential networks as models with heterogeneity similar to a variety of social networks [5], and Pareto networks (scale-free networks) as models with the extreme levels of heterogeneity that might be found in sexual contact networks [60]. We used the Erdős-Rényi algorithm [26] to generate Poisson networks and an edge-shuffling algorithm [102] to generate the regular, exponential, and Pareto networks.

We simulated epidemics and genealogies in continuous time using a method based on the Stochastic Simulation Algorithm [36, 37]. Epidemics began with one node infectious and the rest of the nodes susceptible. Infectious nodes recovered at a set rate and transmitted infection to susceptible neighbors (nodes sharing an edge) at a set rate. We drew the time to the next event from an exponential distribution with a rate equal to the sum of the rates of all possible events. We then selected an event with probability proportional to its rate, updated the state of the network accordingly, and drew the time until the next event. This process was iterated until either the time-evolution of the epidemic reached a set time point or no more events were possible.

Simulation source code is available from the authors upon request. The code made use of the GNU scientific library [35, version 1.13+dsfg-1] to generate random numbers and the igraph library [21, version 0.5.3-6] to construct networks.

The output of a simulation included a time series of prevalence, that is, the count of infected nodes (given a fixed population of 10,000 nodes), and incidence, that is, the sum of the rates of all possible transmissions. Simulations

also generated infection trees in which each transmission was a bifurcating node, each recovery was a terminal node, and branch lengths were equal to the time between events. We sampled the full infection trees to generate the trees for input in the skyride coalescent analyses. We sampled by selecting a set of nodes uniformly at random from the full infection tree to become tip branches of an infection subtree. To generate the subtree, we cut the branches of the full infection tree at the subset of randomly selected nodes that had no descendants in the set of randomly selected nodes, and we pruned off any paths that did not terminate in this subset of nodes.

Using the sampled infection trees as genealogies, we obtained a posterior distribution for the skyride population sizes with the time-aware method of Minin et al. [71], implemented in BEAST [23, version 1.5.4]. The MCMC chain lengths were 100,000 states and every 10th state was written to a log file. We discarded the first 10,000 states as burn-in. In all cases, effective samples sizes were well above 200. Thus, convergence had occurred. Example BEAST XML input files are available from the authors upon request.

Using the posterior skryride–population-size distributions, we obtained the skryride trajectories with Tracer [80, version 1.5]. Using the framework of Volz et al. [107], we calculated a predicted skyride as described next in the Results.

To plot time series from different stochastic simulations on a common time scale, we used the time at which growth became nearly deterministic in each simulation as time zero for that simulation.

Results

Theory

Coalescent theory is an area of population genetics that models the structure of genealogies backward in time from a set of lineages sampled from a large population. A simple coalescent process turns out to be a good model for the genealogies of a wide range of scenarios in population genetics [54]. In the coalescent process that occurs in the limit of a large population and a much smaller sample, each pair of lineages in the sample coalesces into a common ancestral lineage at a constant rate. When time is measured in units of generations, this rate is equal to the reciprocal of the effective population size. The precise definition of the effective population size depends on the model of the population. In a Wright–Fisher model, the effective population size is equal to the census population size (i.e., the number of individuals in the population). So assuming such a model the rate at which any of the pairs coalesces is equal to the number of pairs of lineages divided by the effective population size.

The skyride uses this simple relationship between effective population size and the expected time before coalescence to estimate population size from the length of intracoalescent intervals in a genealogy. The median of a skyride reconstruction y_{rec} within an intracoalescent interval is approximately

$$y_{\text{rec}} = N_e\tau = \binom{n}{2}u, \tag{1.1}$$

where N_e is the effective population size, τ is the generation time, $\binom{n}{2}$ is the average number of pairs of lineages in the sample within the intracoalescent interval, and u is the length of the intracoalescent interval.

Predicting a skyride from the dynamics of an epidemic model is simply a matter of calculating the rate at which a pair of lineages will coalesce, that is, the rate at which two chains of infection merge into a single chain. Volz et al. [107] have described how coalescence rates follow from prevalence and incidence. Prevalence, given a fixed population size, refers to the count of cases of infection, and so we denote it with I. Incidence refers to the rate at which new cases are occurring, and so we denote it with r_i. The rate of coalescence of a single pair of cases is

$$r_i P, \tag{1.2}$$

where P is the probability that we can trace a particular pair of cases back to a single case before the last transmission event. We have

$$P = 1/\binom{I}{2}, \tag{1.3}$$

making the approximation that the last transmission event was equally likely to have taken place between any pair of current cases. Therefore, the predicted skyride y_{pred} satisfies

$$y_{\text{pred}} = 1/(r_i P) = \binom{I}{2}/r_i. \tag{1.4}$$

If each pair of lineages in our sample were a randomly selected pair from the entire set of pairs in the population, then $P = 1/\binom{I}{2}$ would not be an approximation. Provided that lineages are sampled at random, P is exact up to the first coalescent event. The coalescent event may alter the set of lineages in our sample such that the probability that the next coalescent event occurs between one of the pairs in our sample is different from a randomly selected pair. For the parameters considered in our simulations, the effect of this error was not noticeable.

The similarity of Eqs. (1.4) and (1.1) reflects the similarity of the coalescent process to the transmission process in a continuous-time epidemic model. N_e and τ, however, are often considered as parameters of a discrete-time population model that has non-overlapping generations. The coalescent process describes the genealogy in such a model when we sample a small fraction of the lineages in a population. So how do we interpret N_e and τ in the terms of a continuous-time epidemic model that has overlapping generations? Following Frost and Volz [34] and the general theory of Wakeley and Sargsyan [108], we say that generation time τ is equal to the expected time before an infected individual transmits infection:

$$\tau = I/r_i. \tag{1.5}$$

Then from Eqs. (1.1), (1.4), and $y_{\text{rec}} = y_{\text{pred}}$, we have

$$N_e = (I-1)/2 \approx I/2. \tag{1.6}$$

Simulation

To determine the effect of sampling on the ability of the skyride to reconstruct prevalence history, we simulated genealogies and pruned off a variable number of branches from the genealogies. We found that small amounts of pruning rapidly reduced the number of coalescent events in the sampled genealogy that occurred in the peak and late phases of the epidemic, thereby restricting accurate reconstruction to the early phase of the epidemic (Figure 1.1).

To demonstrate the effect of network structure on the reconstruction of prevalence history, epidemics were simulated on networks with varying heterogeneity. Keeping the extent of sampling equal and increasing heterogeneity

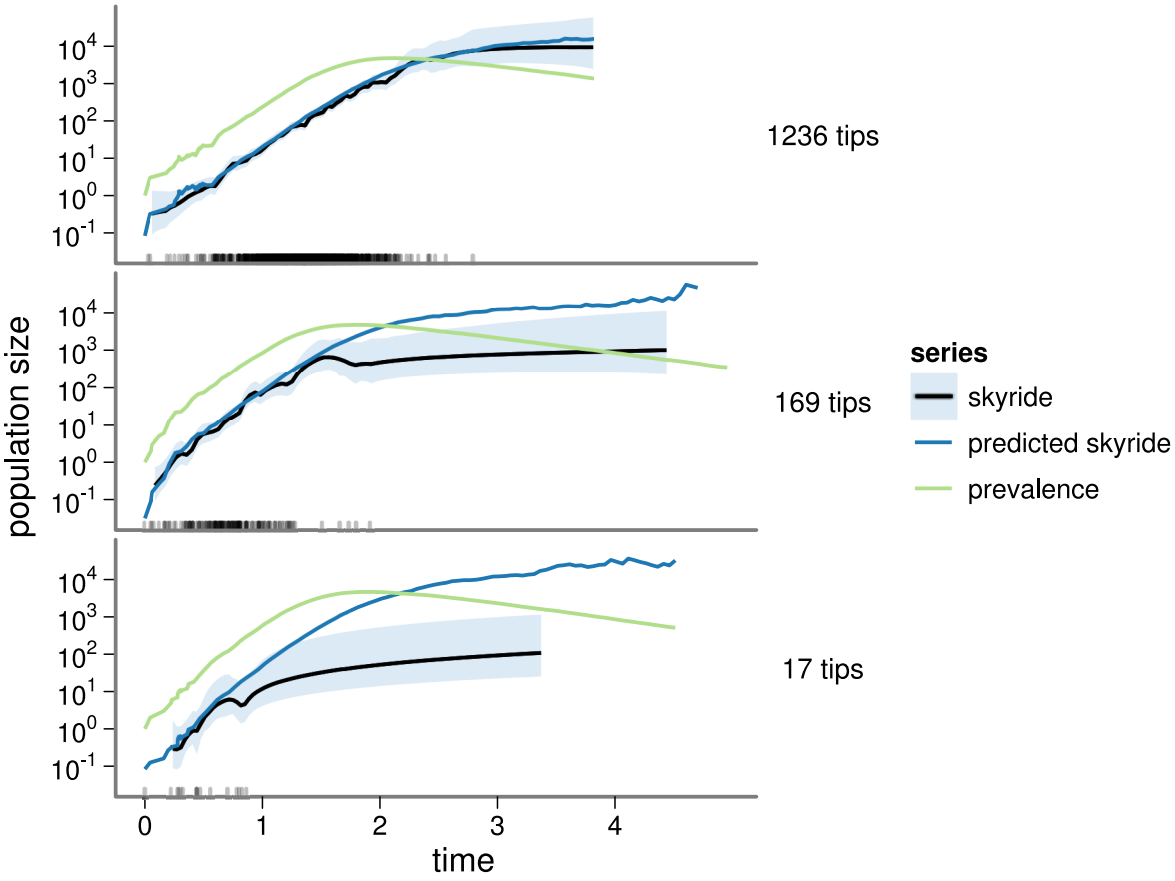

Figure 1.1: Low levels of proportional sampling may prevent accurate reconstruction of prevalence during and after the epidemic peak. We consider reconstruction accurate when the skyride and the predicted skyride match. The light-blue ribbons are the middle 95% of the posterior density of the skyride reconstruction. The small bars on the x-axis represents the times of coalescent events in the sampled genealogy. Panel labels on the right indicate the number of tips in the sampled genealogy. Parameters: contact-network size = 10,000, Poisson degree distribution with mean = 4, transmission rate = 2, recovery rate = 1, proportion of nodes sampled = $\{0.1, 0.01, 0.001\}$ (top, middle, and bottom panels).

compressed the coalescent events in the sampled genealogy into the beginning of the epidemic. Figure 1.2 shows a representative example of this general trend that holds across intermediate levels of sampling. Consequently, increasing heterogeneity has a similar effect to reducing the proportion of nodes sampled: the time at which the prediction of the skyride based on prevalence and incidence diverges from the estimated skyride based on the genealogy occurs earlier.

Figure 1.3 shows how differences in the scaling of prevalence to the skyride follows from differences in trajectories of prevalence and incidence. The ratio of prevalence to incidence is the expected time until an infected host transmits infection, and we here define it as the generation time (1.5). In Figure 1.3, we see that generation times are at, or quickly reach, a minimum after an epidemic begins and then gradually increase until the epidemic ends. In the regular networks, the decline in the number of susceptible hosts over the course of the epidemic causes this increase to happen. In the other networks, which have hosts of varying degree, infection first moves to the high-degree hosts and then to progressively lower- and lower-degree hosts [6, 7, 103]. Because the degree of a host determines how much his/her infection increases incidence, this movement of infection from high- to low-degree hosts translates into generation times being at first shorter and then longer in heterogeneous networks relative to regular networks (Figure 1.3).

Discussion

The effects of contact heterogeneity can be important in relating the structure of genealogies to infectious disease dynamics (Figure 1.3). The strength of the effect will vary from system to system, and for some systems other aspects of contact-network structure such as the frequency of short paths [70]

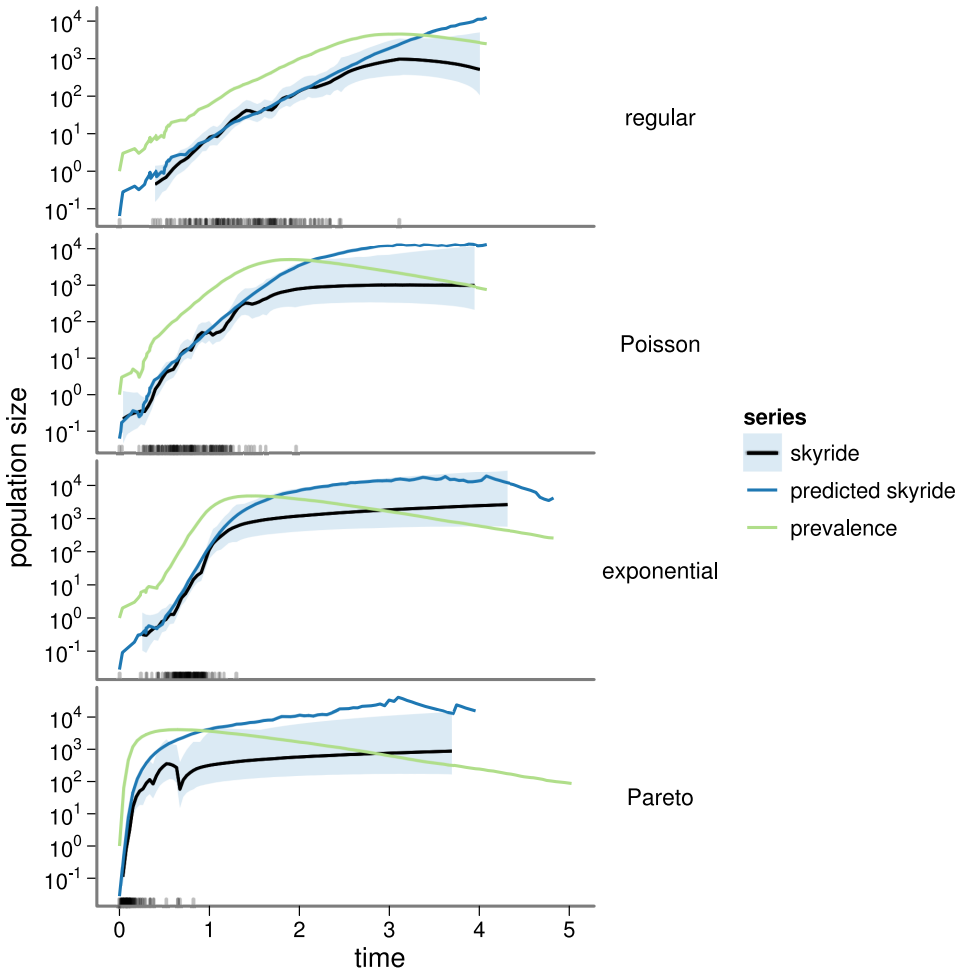

Figure 1.2: Contact heterogeneity determines the amount of time over which the skyride estimated from the geneaology is informative of the skyride predicted by prevalence and incidence. Contact heterogeneity also affects the relationship between the skyride and prevalence trajectories. The light-blue ribbons are the middle 95% of the posterior density of the skyride reconstruction. The small bars on the x-axis represent the times of coalescent events in the sampled genealogy. Panels labels on the right indicate the approximate degree distribution of the contact networks. The variance of the degree distributions increases from the top panel to the bottom panel. Parameters: contact-network size = 10,000, degree distribution mean = 4, transmission rate = 2, recovery rate = 1, proportion of nodes sampled = 0.01.

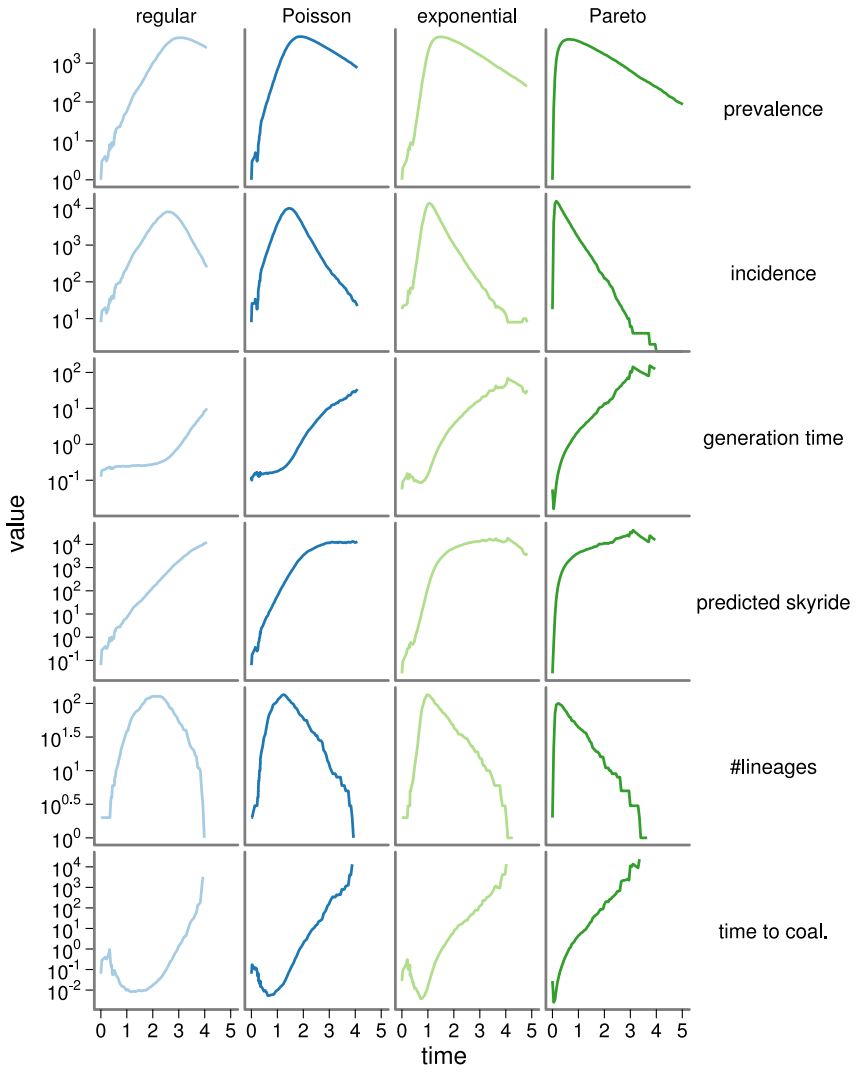

Figure 1.3: Contact-network structure, infectious disease dynamics, and genealogical structure interact. The panels of this figure plot several related variables from our simulations. The ratio of prevalence to incidence is the generation time, which scales prevalence to the predicted skyride (up to a constant factor). Dividing the predicted skyride by the number of pairs of lineages backs out a smoothed expected length of intracoalescent intervals in the genealogy. Panels labels on the top indicate the approximate degree distribution of the contact networks. The variance of the degree distributions increase from left to right. Parameters: contact-network size = 10,000, degree distribution mean = 4, transmission rate = 2, recovery rate = 1, proportion of nodes sampled = 0.01.

and the dynamics of edge formation [1, 72, 105, 106] may also be important. More generally, models may also require more detailed models of the course of infection within hosts (including incubation periods, for example), the effects of natural selection [76, 110], and other additions before they can make precise predictions in real-world systems.

But are the data requirements of these more complex models feasible? To begin answering this question, we next discuss the implications of obtaining the equivalent of our simulated data from a real-world system.

We knew the true infection tree in our simulations. In typical coalescent analyses of an infectious disease [e.g., 24, 71] we do not know the true genealogy and so we must infer it along with the dynamics of the effective population size. Although there is a large set of methods for the inference of trees from sequences [30, 59, 115], the variety of methods available reflects the difficulty of the task. Additionally, as is well known by practitioners of phylogenetics, substitution rates set fundamental limits on the amount of phylogenetic information that sequences may contain. Sequences with common ancestors that are very recent may not have any polymorphic sites that could suggest the structure of the branching of the tree connecting them. Sequences with common ancestors that are too distant similarly contain little information about the true genealogy [40].

It may be possible to work around the second problem by collecting sequences over time such that there are no branching points in the tree that are too far from every pair of tips. For the first problem, there is simply no information that the sequences alone can provide and additional knowledge of events in the chain of infection is necessary to determine the infection tree. The panels labeled "time to coal." in Figure 3 show that this additional infor-

mation is most likely to be needed early in the epidemic and when there is a large amount of variance in the contact network. It is then perhaps fortunate that contact-tracing methods are practiced by many health departments for sexually transmitted diseases (STDs) [67, 97], which are thought to have have higher contact heterogeneity than airborne diseases [60]. However, we probably need more widespread practice of contact tracing for large genealogies to be assembled. A recent survey of physicians in the U.S. [95] found that less than one-third of physicians routinely screen patients for STDs and many physicians relied on patients to notify health departments and partners, and similar surveys in other countries [19, 43, 67] likewise indicate that contact tracing is not routine in general medical care of STDs.

There also may be a need for contact tracing to establish the genealogy for airborne infections because many airborne transmissions may occur in a single day during which a single strain may be dominant in a host, as the super-spreading events in the 2003 SARS-coronavirus outbreak demonstrated [87]. Contact-tracing is also practiced for airborne diseases. It has been used to help contain the SARS-coronavirus outbreak [22], smallpox [31], and tuberculosis [86]. Given that contacts for airborne disease can be quite transient, it seems that, even with the addition of contact-tracing data, we may generally know less about parasite genealogies for airborne diseases compared to STDs. On the upside, our results suggest that the ability to reconstruct early parts of the epidemic is robust to much pruning of the full genealogy (Figure 1.1). However, this robustness may depend on our sampling scheme. Using discrete-time simulations, Stack et al. [96] found that the difference between reconstructed prevalence and simulated prevalence depended largely on how the samples were distributed over the course of the epidemic. Also, it is unclear

how any of our sampling levels might compare to realistic amounts of contact tracing and molecular data for a specific infectious disease.

In addition to being necessary to fill gaps in molecular data, contact-tracing may be be necessary because genealogies do not always match infection trees. Such discordance is likely to occur when there is relatively little time between transmissions. When there is little time for a mutant to become fixed between transmissions, the order in which alleles at loci of a sequence appear in transmitting inocula (or sequence isolates) need not match the order in which the alleles appeared in the within-host population. Measures of within-host viral load and sequence diversity may be informative of the chance of such discordance. If populations tend to be large and diverse, then sequence data may be useless for reconstructing the recent details of chains of infection but still useful in reconstructing deeper branches in the tree. Sequence data from diverse within-host populations could also be useful in parameter estimation for coalescent models [e.g., 25] that include the within-host dynamics of the parasite. Two properties that parasites may have that would help increase the chance that infection trees and genealogies match are a low level of diversity in tranmsitting inocula (i.e., a strong bottleneck effect at transmission) and reduction of diversity in an incubation period that precedes all transmission.

In our simulations, we also knew the variance of the degree distribution. We do have some data about the structure of contact networks for some systems. We have survey data about human sexual–contact-networks [e.g., 78, 85] and survey data about networks of close, but not sexual, human contacts [45, 73, 109]. Researchers have used field data to construct hypothetical contact networks for wildlife and vector-borne diseases [e.g., 20, 88], and researchers have also used census data to construct hypothetical contact net-

works for human diseases [e.g., 27, 69]. It seems likely, however, that in the analysis of real sequence data the heterogeneity of the contact network will be at least as uncertain as disease incidence and prevalence. Thus, estimation of contact heterogeneity may be an important goal of the analysis. We note that previous work [e.g., 89] has also discussed the potential use of sequence data to estimate contact heterogeneity.

Conclusions

Contact heterogeneity is well-known to have a strong effect on infectious disease dynamics. We have shown how the relationship between infectious disease dynamics and genealogies is similarly sensitive to the contact heterogeneity specified by a network. We have argued that direct knowledge of the tree of infections is likely needed in addition to sequence data for the accurate inference of prevalence from sequence data. Thus, it seems that understanding the structure of the contact networks for various diseases will be important for progress in phylodynamics.

Chapter 2

Joint estimation of transmission rate and initial growth rate with data aggregated from multiple norovirus outbreaks

Introduction

A common and difficult problem in epidemiology is to estimate rates of disease spread. Accurate estimates of these and other population parameters are crucial in the evaluation of disease control measures [2, 41, 53] or biological hypotheses [61]. Heterogeneity complicates the problem of obtaining such estimates. For example, a person's risk of infection depends on contact rates and acquired immunity, and these quantities can vary widely between people and outbreaks.

Norovirus (NoV) epidemiology provides a fine case in point of the need for models to accommodate heterogeneity. Noroviruses are the most common cause of diarrheal disease in the United States, causing an estimated 21 million cases [90] and 71,000 hospitalizations per year [63]. A genetically diverse group of strains is often circulating within a population. New strains of the predominant genogroup 2 genotype 4 (GII.4) taxon appear regularly over time [38], and a person's risk of infection, given exposure, likely depends on both the antigenicity of the virus and the type-specific immunity developed from the person's previous exposure [17]. Other important heterogeneities include innate susceptibility (which depends on a person's histo-blood group antigens

and secretor status) and age-specific risks of exposure. Outbreak investigations [28, 99, 112] have provided convincing evidence that single vomiting incidents in crowded settings can lead to scores of secondary cases. Models that account for both between-individual and between-population heterogeneity are needed to obtain the accurate parameter estimates required for predicting outbreak dynamics and implementing effective controls. At present, control measures are based on general infection-control principles [18] and thus are likely to be somewhat inefficient.

A further complication for modeling norovirus is that it often occurs in small outbreaks. The transmission and recovery times of cases in small outbreaks are correlated [82], which makes estimation difficult when using data from a single outbreak. Some previous work has developed methods for estimating parameters with data from multiple small outbreaks in different households [8, 10]. The approach we take here differs from that work in using maximum likelihood, which is known to be optimal for large data sets; in being based on full observation of the outbreaks rather than final sizes; and in modeling variation in the parameters via a linear predictor. Our regression approach, however, makes this work similar in spirit to Höhle [46].

We account for a simple but fundamental type of heterogeneity between outbreaks—variation in the initial growth rate of the outbreak. In our model, the initial growth depends on the number of initially susceptible individuals. In the case of norovirus, this number is difficult to know as there is no serological correlate of protection. The incidence rate of new cases in our model is proportional to the product of the number of susceptibles and the transmission rate, so estimates of the transmission rate will be highly sensitive to those of the number of susceptibles in the model. Recent work on joint estimation

of transmission rates and the initial number of susceptibles with data from a single outbreak [42, 49, 56, 57] has shown that estimates of the initial number of susceptibles tend to be low when data sets are small. We see the same bias when the estimates are based on multiple outbreaks, but obtain accurate estimates in the limit of a large number of outbreaks even if all outbreaks are small.

Fitting our model to a large number of outbreaks, we find a distinct increase in transmission and initial growth rates in long-term–care facilities relative to hospitals. A simulation experiment shows that our methods perform well even when some of the data are missing.

Methods

Model

At the beginning of an outbreak, $t = 0$, a population is made up of $Y(0)$ *infective* people and $X_i(0)$ *susceptible* people of type i for one or more types. A type of person in this model is defined by the instantaneous rate at which infectives may infect her, which we call the *transmission rate,* and by the mean and dispersion parameter of her gamma-distributed infectious period, should she become infected. As time moves forward, each infective person transmits infection to each susceptible person at the points of a Poisson process such that the rate at which new type-i infections appear is $\beta_i X_i(t)Y(t)$, where β_i is the transmission rate for type i. A susceptible person that contracts infection becomes infective after being in a *latent* state for a fixed time period. She then becomes a *recovered* person after a randomly-distributed infectious period. No other transitions in state occur.

Our outbreak model departs from the stochastic counterpart of the

common Kermack–McKendrick susceptible–infective–recovered (SIR) model in two ways besides the addition of a latent state and multiple types. First, we do not make our transmission rate depend on the total number of people N in the population. This departure prevents the need for N to be estimated, and it is appropriate when an infective person may be able to infect every susceptible person in the population with approximately the same probability. Second, we do not assume that latent periods and infectious periods are exponentially distributed, which is more realistic because it allows the probability of a person leaving a latent or infectious state to depend on how long she has been in that state.

It is natural to introduce the likelihood of the data for each type of person separately because each type of person is defined by different parameters. As indicated in our outbreak model description, the rate at which an infective transmits to a susceptible depends on the susceptible's type. The type of a person also determines the parameters of her symptomatic period. With multiple-outbreak data, we further define types as unique to individual outbreaks. In other words, we make no general assumption that people in different outbreaks may be modeled with the same parameters.

It is also natural to introduce the recovery-time and transmission-time parts of the likelihoods for each type of person separately because these parts factor apart into common density functions. The simplicity of these functions belies an involved construction, available in [52], as the product integral of the likelihood of events in infinitesimal time steps, where the likelihood of each time step is conditional on the history of the model up until that time step.

For type i people, the recovery-time part of the likelihood is

$$l_{\text{rec}}(\mu_i, \rho) = \prod_{j=1}^{k_i} \frac{1}{\Gamma(1/\rho)(\rho\mu_i)^{1/\rho}} I_{i,j}^{1/\rho-1} \exp \frac{-I_{i,j}}{\rho\mu_i}, \tag{2.1}$$

where k_i is the number of type-i people infected over the course of an outbreak, $I_{i,j}$ denotes the length of the symptomatic period of the jth type-i infection, μ_i is the mean of the symptomatic period of type-i infections, and ρ is the dispersion parameter, which we take to be the same for all types of infections. Equation (2.1) represents the standard likelihood function for a joint distribution of gamma-distributed random variables. Recall that per our model definition, the symptomatic periods $I_{i,j}$ are gamma distributed within each type.

The transmission-time part of the likelihood for type-i people is

$$\begin{aligned} l_{\text{tr}}(\beta_i, X_i^{(0)}) = {} & X_i^{(0)}!/(X_i^{(0)} - k_i)! \exp[-\beta_i \tau_i (X_i^{(0)} - k_i)] \\ & \times \prod_{j=1}^{k_i} \beta_i Y_{i\,j} \exp(-\beta_i h_{i,j}), \end{aligned} \tag{2.2}$$

where τ_i is the cumulative exposure of such people at the end of an outbreak (i.e., the total area under $Y(t)$), $Y_{i,j}$ is the number of infectives present when the jth such person becomes infected, $h_{i,j}$ is the cumulative exposure of the jth such person when infected. Further discussion of this likelihood function is provided in the Appendix.

Examination of (2.2) reveals a few ways in which the data requirements for estimation may be minimized. The values $Y_{i,j}$ disappear on differentiation of the log likelihood, which means that they do not need to be known to find maximum likelihood estimates or Hessian-based (Wald) confidence intervals. Additionally, the $h_{i,j}$ only affect the likelihood through the sum $\sum_j h_{i,j}$. Thus

some error in our calculation of $h_{i,j}$ should not bias our estimates too much as long as the average error is close to zero.

The likelihood (2.2) can be parameterized differently as

$$l_{\mathrm{tr}}(\beta_i, r_i) = (r_i/\beta_i)!/(r_i/\beta_i - k_i)! \exp[\tau_i(\beta_i k_i - r_i)] \times \prod_{j=1}^{k_i} \beta_i Y_{i\,j} \exp(-\beta_i h_{i,j}), \tag{2.3}$$

where $r_i = \beta_i X_i^{(0)}$ is the initial per-infective incidence rate. In our application, we choose to estimate r_i instead of $X_i^{(0)}$ because r_i is easier to interpret in the context of our data. For brevity, we refer to r_i as the initial growth rate.

The full likelihood function that we use for an n-outbreak data set is then

$$l(\boldsymbol{\beta}, \mathbf{r}, \boldsymbol{\mu}, \rho) = \prod_i l_{\mathrm{tr}}(\beta_i, r_i) l_{\mathrm{rec}}(\mu_i, \rho), \tag{2.4}$$

where we use boldface to denote vectors with elements equal to the parameters for each type i.

To make use of previous results from statistical theory as well as to use conventional language when writing about our model, we shall next present our model as a generalized linear model (GLM). GLMs are a broad class of statistical models that includes many commonly used regression models. A GLM consists of three components: (i) a density function from the exponential family, (ii) a linear model that maps predictive variables to a predictor, and (iii) a link function that maps the predictor to the mean of the density function.

Our likelihood functions, (2.1) and (2.3), fit the definition of exponential family densities. That is not to say that the transmission and recovery times from a small outbreak are independent random variables with those

densities. In fact, they may be highly correlated [82]. In the limit of a large number of outbreaks, however, we may be assured of sufficient independence for asymptotic results to apply [52].

We obtain a linear model by associating each infective type with a set of predictive variables. In the application to norovirus we describe here, such predictive variables are for example facility type in which an outbreak occurred (hospital or long-term–care facility) or case type (patient vs. facility staff). We combine these predictive variables into a design matrix $\mathbf{Z}$, which has a row for each type i and a column for each predictive variable. The linear mapping from multiple predictive variables to a linear predictor is achieved by multiplying the design matrix with a vector of regression parameters $\mathbf{c}$.

As link function, we chose the natural log, which tended to perform better than other potential link functions in our application. For example, for transmission-rate estimates β_i, we let $\log \beta_i = \mathbf{Z}_{i,*}\mathbf{c}_\beta$, where $\mathbf{Z}_{i,*}$ is row i of the design matrix and $\mathbf{c}_\beta$ are our regression parameters for the transmission rates.

Fahrmeir [29] gives conditions for consistency and asymptotic normality of parameter estimates for GLMs. In the case of our model, asymptotic normality will not occur for the transmission and growth rate parameters when the data include an outbreak in which everyone was infected, because in that case the true value of the parameter lies on the boundary of parameter space. On the one hand, the probability of such a large outbreak occurring in the model will be extremely small for many parameter values. On the other hand, some of the estimates in our application are on the boundary of feasible parameter space. For this reason, we assessed confidence in our estimates by parametric bootstrap rather than by relying on asymptotic normality.

We are not aware of previous work giving conditions for consistency

that is directly applicable to our model. In the Appendix, we provide a proof of consistency for our model in the simple case that all outbreaks share the same parameters. Evidence that the model performs well in realistic situations appears in the Results section. We are able to recover from simulated data the parameters for the non-trivial model that we fit in our application.

We fit our model to outbreak data by maximizing (2.4) given the data, using the Newton–Raphson method as implemented by AD Model Builder [32]. To keep the Newton–Raphson search for maximum likelihood estimates in the feasible parameter space, we add a penalty to the log likelihood whenever the implied final number of susceptibles $x = X_i^{(0)} - k_i$ for an outbreak is too close to zero, $x < \epsilon$. The penalty is of the form $C(x - \epsilon)^2$, where C is an arbitrary numeric constant which we set to $C = 0.01$. Likewise, whenever $x < \epsilon$, we replace x by $\epsilon/(2 - x/\epsilon)$. Throughout this work, we use $\epsilon = 0.001$.

Data

The norovirus (NoV) data we analyze here originated in a prospective surveillance program in hospitals and long–term-care facilities in England [64, 65]. We analyzed the dynamics of 77 outbreaks laboratory-confirmed to be caused by NoV in which a total of 1568 cases of gastroenteritis occurred among patients and staff. We selected these data from the larger data set produced by the surveillance program as follows.

In the original data, one outbreak contained three reported symptomatic periods that had a strong influence on the model fit and appeared likely to be data-entry errors. One symptomatic period began and ended several months before the others, one was 34 days long, and the other was 120 days long. Both of these long symptomatic periods were for staff members,

which makes it unlikely that these protracted periods were real effects of the frailty of the patients. Rather, it appears that the day and the month were reversed in two cases and a zero was entered as a one in another case. We therefore corrected these putative data-entry errors before analyzing the data. Two other data-entry errors were noticed in which the date of relief from symptoms was later than the date of onset of symptoms. We discarded these two records.

Most records of infections that were attributed in whole or in part to norovirus included the dates of both the onset of and the relief from symptoms. However, in many records both dates were missing, and in most outbreaks some records lacked at least one date.

We discarded all records from outbreaks in which more than 55% of the dates of relief were missing. In the remaining outbreaks, we replaced missing dates of relief with the corresponding onset date plus the median symptomatic period from complete records in that outbreak. These replacements were done as a preparation for the estimation of the transmission rates and were not included when estimating symptomatic periods.

We discarded all records where the onset date was missing. This practice is unlikely to introduce a large bias as long as a relatively small number of onset dates are discarded. So data from outbreaks where the number of discarded records would have exceeded 7% of the number of retained records are excluded from analysis.

The thresholds of 55% and 7% were chosen because they allowed the bulk of the data to be included and not much more data could be included without drastically increasing these thresholds.

We made several simplifying assumptions. We assumed a person is infective only when symptomatic. We further assumed that staff are infective

only on day one of a symptomatic period, in accordance with an infection control policy. We also assumed that the latent period is fixed at 24 hours. We assumed a small, background hazard of infection (10^{-8} that of an infective) triggered illness in cases when no infectives were present. We also assumed that the number of initial infective people was equal to the number of people reporting symptoms on the first day of the outbreak. Finally, we assumed that any changes in state happen at the same time each day.

Predictive variables

The predictive variables that determined our design matrices were as follows. The data were collected over the course of a 1-year period beginning in April 2002, and we categorized the data into two groups by the period in which they began: spring–summer refers to outbreaks that started between April 1 and October 1 of the study year; fall–winter refers to outbreaks that began in the remainder of the study year. The period variable allows for variation in transmission rate as a result of seasonality of NoV.

As an additional predictive variable, we include what type of facility the outbreak occurred in, hospital or long-term–care facility (LTCF).

The third predictive variable we use is size class. We classify units in which the number of beds is less than or equal to the median number of beds as small. We classify the other units as large. This classification was done separately for hospital and LTCF units because LTCF units are usually larger than hospital units. For the hospitals, the small units have 6–22 beds and the large units have 24–33 beds. For the LTCF units, the small units have 6–34 beds and the large units have 36–66 beds. The size class variable allows the number of initial susceptibles to depend on the approximate total number of

people in each unit. The variable also allows population sizes to affect contact rates.

The fourth predictive variable we use is case type, the two types being patient and staff. Case type is the only predictive variable that varied within outbreaks.

We use a facility–size–period–case-type combination with a relatively large amount of data as the reference group. Specifically, the reference group comprises outbreaks that occurred among patients in large care-units of hospitals that began between October 2002 and April 2003. The estimated rate parameter for the reference group serves as the coefficient of the intercept of the linear model. Estimates for other coefficients then inform us of how moving away from the reference group changes rate estimates.

Confidence intervals

To obtain confidence intervals for the estimates, we performed a parametric bootstrap. Data were simulated according to our outbreak model with the estimated parameters. Each simulation produced data from a set of outbreaks equal in size to the set that we fit, with each outbreak in the simulation matching an outbreak in the fitted data in terms of initial number of infectives, predictive variables, fraction of case records with missing onset and recovery times, and fraction of cases with missing recovery times. Percentile confidence intervals for regression coefficients were estimated from 10,000 simulation replicates.

Diagnostics

As a general test of model fit for the transmission rate and growth rate likelihoods, we calculated the percentile of the log likelihood of the fit to the real data in the distribution of log likelihoods generated by bootstrapping. Out of 10,000 bootstrap replicates, our optimization code found estimates in 9809 cases. The log likelihood of the fit to the real data was in the 25^{th} percentile of the log likelihoods from these estimates. Thus, the log-likelihood of our fit to the real data is not extreme, consistent with a good model fit.

Our use of the moments estimator for the dispersion parameter ρ in (2.1) precluded a similar assessment of model fit for the symptomatic periods. However, inspection of the default diagnostic plots for glm objects in R did not indicate any problems.

Simulation

We used simulation to investigate how the bias and variance of our estimates depend on the number of outbreaks that they are based on as well as the amount of missing information. We also used simulation to generate bootstrap confidence intervals.

Simulations began with some initial number $X_i^{(0)}$ of susceptibles of type i. To initiate the outbreak, some of the susceptibles were moved into a latent state. All people entering the latent state moved to the infective state after a fixed time period. People entering the infective state moved on to the recovered state after a gamma-distributed time period with mean μ_i and dispersion parameter ρ. Every time the number of infectives or susceptibles changed, the time of a potential transmission event was calculated by drawing from an exponential distribution with rate $Y \sum_i \beta_i X_i$, where Y is the number of infec-

tives and β_i is the transmission rate for susceptibles of type i. If the potential transmission was sooner than the next change in Y, a type of susceptible was chosen with probability proportional to $\beta_i X_i$ and moved into the latent state. Simulations stopped when the number of latent and infective people reached zero. The output of the simulations was a record for each person infected giving transition times.

Our simulation experiment had a full factorial design, with the number of outbreaks n being 1, 10, or 100; the fraction of recovery times imputed being either zero or approximately the highest such fraction in our real data (0.53); the fraction of records missing both onset and recovery times being either zero or approximately the highest such fraction in our real data (0.05); and onset and recovery times being either rounded to days or exact. For each combination of factor levels, we simulated data and attempted to fit it 10,000 times. Simulations were initiated with one infective and had a transmission rate β of 0.0037, an initial growth rate r of 0.2664, a latent period of 24 hours, and infectious periods with a mean μ of 3.32 days and a dispersion parameter ρ of 0.58.

Once-daily observation of the outbreak was simulated by rounding transition times down to the nearest whole day. Outbreaks were started at random times in the first day to prevent the rounding from having artificial effects on the data from small outbreaks.

Our gradient-based optimization code for model fitting, which worked well at estimating transmission rate parameters when the number of outbreaks was large, did not work well when the number of outbreaks was small. So we used specialized code to fit the models of the simulation study, which were more analytically tractable by virtue of not having linear predictors. The Appendix

describes the basis for this code, which always finds the maximum likelihood estimate if it exists and identifies cases in which no such estimate exists.

Software

Our outbreak simulation code made use of the SimPy [100] python module. The RngStreams C library [58] allowed for the simulations to run in parallel. We used the AD Model Builder [32] and R2admb [13], an R [79] interface for it, to optimize the log likelihood. We prepared graphics with the R package ggplot2 [111]. Code capable of reproducing the results is available from the authors on request.

Results

Simulation

We used simulation to see how many outbreak data sets may be required for estimates to be approximately normally distributed around the true parameter values. The simulations also allowed us to gauge the effects of the imputation and rounding necessary for our application.

Much previous work has shown that estimation with data from a single, small outbreak is unreliable [92, and refs. therein]. Thus one benefit of aggregating data from multiple outbreaks is that it allows for data from minor outbreaks to produce reliable estimates. However, using data from minor outbreaks does represent a worst-case scenario in the sense that each such outbreak contributes only a small amount of information. For those two reasons, and to keep the simulation study at a manageable size, we restricted our simulations to one set of parameters that is guaranteed to result in small outbreaks. To allow for comparison with our fits to the norovirus data, we used

the parameters estimated for our baseline regression group.

As expected, the estimates were not very good when using data from single outbreaks (Fig. 2.1). In about 49% of these simulations, the initial infective failed to infect anyone, limiting estimation to the length of the symptomatic period. In about 13% of these simulations, only one transmission occurred and the transmission and growth rate parameters were unidentifiable. In about 21% of these outbreaks, the estimate of r was on the lower bound of parameter space, preventing calculation of Wald confidence intervals. In the remaining 17% of replicates, the coverage probability of the 95% Wald confidence intervals ranged from 80 to 90% (Table 2.1) and the bias and average standard error for the transmission rate was almost 100 times the true value of the parameter. The average correlation between the transmission rate and initial growth rate estimates was 94%. Estimates for the symptomatic period, although obtained for all replicates, were also not accurate (Fig. 2.1 and Table 2.2).

Rounding, deleting 5% of case records, and imputing 53% of recovery times all generally increased the average standard error of estimates, with effects in that order. Effects on the bias were somewhat more variable, but the asymptotic effects of these procedures on the bias of the estimates appears to be zero. However, even in the 100-outbreak scenario the imputation caused coverage probabilities for r to deviate by as many as 13 percentage points from 95%. (Table 2.1), which recommends the use of confidence intervals that account for the imputation, such as the ones we used in our application.

On the whole, the estimates were much more accurate in the 10- and 100-outbreak scenarios (Fig. 2.1, Tables 2.1 and 2.2). They were also more robust. Estimates for r were on the lower bound 5% of the time in the 10-

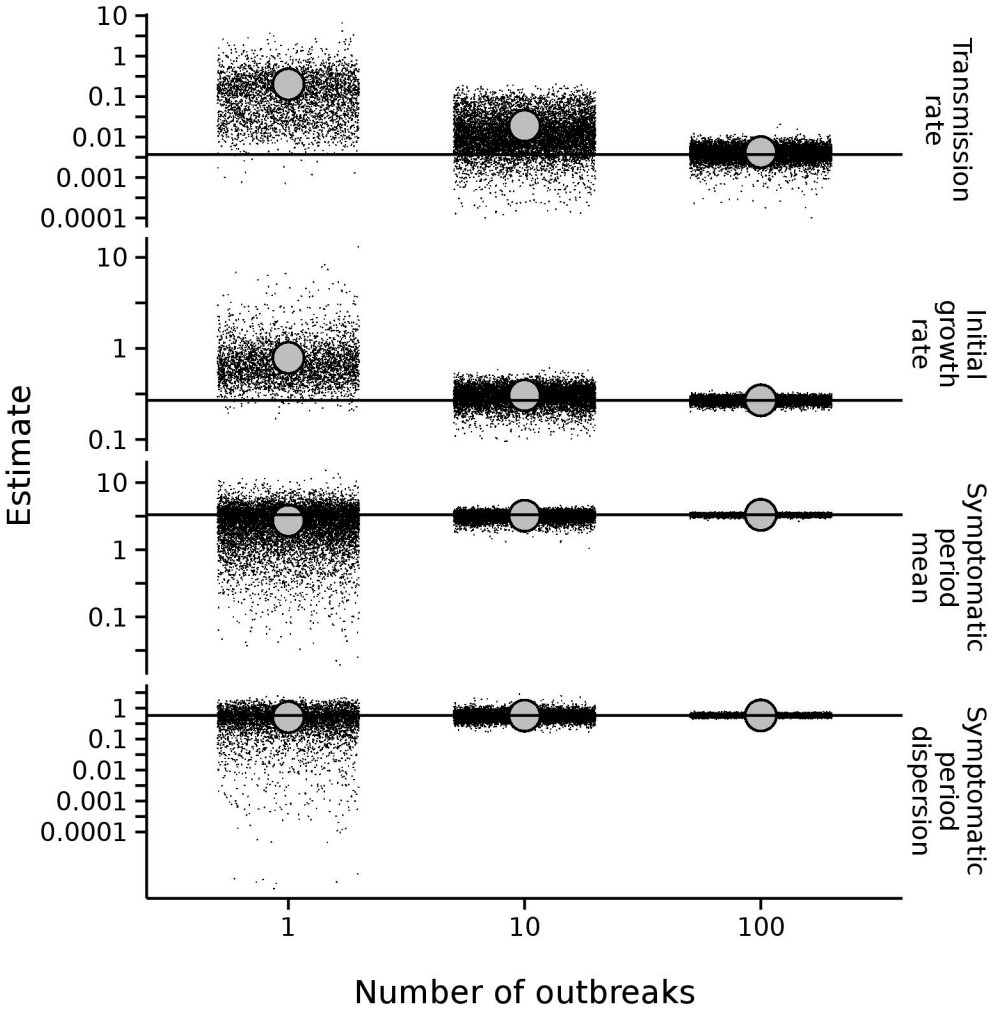

Figure 2.1: Estimates versus number of outbreaks. The row names indicate parameters. Each small black point represents an estimate. The larger gray points represent the means of the estimates. The horizontal lines represent the values of the parameters used to simulate the data.

Table 2.1: Simulation results for transmission rate β and initial growth rate r. n denotes the number of outbreaks simulated for an estimate. Imputed refers to the fraction of recovery times deleted and then imputed as described in the Methods. Missing refers to the fraction of case records deleted before fitting the data. Rounded indicates whether the onset and recovery times were rounded to whole days. In the simulations, β was set to 0.0037 transmissions per infective–susceptible day and the r was set to 0.2664 transmissions per infective day.

n	imputed	missing	rounded	bias($\hat{\beta}$)	av. s.e.($\hat{\beta}$)	β cover. (%)	bias($\hat{r}$)	av. s.e.($\hat{r}$)	r cover. (%)
1	0.00	0.00	0	0.198	0.131	82	0.52	0.834	87
			1	0.196	0.129	82	0.52	0.830	88
		0.05	0	0.21	0.130	81	0.55	0.84	87
			1	0.206	0.127	83	0.53	0.83	90
	0.53	0.00	0	0.251	0.132	80	0.67	0.958	81
			1	0.234	0.137	80	0.65	0.966	81
		0.05	0	0.231	0.149	83	0.63	0.98	84
			1	0.214	0.134	81	0.61	0.96	84
10	0.00	0.00	0	0.0154	0.0456	89	0.0389	0.2882	95
			1	0.0155	0.0457	90	0.0378	0.2888	95
		0.05	0	0.0160	0.0471	90	0.0370	0.2932	96
			1	0.0160	0.0470	89	0.0373	0.2919	96
	0.53	0.00	0	0.0158	0.0525	93	0.0602	0.3116	91
			1	0.0162	0.0533	93	0.0608	0.3134	91
		0.05	0	0.0168	0.0558	92	0.0583	0.319	92
			1	0.0170	0.0562	93	0.0587	0.319	92
100	0.00	0.00	0	0.00055	0.00571	94	0.0027	0.07380	95
			1	0.00053	0.00569	94	0.0025	0.07370	95
		0.05	0	0.00066	0.00614	94	-0.0001	0.07533	95
			1	0.00067	0.00617	93	-0.0004	0.07535	95
	0.53	0.00	0	-0.00033	0.00668	96	0.0195	0.08036	82
			1	-0.00034	0.00653	96	0.0131	0.07844	86
		0.05	0	-0.00018	0.00724	97	0.0174	0.08222	85
			1	-0.00024	0.00707	97	0.0103	0.08018	87

Table 2.2: Simulation results for symptomatic period mean μ and dispersion parameter ρ. n denotes the number of outbreaks simulated for an estimate. Missing refers to the fraction of case records deleted before fitting the data. Rounded indicates whether the onset and recovery times were rounded to whole days. Cover. refers to the coverage probability of Wald confidence intervals. Lower $\hat{\rho}$ and upper $\hat{\rho}$ refer to the bounds of a bootstrap confidence interval. In the simulations, μ was set to 3.32 days and the ρ was set to 0.58.

n	missing	rounded	bias($\hat{\mu}$)	av. s.e.($\hat{\mu}$)	cover. (%)	bias($\hat{\rho}$)	lower $\hat{\rho}$	upper $\hat{\rho}$
1	0.00	0	-0.60	3.76	83	-0.067	0.01	1.46
		1	-0.59	3.75	84	-0.037	0.00	2.00
	0.53	0	-0.65	4.13	80	-0.096	0.01	1.47
		1	-0.59	4.26	80	-0.027	0.00	2.00
10	0.00	0	-0.104	1.511	91	-0.005	0.31	0.98
		1	-0.106	1.534	91	0.008	0.32	0.98
	0.53	0	-0.108	2.123	90	-0.016	0.23	1.12
		1	-0.106	2.161	90	0.002	0.24	1.15
100	0.00	0	-0.010	0.4703	94	-0.0006	0.49	0.68
		1	-0.010	0.4772	95	0.0157	0.50	0.70
	0.53	0	-0.007	0.6853	94	-0.0012	0.46	0.73
		1	-0.010	0.6928	94	0.0132	0.47	0.76

outbreak scenario and never on the lower bound in the 100-outbreak scenario. The likelihood was divergent about 7–10% of the time in the 10-outbreak scenario 0.1–2% of the time in the 100-outbreak scenario. The average correlation between the estimated transmission rate and growth rate was about 0.83 and 0.74 for replicates in the 10- and 100-outbreak scenarios, respectively.

In sum, the method works well with a sufficiently large data set. Moderate amounts of imputation, missing data, and rounding will have mostly modest effects on estimates.

Estimates for norovirus in health-care settings

We fitted our generalized linear model to data from a large prospective study of gastroenteritis in health-care settings [65]. In this one-year study, patients and the care staff assigned to any of about 4500 beds in health-care facilities in the county of Avon, England, were under active surveillance. Trained staff members recorded the dates over which people were symptomatic and took samples that allowed for laboratory confirmation of the causes of outbreaks. Fig. 2.2 shows the cases histories that were used to fit our model.

The predictors in our model were facility type, which indicated whether an outbreak took place in a long-term–care facility (LTCF) or a hospital; size class, which indicated the number of patients and staff in the unit; period, which indicated the time of the study year when the outbreak began; and case type, which indicated whether a case was a patient or a member of the care staff. See the Methods for more details.

For our baseline regression group of patients in large hospitals in the fall and winter, the estimates (95% bootstrap confidence interval) of the transmission rate was 0.0037 (0.0026–0.0052) transmissions per infective–susceptible

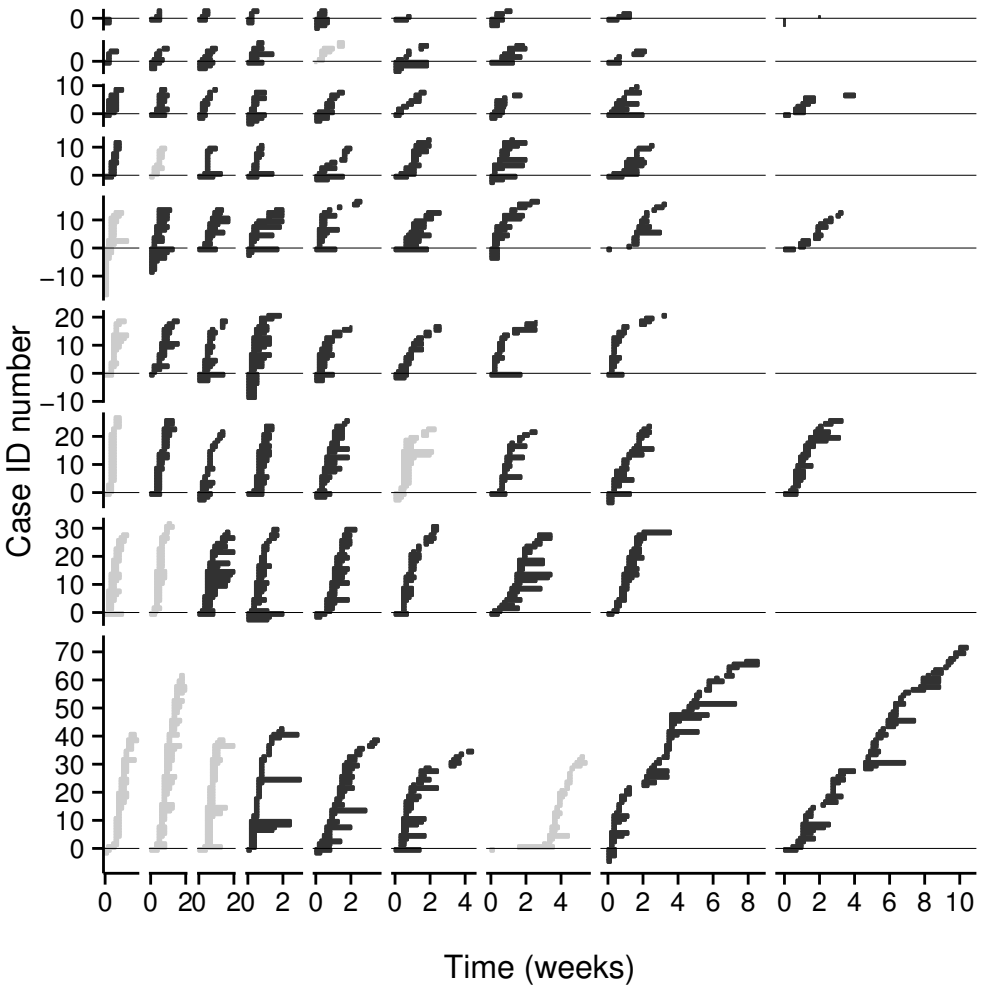

Figure 2.2: Case histories. Each horizontal bar represents the history of a person. The interval between onset of symptoms and relief from symptoms is filled in. Case IDs were assigned by sorting the cases first by onset time, then by relief time, and then by a random ordering. Initial infectives were given negative case IDs. The panels are arranged so that the outbreak size increases from top to bottom and the outbreak length increases from left to right. Case histories from long-term–care facilities (LTCFs) are in light gray. Some of the times of relief from symptoms were imputed as described in the Methods.

day, that of the initial growth rate was 0.27 (0.23–0.30) transmissions per infective day, that of the symptomatic period was 3.35 (3.09–3.57) days, and that of the dispersion parameter ρ for the symptomatic period was 0.57 (0.54–0.65).

Fig. 2.3 shows the effects on these estimates of moving away from the reference group with respect to a predictive variable. The largest effects are the increase in transmission and growths rates in long-term–care facilities (LTCFs) and the reduction in these rates in staff. It appears that transmission rates are higher in the smaller units. Symptomatic periods were estimated to be about 25% shorter for outbreaks in LTCFs and 20% shorter for cases among staff.

Discussion

We have shown that estimation of parameters from many small outbreaks can be done using a generalized linear model. A simulation study demonstrated that we are able to accurately estimate parameters when the data stem from small outbreaks even when the data set is missing some data and about half of recovery times are imputed. Fitting the model to a large number of outbreaks of norovirus, we found that facility type, facility size, and case type seem to have significant effects on outbreak dynamics.

The most striking result of our regression estimates (Fig. 2.3) are the approximately 7-fold increase in transmission rates and 3-fold increase in initial growth rates in the long-term–care facilities (LTCFs) relative to hospitals. Fig. 2.2 shows that LTCF outbreaks do indeed include many of the larger and faster growing outbreaks in the data set.

The higher transmission rates for occupants of LTCFs may be a consequence of occupants having more opportunity to socialize in large groups.

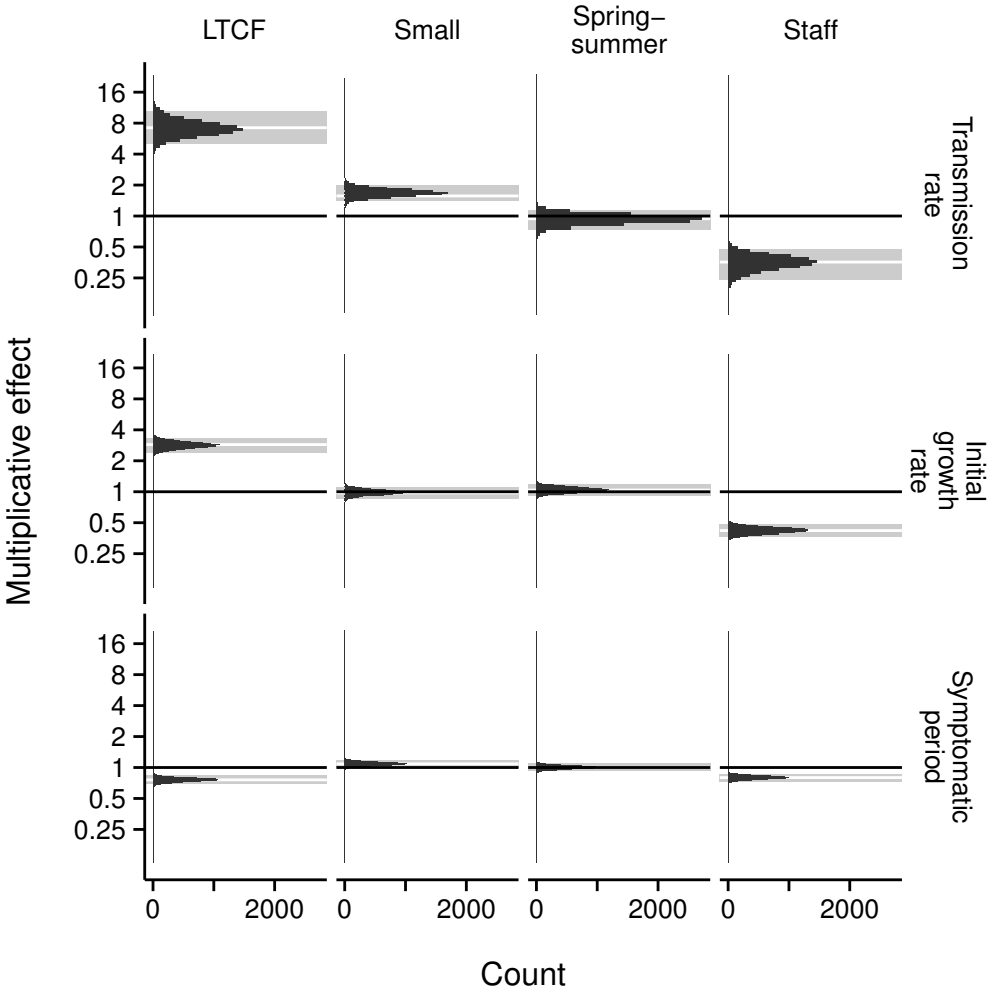

Figure 2.3: Regression effect estimates. The column names indicate predictive variables and the row names indicate parameters. The histograms display the distributions of estimates obtained in a parametric bootstrap. Gray rectangles indicate a 95% confidence interval based on the percentiles of bootstrap estimates. The white horizontal line inside each rectangle indicates the ML estimate. LTCF stands for long-term–care facility.

Alternatively, we may be seeing the effects of our model assumptions being violated. Perhaps foodborne transmission is more common in LTCFs. Hospitals have more rapid turnover of patients, and the exposure of people who arrived in the care unit after the outbreak started will be overestimated in our model. Occupants of LTCFs may vary more in contact rates by virtue of personality differences, and such variation in exposure could lead to a higher initial growth rate [9, pp. 133–138].

The estimates for NoV transmission dynamics we calculated complement results from previous epidemiological analyses of NoV in health-care settings. Previous analyses of our data set [64, 65] had examined how risk of NoV infection or particular symptoms of NoV infection varied with age and other characteristics of people. The current analysis adds to these results by providing estimates for a mechanistic transmission model.

Analysis of a 2003–2006 study of NoV outbreaks in long-term–care facilities (LTCFs) in Oregon [83] suggested that larger facilities may have a higher risk of experiencing outbreaks. Our result that transmission rates are lower in larger facilities suggests that any increased risk that larger facilities have is not caused by increased transmission rates.

A few previous studies have estimated individual-level parameters for NoV that are comparable to our estimates. Using data from a NoV outbreak in a primary school and nursery in Derbyshire, England, O'Neill and Marks [77] estimated that the probability of a susceptible person avoiding infection from an infective person in the school for a day was 0.998. Using the formula $\Pr(\text{avoidance}) = \exp(-\beta \times 1\,\text{susceptible} \times 1\,\text{infective} \times 1\,\text{day})$, our estimates yield $\Pr(\text{avoidance})$ that ranges from about 0.959 for patients in small LTCFs to 0.999 for staff in large hospitals.

Heijne et al. [44] estimated the basic reproduction number of NoV in boy-scout camps to be about 14 and 7, respectively, under two different sets of assumptions. The basic reproduction number R_0 is the initial number of new infections that a single infection will cause. Using the formula $R_0 = (r_{\text{patient}} + r_{\text{staff}})\mu_{\text{patient}}$, our highest R_0 was approximately equal to 3. The relative lowness of our R_0 might reflect contact rates being higher in the camp setting, and it may also reflect the effect of better hygiene in the health-care settings. Heijne et al. [44] estimated that the implementation of an enhanced hygiene protocol drove the reproduction number in the camps down to about 2 and 1, values on par with our own estimates. Our estimates may be more generalizable than the estimates from the boy-scout outbreaks because our data set was larger and included data from both large and small outbreaks.

Zelner et al. [116] used data from a Stockholm outbreak to estimate that the average infectious period was 1.2 days. The setting of these outbreaks was households that included children in daycare centers. Thus, the infectious period may have been shorter in these outbreaks because many of the infectives were likely healthy people between the ages of 5 and 70, whereas people below the age of 5 and, to an even greater degree, people over the age of 70 were overrepresented in our data [65]. In our data, people in these extreme age groups had average symptomatic periods of 3 days [65]. Moreover, the Stockholm estimate is based on imputed infectious periods rather than symptomatic periods, which were not reported. As a result, if the assumed initial number of susceptibles for the Stockholm analysis was too high, the infectious period would have been underestimated.

Although our estimates of the symptomatic period may be relatively long, it is possible that some of the patients were discharged into the commu-

nity before they became asymptomatic. Thus, for patients, our estimates most accurately reflect the period of being symptomatic while simultaneously being in a health-care facility.

The daily transmission rates estimated from the Stockholm data, 0.14 transmissions per infective–susceptible day, are more than 3-fold higher than our highest estimated transmission rate, which was 0.04 transmissions per infective–susceptible day for patients in small LTCFs. The joint estimation approach we used could be applied to the Stockholm data to determine whether the higher transmission-rate estimates may have resulted from underestimation of household sizes. However, the transmission rates may well be different because of differences in hygiene measures and contact rates. Additionally, time-series analysis of outbreak incidence [62] has suggested that transmission rates generally may vary with host, weather, and virus factors. Taken together, these differences may explain the large discrepancy in estimated transmission rates.

In our application, we made the simplifying assumption that the latent period was fixed at its mean, which allowed us to directly calculate infection times from the reported onset of symptoms. The infection times determine the cumulative exposures $h_{i,j}$ in (2.3). Because the cumulative exposure is a non-linear function of time and the mean of a non-linear function of a random variable does not always equal the function evaluated at the random variable's mean, the extent to which latent periods varied in reality likely introduced bias into our calculated cumulative exposures and the estimates based on them. The bias could be either positive or negative depending on whether cumulative exposure usually increases more quickly before or after the assumed transmission times.

Another simplifying assumption we made was that people were only infectious when they were symptomatic. The effect of this and the fixed latent period assumption could be tested by making the infectious period a latent variable that we integrate over to evaluate the likelihood, as in Hohle et al. [47], or by using a kernel-smoothing method, as in Lau and Yip [57]. However, even without such calculations it is clear that if, in reality, the infectious period extends beyond the symptomatic period, our estimates of transmission rates have been inflated by our underestimation of exposure.

From the numerical results displayed in Tables 2.1 and 2.2, we see that highly reliable estimation depends on collection of an extensive data set. The Centers for Disease Control and Prevention (CDC) has recently established a National Outbreak Reporting System that, with the contributions of state health departments, will provide more comprehensive surveillance for all U.S. gastroenteritis outbreaks [18]. However, the data we have analyzed here is more detailed than what is routinely reported to this system. More outbreak investigations are needed to collect detailed data and further characterize modes of transmission. The collection of NoV genomic data may also be of great value [98].

What data are required? Our method can be applied to estimate transmission rates for any data set for which the total number of cases and the average cumulative exposure of individuals in each outbreak can be estimated. To additionally estimate the initial growth rate, we further require an estimate of the area under the curve of the number of infectives over time. If only the times at which individuals stop being infectious are known, these quantities could be estimated using a kernel-smoothing method [57]. Of course, it is also desirable to have data about important covariates for the regression.

Why bother collecting more data? We submit that, for norovirus and many other diseases, there are several use cases for the types of regression estimates for transmission and initial growth rates that we have presented here. Policy-makers can use such estimates to compare the efficacy of different control strategies such as hygiene protocols, isolation measures, prophylactic treatments, and vaccination policies. Those monitoring the small outbreaks of zoonotic diseases may be able to use such estimates to identify variables that make transmission more likely.

Chapter 3

Maximum likelihood estimation of HIV-risk network dynamics from multiple surveys

Introduction

HIV surveillance is a major component of public health efforts around the world. One hundred eighty-six countries contributed data about the HIV epidemic to the 2012 UNAIDS global report [101]. This great effort reflects the great burden of HIV. Based on these data, UNAIDS estimates that, worldwide, 0.8% of adults in the age range of 15-49 years are living with HIV [101, p. 8].

The UNAIDS report also contains reports about the relative risk of different population segments for HIV. Based on data from 50 countries, female sex workers are estimated to be 13.5 times more likely to have HIV than other women [101, p. 21]. In 49 countries, injecting drug users (IDUs) are at least 22 times more likely to have HIV than the general population [101, p. 34]. However, the report also shows substantial variation in estimates from region to region. For example, there are 11 countries in which prevalence among IDUs is at least 50 times the national average prevalence and 2 countries with almost no difference in prevalence between the groups [101, p. 35]. Thus as valuable as group estimates may be about the overall state of the HIV epidemic, they are not necessarily informative of how one particular person's behavior affects her risk of acquiring HIV. Intervention efforts directed at her that are designed

to change population averages may well be wasted efforts.

Many authors [55, 85, e.g.] have pointed out that a person's risk of acquiring HIV depends critically on her position in the network of transmission routes in relation to infected people. We refer to such network as contact networks. Contact network structure may also determine large scale trends in prevalence as well. For example, Morris and Kretzschmar [72] examined the effect of the frequency of concurrent partnerships on the growth and overall size of epidemics, and Volz and Meyers [105] looked at how epidemic trajectories varied with different rates of partner replacement conditional on a set network structure. It is clear from such studies that both concurrency and rates of partner change both may have strong effects on the course of an epidemic.

Although the work of Morris and Kretzschmar [72] and Volz and Meyers [105] clearly demonstrate the importance of network structure, the models used in some sense invert the processes involved because the behavior of individuals is determined by high-level network parameters. It seems more realistic to model the network as a consequence of decisions made by different people that make no special attempts to coordinate their actions. Thus in this paper our main approach is to not consider the number of edges a node has as a model parameter, but rather to consider that the edges between nodes are the realization of simple stochastic processes. We can thus fit this model to survey data by estimating rates at which nodes acquire new neighbors and rates at which nodes lose neighbors, which we refer to as *on-* and *off-rates*, respectively.

This approach is a good way to model this data for a number of reasons. First, the rate parameters can be interpreted in terms of the behavior of individuals. Because HIV control measures often have the goal of changing the behavior of individuals, the parameter estimates could be useful as a mea-

surement of how effective such programs are or need to be. Second, standard statistical methods are often available or can be adapted to fit these models. Third, there are analytic results about several properties of this type of network model [14, 15]. For example, a skewed degree distribution, which often characterizes sexual networks, occurs in this type of network when there is wide variation in the on-rates of nodes [14].

Britton and Lindholm [14] describe two possible relationships between a node's on-rate and its individual propensity for forming new edges, which is called its *social index*. We use what they call the modified model in our analyses. In this model, a node sprouts new edges at a rate proportional to its social index and the free end of this new edge connects to any node in the network with probability proportional to that node's social index.

By fitting the above model to data, we are able to answer several epidemiological questions: For each of needle sharing and unprotected heterosexual relationships, at what rates are people starting and breaking off risky relationships? How do these rates vary between studies? How much variation is there from person to person?

Methods

Data

The data come from a number of different behavioral surveys: Project 90 [114] in Colorado Spring, CO; the Urban [85], Adolescent [84], and Geography studies in Atlanta, GA; and the HAART and Clustering studies. In each of these studies, participants were interviewed one or more times and asked to report the identities of the individuals with whom they had engaged in sex or needle sharing in the last 3 months (6 months for Project 90).

Data preparation

Sexual relations for which perfect condom use was reported were removed. Sexual relations in which the gender of the participant and the reported contact were listed as male were considered MSM (Males who have Sex with Males). Sexual relations in which the gender of the participant and the reported contact were a male–female pair were considered heterosexual. We did not include the MSM contacts in our analysis here.

Two studies in the original data set, Urban2 and ARRA, were removed from analysis because they had little information about break-up rates in them. The needle-sharing data were zero-inflated, which made them difficult to fit as our model had no zero-inflation parameter. We therefore removed people reporting zero needle-sharing and MSM contacts from the data when estimating rates for those parameters.

Model

We consider the number of contacts that a person has over time as following an immigration–death process where λ is the on-rate, or the immigration rate, and μ is the per-contact off-rate, or the death rate. Therefore, the probability $P_n(a)$ that a person has n contacts at age a satisfies the master equations

$$\frac{\mathrm{d}P_n}{\mathrm{d}a} = -(\lambda + \mu n)P_n(a) + \lambda P_{n-1}(a) + \mu(n+1)P_{n+1}(a). \tag{3.1}$$

The age a in this expression is the age since the person began any risk-taking activity such that a person of age less than zero always has zero edges.

The probability generating function $\pi(z, a)$ of the master equations

satisfies

$$\frac{\partial \pi}{\partial a} + \mu(z-1)\frac{\partial \pi}{\partial z} - \lambda(z-1)\pi = 0 \tag{3.2}$$

with initial conditions $\pi(z,0) = z^{n(0)}$. This initial value problem can be solved using the method of characteristics, yielding

$$\pi(z,a) = \exp[-\lambda(1-\mathrm{e}^{-\mu a})(z-1)/\mu](1+\mathrm{e}^{-\mu a}(z-1))^{n(0)}, \tag{3.3}$$

which shows that the number of contacts at age a is the sum of a Poisson random variable with mean given by

$$\lambda(1-\mathrm{e}^{-\mu a})/\mu \tag{3.4}$$

and $n(0)$ Bernoulli random variables with success probability of $\mathrm{e}^{-\mu a}$.

If we observe a person at two ages, a_1 and a_2, the log-likelihood of the proportion y_1 of the number $n(a_1)$ of contacts observed at a_1 that remain at a_2 and the number of new contacts y_2 that were observed only at a_2 is then

$$\begin{aligned}\ell = \ln \binom{n(a_1)}{n(a_1)y_1} + n(a_1)y_1 \ln p + n(a_1)(1-y_1)\ln(1-p) \\ + y_2 \ln m - m - \ln y!,\end{aligned} \tag{3.5}$$

where $m = \lambda(1-\mathrm{e}^{-\mu(a_2-a_1)})/\mu$ and $p = \mathrm{e}^{-\mu(a_2-a_1)}$. We can, in fact, use a similar log-likelihood for the data from every interview because at the first interview we can assume that the person had zero contacts at some earlier age that we consider to be age zero.

We cannot use exactly Eq. 3.5 for our data because we do not know the identities of contacts at exact time points. Rather, we know all the contacts that a person had within a reporting period of 3 to 6 months preceding the interview. Thus we have to consider that any contacts reported may have

begun either before or during the reporting period. Of course, we do know that contacts that were first reported before the last interview began sometime before they were reported. So the times of up to three successive interviews may figure in to the likelihoods. Let t_a denote the interval between the first interview and the beginning of the reporting period for the second interview, t_b the reporting period for the second interview, t_c the interval between the second interview and the third interview, and t_d the reporting period of the third interview.

For the data from interview k of the w_q interviews of participant q,the log-likelihood of the proportion y_1 of the $\tilde{n}$ contacts present at both the first and second interviews that remain at the third interview, the proportion y_2 of the y_3 contacts that were first reported at the second interview that remain at the third interview, and the number y_4 of contacts that are new at the third interview is

$$\begin{aligned}\ell_{k,q} = \ln \Big[& \sum_{i=0}^{y_3} m_1^i m_2^{y_3-i} \frac{\exp(-m_1 - m_2)}{i!(y_3 - i)!} \\ & \times \sum_{j=0}^{\min(y_3 y_2, i)} \binom{i}{j} \binom{y_3 - i}{y_3 y_2 - j} p_1^j p_2^{y_3 y_2 - j} (1 - p_1)^{i-j} (1 - p_2)^{y_3 - i - y_3 y_2 + j} \Big] \\ & + \ln \binom{\tilde{n}}{\tilde{n} y_1} + \tilde{n} y_1 p_1 + \tilde{n}(1 - y_1)(1 - p_1) \\ & + \mathbf{1}_{[k = w_q]} \ln \sum_{i=0}^{y_4} m_3^i m_4^{y_4 - i} \frac{\exp(-m_3 - m_4)}{i!(y_4 - i)!}, \end{aligned} \tag{3.6}$$

where $m_1 = \lambda(1 - \mathrm{e}^{-\mu t_a})/\mu$, $m_2 = \lambda t_b$, $p_1 = \mathrm{e}^{-\mu(t_b + t_c)}$, $p_2 = \mathrm{e}^{-\mu t_c}(1 - \mathrm{e}^{-\mu t_b})/(\mu t_b)$, $m_3 = \lambda(1 - \mathrm{e}^{-\mu t_c})/\mu$, and $m_4 = \lambda t_d$. We go through each interview of each participant and calculate $\ell_{k,q}$ by considering an interview as the third in the above scheme of three interviews. In the case of the first real interview, we imagine a previous interview that took place at the time when

the person was age zero at which they reported zero contacts. Because $\tilde{n}$ and y_3 are necessarily equal to zero in this case, it does not matter that t_a and t_b are undefined.

To model variation from person to person, we use linear predictors for $\ln \lambda$ and $\ln \mu$. We assume a model in which there is a fixed effect for each survey and a random effect for each person. For example, $\ln \lambda_q = \mathbf{X}_{q,*}\boldsymbol{\beta} + U$, where $\mathbf{X}_{q,*}$ is row q of a design matrix based on the study in which each participant is from, $\boldsymbol{\beta}$ is a column vector of unknown regression coefficients, and U is a normally distributed random variable with mean zero and unknown variance.

For $\ln \mu$, we add one extra term to the predictor when predicting y_1 because we observed that predicted values of y_1 were consistently above observed values without this term. This additional parameter is a change in the off-rate that occurs after the first interview.

Results

Distribution of exposure

The contacts a person has are of interest to us only so far as they inform us of risk for an infectious disease. Such risk would be proportional not to the number of contacts someone reports at a particular time, but rather to the total amount of exposure in contact-time a person has. Thus we calculate the distribution of that exposure in the remainder of this section.

In this section, we denote random variables with capital letters and particular values that they may take with corresponding lower-case letters.

We assume that the number of edges incident to a node N (i.e., the node's degree) evolves as an immigration–death process with immigration rate

λ and death rate μ. Our goal is to find the distribution of the cumulative exposure $C = \int_0^a N(t)\,\mathrm{d}t$ for nodes of a given age a. Because the lifetimes of different edges are independent of each other, we can express that area as a mixture of convolutions of lifetime distributions.

A node of age a may have had two types of edges incident to it: (1) those edges that are currently incident to it, and (2) edges that were incident to it only over some interval in $[0, a)$. For the first type of edge, we require time since it became incident to the node. We refer to this as the partial lifetime T_p of the edge and find that its probability density function (p.d.f.) is

$$f_{T_p}(t) = \mu \mathrm{e}^{-\mu t}[u(t) - u(t-a)]/(1 - \mathrm{e}^{-\mu a}), \tag{3.7}$$

where $u(\cdot)$ is the Heaviside step function. For the second type of edge, we find its complete lifetime T_c has the p.d.f.

$$f_{T_c}(t) = \mu^2 (t-a)\mathrm{e}^{-\mu t}[u(t) - u(t-a)]/(\mu a - 1 + \mathrm{e}^{-\mu a}). \tag{3.8}$$

Using Laplace transforms, it is straightforward to find that the sum of n_p partial lifetimes and n_c complete lifetimes had the p.d.f.

$$\begin{aligned} f_{T_p}^{n_p^*} * f_{T_c}^{n_c^*}(t) = &\left(\frac{\mu}{1 - \mathrm{e}^{-\mu a}}\right)^{n_p} \left(\frac{\mu^2}{\mu a - 1 + \mathrm{e}^{-\mu a}}\right)^{n_c} \\ &\times \Big[\sum_{i=0}^{n_p} \sum_{j=0}^{n_c} \sum_{k=0}^{j} \binom{n_p}{i} \binom{n_c}{j,k} a^{n_c - j} (-1)^{i+2j-k} \frac{[t - a(i+j-k)]^{n_p+n_c+j-1}}{(n_p + n_c + j - 1)!} \\ &\times \mathrm{e}^{-\mu t} u(t - a(i+j-k))\Big]. \end{aligned} \tag{3.9}$$

Henceforth, we abbreviate $N(a)$ as N for brevity. The number of edges N incident to a node of age a is Poisson distributed with mean

$$m = \lambda(1 - \mathrm{e}^{-\mu a})/\mu.$$

This Poisson distribution results from a Poisson number Y of immigration events with mean λa in which the number of surviving immigrants X out of the total immigrants y is binomially distributed with probability $m/(\lambda a)$. The p.d.f. for the cumulative exposure C is then

$$f_c(n,a,t) = \sum_{y=n}^{\infty} f_Y(y) f_X(n) f_{T_p}^{n^*} * f_{T_c}^{(y-n)^*}(t)/f_N(n), \tag{3.10}$$

or, more explicitly,

$$\begin{aligned} f_c(n,a,t) = {} & \frac{n!}{m^y \mathrm{e}^{-m}} \\ & \times \sum_{k=0}^{\infty} \frac{(\lambda a)^{k+n} \mathrm{e}^{-\lambda a}}{(k+n)!} \binom{k+n}{n} [m/(\lambda a)]^n [1 - m/(\lambda a)]^k f_{T_p}^{n^*} * f_{T_c}^{k^*}(t) \end{aligned} \tag{3.11}$$

$$= \mathrm{e}^m \sum_{k=0}^{\infty} \frac{(\lambda a)^k \mathrm{e}^{-\lambda a}}{k!} [1 - m/(\lambda a)]^k f_{T_p}^{n^*} * f_{T_c}^{k^*}(t). \tag{3.12}$$

The distribution function that follows from f_c agrees well with simulation (Fig. 3.1).

Goodness of fit

We assessed goodness of fit as follows. After obtaining the maximum likelihood (ML) estimates for the regression coefficients and the variance of the random effect, we obtained empirical Bayes estimates for the specific values of the random effect for each individual. As the left most column in Fig. 3.2 shows, these values were approximately in agreement with the assumed normality of the model. There are probably fewer lower quantiles than expected for both fits as an artifact of the methods; the likelihood of a Poisson distribution producing zero counts does not always increase as fast as the likelihood of a normal deviate decreases as it moves away from the normal's mean.

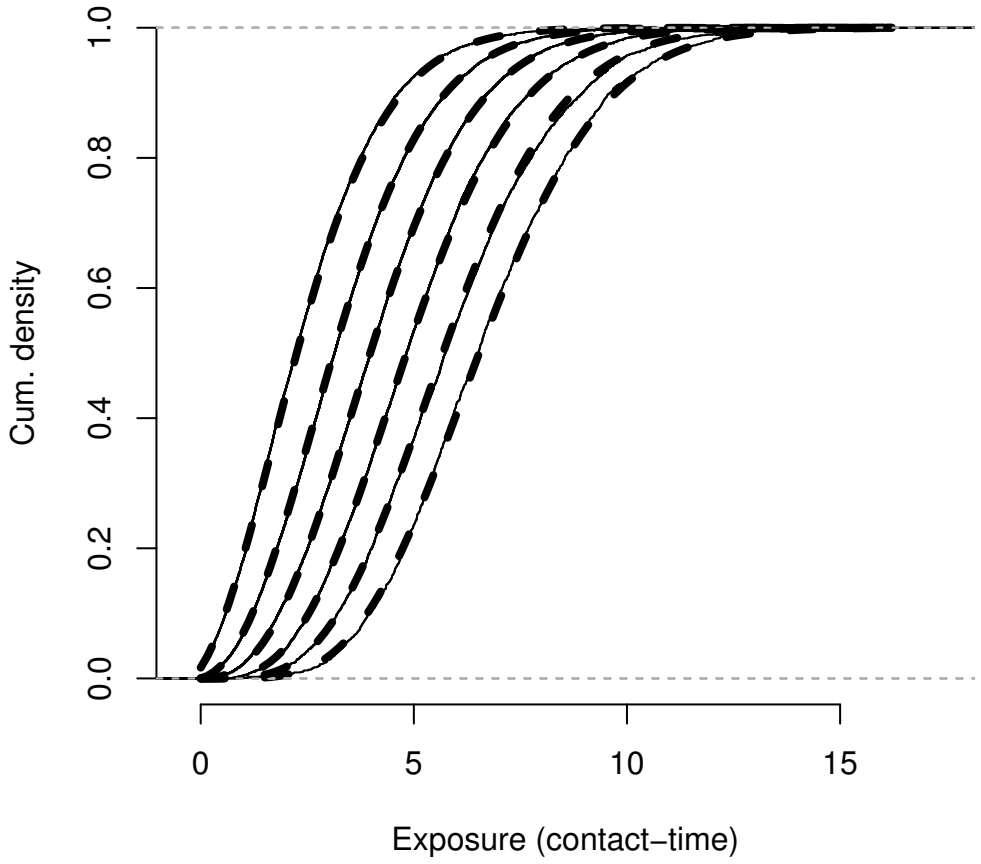

Figure 3.1: Simulated and analytically calculated exposures agree. The dashed lines are the empirical cumulative distribution functions of the areas under the number of contacts from age 0 to 3 conditional on observing from 0 to 5 contacts (lines from left to right) at age 3. The solid lines are distribution function based on f_c (Eq. 3.12). The on-rate is 2 and the off-rate is 1.

According to our model, the observed numbers of contacts and proportions of remaining contacts all may come from different distributions. Therefore, quantile-quantile plots are not of any use for checking distributional assumptions. However, we can put the information in the quantiles on a similar scale by plugging the quantiles into a distribution function. The resulting probabilities should be approximately uniformly distributed. Columns 2 to 4 in Fig. 3.2 thus provide confirmation that no distributional assumptions are grossly violated. We used the empirical Bayes estimates of on-rates and off-rates as the parameters in numerically-evaluated distribution functions.

Estimates

For Project 90, which was the intercept in our linear predictor, the maximum likelihood (ML) estimate of the median daily on-rate (Wald standard error) for heterosexual and needle-sharing contacts were 0.005 07 (0.000 22) and 0.005 21 (0.000 38), respectively, where we have used the delta method to calculate standard errors. The corresponding daily off-rates were 0.004 13 (0.000 15) and 0.004 91 (0.000 33). The standard deviation of the random effect (the standard deviation on the log scale) was about 0.760 (0.022) for heterosexual contacts and 0.755 (0.052) for needle-sharing contacts. Fig. 3.3 illustrates the fitted density of heterosexual on-rates for Project 90.

The estimates for changes from these baseline values are displayed in Fig. 3.4. For heterosexual contacts, the ML on-rates for all other studies were significantly lower than those of Project 90. The off-rates for heterosexual contact appear to be about the same in the other studies except for the Urban and Clustering studies, which have higher off-rates. The small number of significant difference in off-rates as compared to on-rates may reflect the smaller

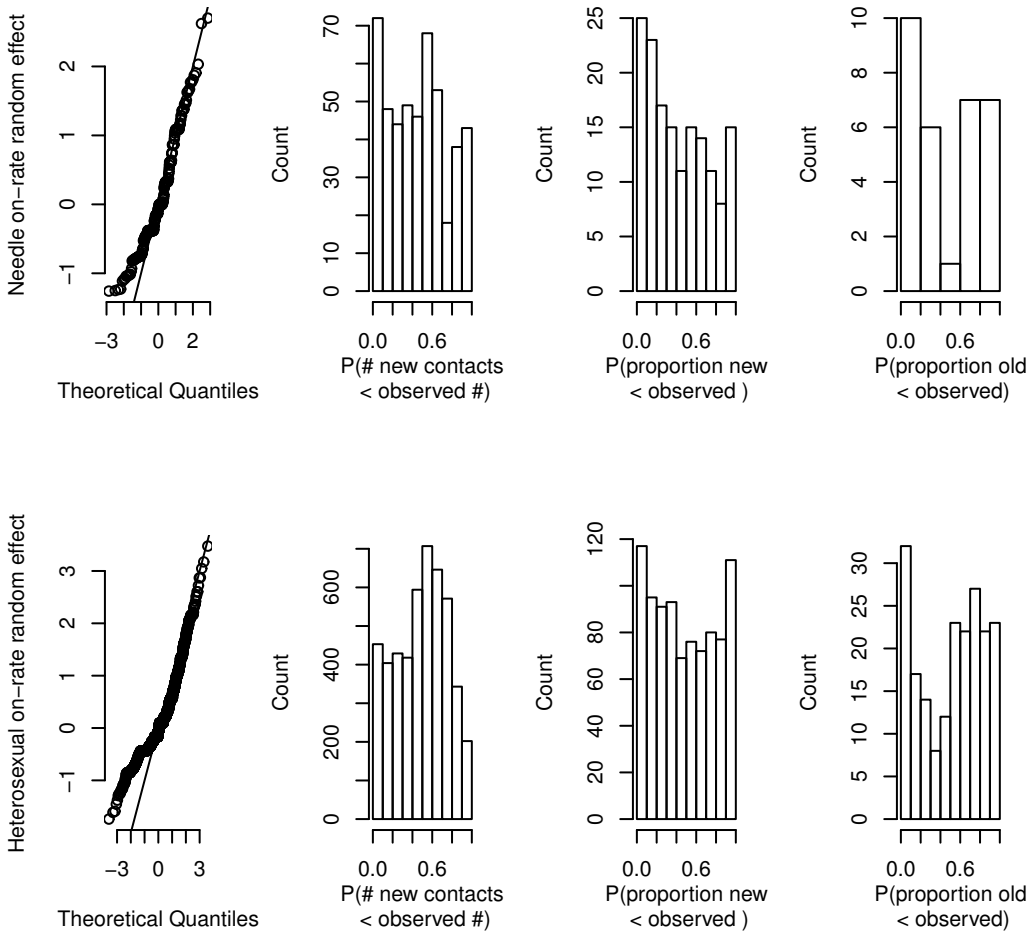

Figure 3.2: Diagnostic plots of the model fits. In column 1, studentized quantiles from the estimated random effect are plotted against standard normal quantiles. All of the panels in each row are from the same model. The histograms in columns 2 to 4 display how the observed data is distributed on the fitted models' distribution functions.

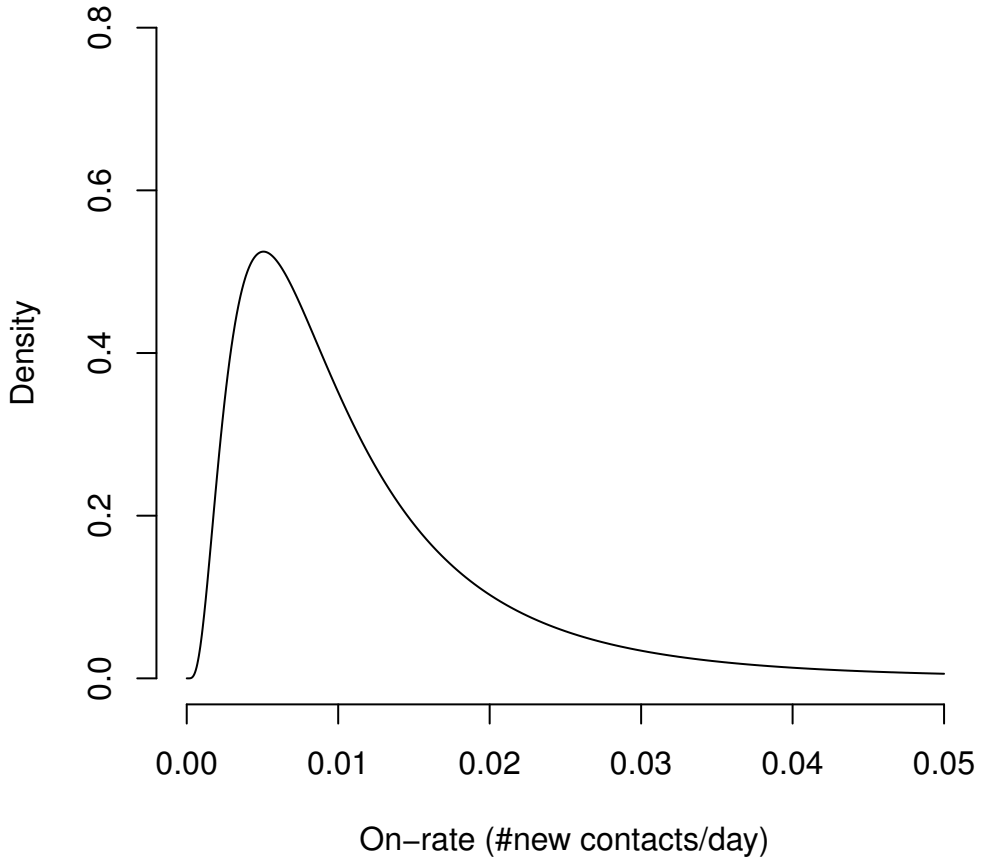

Figure 3.3: Fitted density of on-rates for heterosexual contacts in Project 90.

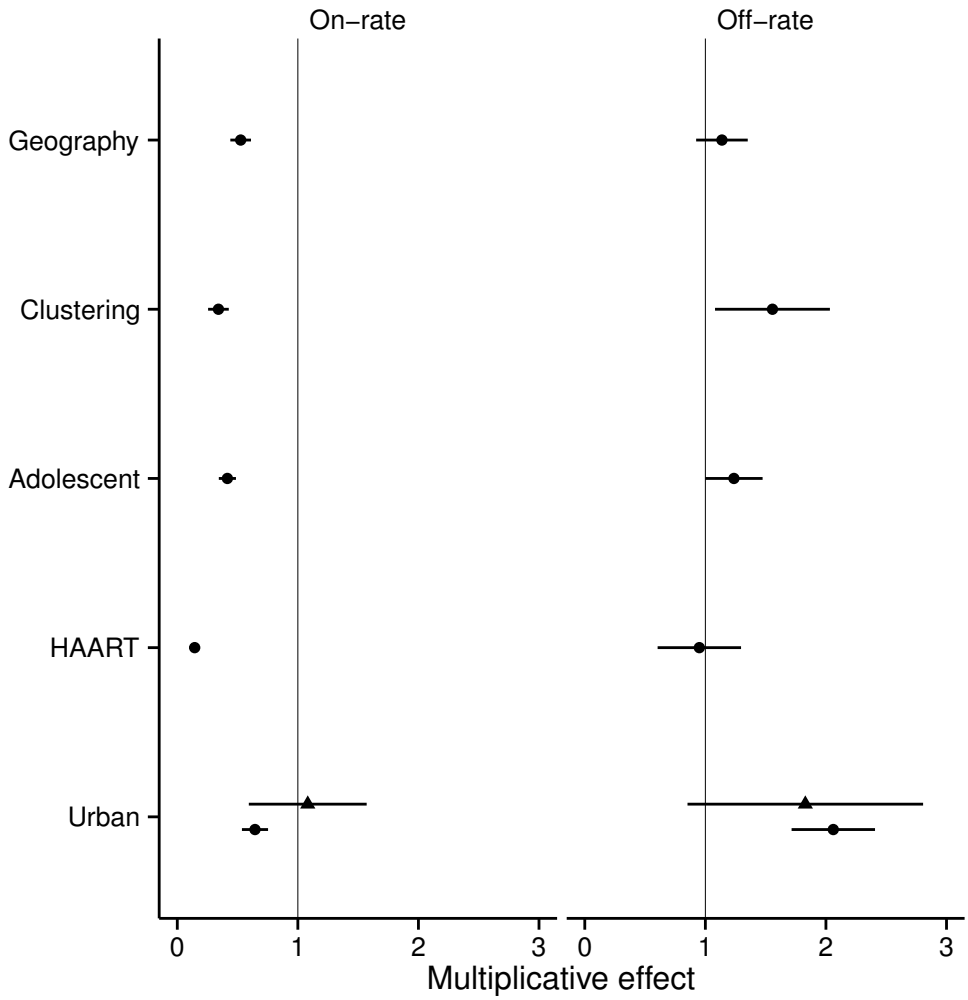

Figure 3.4: Regression estimates for on- and off-rates. Labels on the y axis name which study regression estimates at that height correspond to. Circular marks are for heterosexual contacts, triangular for needle-sharing. The horizontal lines are 95% Wald confidence intervals.

amount of information in the data to estimate them, which is evident in the larger error bars for off-rates (Fig. 3.4). For the needle-sharing contacts, the rates in the Urban study did not differ significantly from those of Project 90.

A significant second-interview effect was found for both types of contact. For heterosexual contacts, the multiplicative effect was 0.312 (0.030). For needle-sharing, the multiplicative effect was 0.374 (0.082). Thus our model detects the presence of distinct long- and short-term risky contacts.

Discussion

We have derived an expression for how the total contact-years a person has may vary given her total time active, her current number of risky contacts, and her rates of gaining and losing partners (on- and off-rates). This result contributes to our ability to account for the volatility of people's contacts when investigating the relationship between risk-taking behavior and infection. Further to that end, we have estimated on- and off-rates with data from multiple surveys and identified significant differences in these parameters between surveys, between people within surveys, and across interviews for the same person. The on-rates in Project 90 tend to be about twice those of the other studies. The off-rates are about the same in all studies with the exception of the Urban study, which had an off-rate that was roughly twice that of Project 90's. The dispersion of on-rates among people within a survey was about the same for both sexual and needle-sharing contacts. Also, a significant decrease in off-rates was found for both needle-sharing and sexual contacts following the second interview in which they were reported.

The estimated on- and off-rates for our baseline group of Project 90 participants were often in an approximately one-to-one ratio and the median

of on-rates was about 1 partner gained every 200 days. Parameters in this range result in the number of contacts over time a person has being highly volatile. The situation is similar to what is depicted in Fig. 3.1, wherein inter-quartile ranges of cumulative exposure in contact-years are on the order of one to five times the medians. The implication is that a stochastic model of exposure is vital to accurately calculate the variance of estimates of exposure. Underestimating the variance could lead to underestimating the association between contact and HIV infection, as a consequence of regression dilution [51].

Although Britton and Lindholm [14] discussed the importance of allowing for both long- and short-term contacts when fitting the model we used to real data, they did not attempt to do so in the toy data analysis they included in the paper that introduced the model. And although the surveys collect data about the nature of the reported relationships which provides some indication of whether they are likely to be long-lived, our estimates are likely the first that quantify the difference in the lengths of these relationships for these data. Such quantification is necessary to accurately model the spread of HIV through a contact network.

Our estimated on-rates seem low compared to comparable estimates from some previous studies. For example, in a sample of 78 MSM AIDS patients early in the U.S. epidemic, more than 64.1% of patients reported having 50 or more partners in the year before AIDS onset, and 52.6% of patients reported 1000 or more sexual partners in their lifetime [4]. Klovdahl [55] cites William [113] as estimating the average number of partners for MSM to be 1000 over the course of their lives. If we assume activity takes place over a 50 year period, we obtain an on-rate estimate of 0.055, whereas our ML means are about 0.007 for

both needle and sexual contacts. Of course, we may expect MSM, particularly those infected early in the epidemic, to have higher on-rates than heterosexuals or needle-sharers. However, preliminary analysis of the MSM contacts in our data has not led to estimates 10-fold higher than what we present here. What seems to explain the difference better is that we remove a significant fraction of sexual contacts because condoms were always used.

If we applied a correction to our estimates that accounted for how people with many contacts are more likely to be participants in the surveys, our mean estimates would almost certainly be lower. We have decided not to apply such a correction for several reasons. First, methods of recruitment varied between studies. For example, the Urban study had two sampling chains of respondents [85]. In one, the chains were extended by selecting a contact at random from the respondent's set of contacts. In the other, the chains were extended by contacts selected by the investigators. In the HAART study, patients at a large metropolitan hospital constituted the sample. None of the studies had sampling methods that fully satisfied the assumptions of respondent-driven sampling (RDS) estimators [104], which use a person's network size (i.e., degree) as an importance sampling weight to calculate overall population averages. For example, putative partners that are interviewed do not always report each other as contacts, which can lead to significant bias [66]. Another reason we decided not to attempt correcting for sampling is that a recent evaluation of RDS [68] in one real-world population found that it failed to remove the bias caused by non-representative sampling. Altogether, it seems likely that correcting for sampling may make our estimates worse. Nevertheless, caution must be taken not to generalize our estimates beyond the populations they are sampled from, which are admittedly somewhat haphazardly determined but

are surely high-risk populations.

There are many possible covariates for our rates parameters in the data that were not included in the model. For example, the age of the respondent, whether the respondent identified themselves as a person who receives money or drugs in exchange for sex, and the year of the study could all conceivably affect the parameters. Multi-model inference (using the model-averaging methods described by Burnham and Anderson [16], for example) on a collection of models that are epidemiologically plausible would lead to more reliable estimates.

The network information in the data also remains untapped by our current methods of estimation. Many social networks exhibit assortativity [75], with like people being more likely to be found together. Thus we might expect people with high on-rates to be more likely to have relationships with other people with high on-rates. Alternatively, we might expect disassortativity if, for example, the highest on-rates are estimated for sex workers that rarely have relationships with each other. We could test such hypotheses by using methods from spatial statistics. One approach would be a geospatial approach of adding a random effects to the on-rates with covariances that depend on the geodesic distance between people in the contact network. Another would be a conditional autoregressive (CAR) approach of making the random effect of a person's on-rate have an expected value that is equal to the on-rates of that person's neighbors. Simulation may be necessary to determine which of these approaches represents the best balance of computational tractability and estimation performance. These spatial methods could of course also be used with the geographic data collected to estimate correlations in physical space.

Appendix: Supplement to Chapter 2

First, we describe a simple method of finding maximum likelihood estimates for the transmission rate and initial number of susceptibles in the case of a single outbreak. Second, we extend this method to the case of multiple outbreaks. We finish with a proof of consistency for the multiple outbreak case.

Model

To make the appendix self-contained, we repeat some of the model description of the main text here, except that we do not distinguish between multiple infective types for simplicity.

At the beginning of an outbreak, $t = 0$, a population is made up of $Y(0)$ *infective* people and $X(0)$ *susceptible* people. Each person in this model has the same instantaneous rate at which infectives may infect her, which we call the *transmission rate.* Each person, if infected, will experience an infectious period of random length drawn from a gamma distribution. As time moves forward, each infective person transmits infection to each susceptible person at the points of a Poisson process such that the rate at which new infections appear is $\beta X(t)Y(t)$, where β is the transmission rate. A susceptible person that contracts infection becomes infective after being in a *latent* state for a fixed time period. She then becomes a *recovered* person after a randomly-distributed infectious period. No other transitions in state occur.

Our outbreak model departs from the stochastic counterpart of the common Kermack–McKendrick susceptible–infective–recovered (SIR) model,

which is often called the general stochastic epidemic model, in two ways besides the addition of a latent state. First, we do not make our transmission rate depend on the total number of people N in the population. This departure prevents the need for N to be estimated, and it is appropriate when an infective person may be able to infect every susceptible person in the population with approximately the same probability. Second, we let the latent period and the infectious period be non-Markovian, which is more realistic.

We have described our model above in terms of how infection spreads. To analyze the model, we find it helpful to formulate it differently. We imagine that each of the $X(0)$ susceptibles at time zero has a threshold to exposure drawn from an exponential distribution with rate β. We refer to these values as the *scaled transmission times* for each person. The time-evolution of the cumulative exposure of susceptible people to infective people relates the scaled transmission times to observed transmission times. The value of that cumulative exposure at some time t is given by the area under the curve Y from the beginning of the outbreak up until t. We use $h(t)$ as shorthand for that integral. A person with a scaled transmission time b becomes latently infected when h is equal to b. This view of an epidemic model was described by Sellke [91] and has been called a Sellke construction [3].

The single outbreak case

As described in the main text, transmission and recovery times observed with data from a single outbreak are not, in general, independent random variables. However, in the case of a major outbreak the correlations become small. Loosely speaking, a major outbreak is an event in which a small number of infective people cause a significant fraction of a large population of initial

susceptibles to become infected. In this case, the scaled transmission times that we can observe are approximately independent and identically distributed (IID) exponential random variables. The likelihood of a parameter vector $\theta = (\beta, X^{(0)})$ is then

$$l(\theta) = X^{(0)}!/\left(X^{(0)} - k\right)! \exp\left(-\beta\tau(X^{(0)} - k)\right) \times \prod_{i=1}^{k}[\beta Y(e_i) \exp\left(-\beta h(e_i)\right)], \tag{A.1}$$

where $X^{(0)}$ is shorthand for the number of initial susceptibles $X(0)$, τ is shorthand for the cumulative exposure of people who remain susceptible throughout the entire outbreak, k is the number of initial susceptibles that are infected during the outbreak, and e_i is the time at which the ith infection occurs in the outbreak. The product over i in Equation (A.1) is the joint density of the transmission times. The factor of $\exp\left(-\beta\tau(X^{(0)} - k)\right)$ is the probability that all $(X^{(0)} - k)$ susceptible people remaining at the end of the outbreak were able to avoid infection for as long as they did. The factor of $X^{(0)}!/(X^{(0)} - k)!$ is the number of ways to label the $X^{(0)}$ initial susceptibles with k unique labels (one label for each transmission event) and $X^{(0)} - k$ identical labels (the label for remaining susceptible throughout the outbreak).

We maximize the likelihood in Equation (A.1) by minimizing the negative log-likelihood function $f(\theta)$:

$$\begin{aligned} f(\theta) = -\ln l(\theta) = &-\ln\Gamma(X^{(0)} + 1) + \ln\Gamma(X^{(0)} - k + 1) \\ &+ \beta\tau(X^{(0)} - k) - \textstyle\sum_i \ln Y(e_i) \\ &- k\ln\beta + \beta\textstyle\sum_i h(e_i), \end{aligned} \tag{A.2}$$

where $\Gamma(\cdot)$ is the Gamma function. As will become clear below, the maximum

of $l(\theta)$ occurs at one of the two integers that bracket the $X^{(0)}$ that minimizes $f(\theta)$.

Equation (A.2) is defined on the space $X^{(0)} \geq k, \beta > 0$ and the model definition implies that $k \geq 0$, $\sum_i h(e_i) > 0$, and $\tau > 0$. To find the minimum, we use the partial derivatives

$$f_\beta = \frac{\partial f}{\partial \beta} = -k/\beta + \tau(X^{(0)} - k) + \textstyle\sum_i h(e_i), \tag{A.3}$$

$$f_{X^{(0)}} = \frac{\partial f}{\partial X^{(0)}} = -\psi(X^{(0)} + 1) + \psi(X^{(0)} - k + 1) + \beta\tau, \tag{A.4}$$

where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is the digamma function. The critical points $(X^{(0)*}, \beta^*)$ satisfy

$$\beta^* = m(X^{(0)*}), \tag{A.5}$$

$$\beta^* = [\psi(X^{(0)*} + 1) - \psi(X^{(0)*} - k + 1)]/\tau, \tag{A.6}$$

where

$$m(X^{(0)}) = k/[\tau\left(X^{(0)} - k\right) + \textstyle\sum_{i=1}^{k} h(e_i)]. \tag{A.7}$$

We next deduce from $f_{\beta\,\beta} = \partial^2 f/\partial\beta^2 = k/\beta^2 > 0$ that $m(X^{(0)})$ gives the value of β that minimizes f along the line $X^{(0)}$. Thus, our minimization problem is effectively one-dimensional along $X^{(0)}$.

Equations similar to (A.5) and (A.6) have been used in earlier work. For example, after some algebra, it is possible to see that Equations (A.5) and (A.6) are approximately the same as the equations that Huggins et al. [49] derived using martingales to jointly estimate β and $X^{(0)}$. Rida [82] demonstrates that Equations (A.5) and (A.6) do not hold in the case of a minor outbreak. However, as described in the main text, generalizations of Equations (A.5) and (A.6) become valid in the limit of a large number of outbreaks.

Minimization procedure

We wish to find a reliable way of minimizing the negative log-likelihood function Equation (A.2) on the space $X^{(0)} \geq k, \beta > 0$. We know from the definition of our model that $k \geq 0$, $\sum_i h(e_i) \geq 0$, and $\tau > 0$. All symbols are as defined in the description of the model.

The Jacobian of f is given by Equations (A.3) and (A.4). In the case that $k = 0$, we have simply

$$f_\beta = X^{(0)}\tau, \tag{A.8}$$

$$f_{X^{(0)}} = \beta\tau. \tag{A.9}$$

Thus f increases with both β and $X^{(0)}$ in this case, and we minimize f by minimizing $X^{(0)}$ and β.

When $k > 0$, stationary points occur at the points $(X^{(0)*}, \beta^*)$ that satisfy Equations (A.5) and (A.6). Because we are considering $k > 0$, $f_{\beta\,\beta} = k/\beta^2 > 0$. Therefore, f is a convex function of β along the line $X^{(0)} = C$ for some $C \geq k$. Therefore, $m(X^{(0)})$ gives the value of β that minimizes f along the line $X^{(0)} = C$. Thus, our minimization problem is effectively a matter of finding the $X^{(0)}$ that minimizes

$$\begin{aligned}\tilde{f}(X^{(0)}) = &- \ln\Gamma(X^{(0)}+1) + \ln\Gamma(X^{(0)}-k+1) \\ &+ \frac{(X^{(0)}-k)k\tau}{X^{(0)}\tau - k\tau + \sum h(e_i)} - \sum_i \ln Y(e_i) \\ &- k\ln\frac{k}{X^{(0)}\tau - k\tau + \sum h(e_i)} + k\frac{\sum h(e_i)}{X^{(0)}\tau - k\tau + \sum h(e_i)}\end{aligned} \tag{A.10}$$

on the interval $[k, \infty)$.

Taking the derivative of $\tilde{f}$ with respect to $X^{(0)}$, we have

$$\tilde{f}_{X^{(0)}} = \frac{k}{X^{(0)} - k + U} - \psi(X^{(0)}+1) + \psi(X^{(0)}-k+1), \tag{A.11}$$

where $U = \sum_i h(e_i)/\tau$. When $k = 1$, $\tilde{f}_{X^{(0)}} = 0$ for all feasible $X^{(0)}$, and f is equally low all along a trough traced out by the parametric curve $\big(X^{(0)}, m(X^{(0)})\big)$.

As $X^{(0)}$ increases, the sign of $\tilde{f}_{X^{(0)}}$ may change from negative to positive, but it never changes from positive to negative. To see this, we first use a few identities to rewrite some terms in $\tilde{f}_{X^{(0)}}$:

$$-\psi(X^{(0)}+1) + \psi(X^{(0)}-k+1) = -\int_0^1 \frac{1-y^k}{1-y} y^{X^{(0)}-k}\,\mathrm{d}y \tag{A.12}$$

$$= -\int_0^1 (1+y+y^2+\cdots+y^{k-1}) y^{X^{(0)}-k}\,\mathrm{d}y. \tag{A.13}$$

Plugging the last expression into Equation (A.11), we obtain

$$\begin{aligned}
\tilde{f}_{X^{(0)}} &= \frac{k}{X^{(0)}-k+U} - \left(\frac{1}{X^{(0)}} + \frac{1}{X^{(0)}-1} + \cdots + \frac{1}{X^{(0)}-k+1}\right) \\
&= \left(\frac{1}{X^{(0)}-k+U} - \frac{1}{X^{(0)}}\right) + \left(\frac{1}{X^{(0)}-k+U} - \frac{1}{X^{(0)}-1}\right) + \cdots \\
&\quad + \left(\frac{1}{X^{(0)}-k+U} - \frac{1}{X^{(0)}-k+1}\right) \\
&= \frac{k-U}{(X^{(0)}-k+U)X^{(0)}} + \frac{k-U-1}{(X^{(0)}-k+U)(X^{(0)}-1)} + \cdots \\
&\quad + \frac{k-U-k+1}{(X^{(0)}-k+U)(X^{(0)}-k+1)}.
\end{aligned} \tag{A.14}$$

The numerators of the terms in Equation (A.14) decrease from left to right such that any positive values will be in the left-most terms. The denominators in Equation (A.14) must all be positive and they also decrease from left to right. Therefore, the farther to the right we go, the more weight terms in the numerator have on the value of $\tilde{f}_{X^{(0)}}$. So more negative terms always have relatively more weight. But as $X^{(0)}$ increases, the relative difference $1-(X^{(0)}-a-1)/(X^{(0)}-a)$ between denominators that are a and $a+1$ terms

to the right of the first term decreases at the rate of $1/(X^{(0)} - a)^2$. Therefore, the weighting of terms becomes more uniformly distributed, and as the more negative terms lose their disproportionate influence, the sign of $\tilde{f}_{X^{(0)}}$ can only change from negative to positive.

The sign of $\tilde{f}_{X^{(0)}}$ will be positive for sufficiently large $X^{(0)}$ if the summation of the numerators in Equation (A.14) is positive. This fact follows from the relative difference of the denominators approaching zero and the weighting of the numerator terms becoming uniform. Thus we have the following necessary condition for $\tilde{f}_{X^{(0)}}$ to be positive: $U < (k+1)/2$.

We now can determine the sign of $\tilde{f}_{X^{(0)}}$ for all $X^{(0)} \geq k$ after considering the sign of $\tilde{f}_{X^{(0)}}$ when $X^{(0)} = k$. When $X^{(0)} = k$, $\tilde{f}_{X^{(0)}} = k/U - H_k$, where the harmonic number $H_k = \sum_{i=1}^{k}(1/i)$. Therefore, for $U \leq k/H_k < (k+1)/2$, $\tilde{f}_{X^{(0)}}$ is non-negative for all feasible $X^{(0)}$, having a root at $X^{(0)} = k$ if $U = k/H_k$. For $k/H_k < U < (k+1)/2$, $\tilde{f}_{X^{(0)}}$ is negative on $[k, X^{(0)*})$ and positive on $(X^{(0)*}, \infty)$. For $U \geq (k+1)/2$, $\tilde{f}_{X^{(0)}}$ is negative for all feasible $X^{(0)}$.

Our procedure for minimizing f when $k > 1$ follows from our knowledge of $\tilde{f}_{X^{(0)}}$. If $U \leq k/H_k$, the minimum of $f(X^{(0)}, \beta)$ occurs at $\big(k, m(k)\big) = \big(k, k/\sum_i h(e_i)\big)$. If $k/H_k < U < (k+1)/2$, the minimum occurs at the unique solution to Equations (A.5) and (A.6), which we can find by using a grid-based search for the root of $\tilde{f}_{X^{(0)}}$. If $U \geq (k+1)/2$, we know that f always decreases along the parametric curve $\big(X^{(0)}, m(X^{(0)})\big)$ as $X^{(0)}$ increases. Figure A1 contains plots of representative curves for each of the three cases in our minimization procedure.

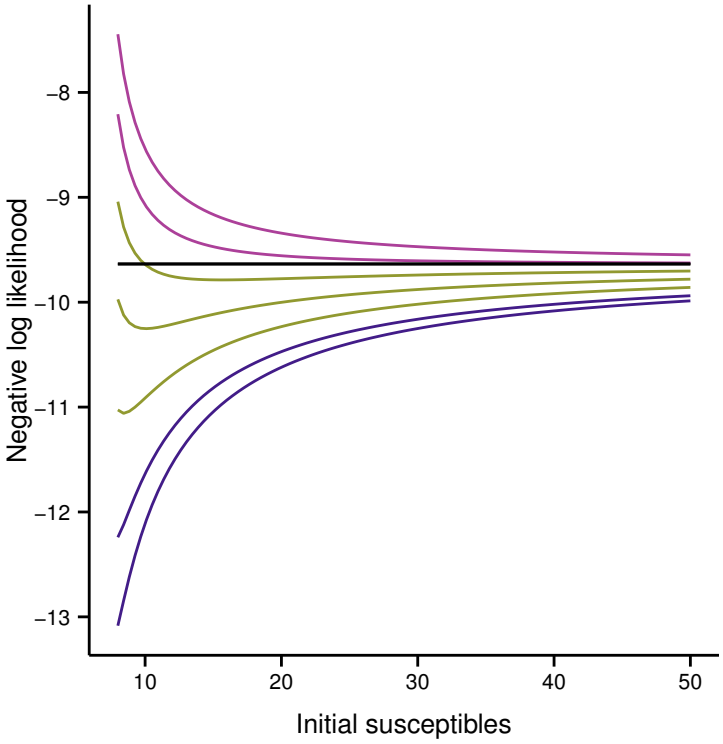

Figure A.1: Representative curves of the negative log-likelihood function, Equation (A.10), for each of the three cases in our minimization procedure. Magenta curves continuously decrease towards an asymptote, which is drawn with a black line. Olive curves have a minimum at an $X^{(0)}$ that is some finite distance above the minimum of the range of $X^{(0)}$. Blue curves have a minimum at the lower limit of the range of $X^{(0)}$. Parameters: $k = 8$; $\tau = 1$; $\sum_i \ln Y(e_i) = 1$; $\sum_i h(e_i) = \{1.1c_2, c_2, c_1 - 0.75(c_1 - c_2), c_1 - 0.5(c_1 - c_2), c_1 - 0.25(c_1 - c_2), c_1, 0.9c_1\}$, where $c_1 = k/H_k$ and $c_2 = (k+1)/2$.

The multiple outbreak case

Extension of the estimators above to the case where we have data from multiple outbreaks, with each outbreak being a realization of a model with the same parameters, is straightforward. In this case, our objective function generalizes to

$$\begin{aligned} f(\theta) = -\ln l(\theta) = \sum_{j=1}^{n} [&- \ln \Gamma(X^{(0)} + 1) + \ln \Gamma(X^{(0)} - k_j + 1)] \\ &+ \sum_{j=1}^{n} [\beta \tau_j (X^{(0)} - k_j) - \textstyle\sum_i \ln Y_j(e_{i\,j})] \\ &+ \sum_{j=1}^{n} [-k_j \ln \beta + \beta \textstyle\sum_i h_j(e_{i\,j})], \end{aligned} \tag{A.15}$$

where the value of the subscript j indicates which outbreak a datum is from and n is the number of outbreaks. The Jacobian of f is now

$$f_\beta = -\textstyle\sum_j k_j/\beta + \sum_j \tau_j (X^{(0)} - k_j) + \sum_{i\,j} h_j(e_{i\,j}), \tag{A.16}$$

$$f_{X^{(0)}} = \textstyle\sum_j [-\psi(X^{(0)} + 1) + \psi(X^{(0)} - k_j + 1)] + \sum_j \beta \tau_j. \tag{A.17}$$

If all k are equal to zero, then

$$f_\beta = X^{(0)} \textstyle\sum_j \tau_j, \tag{A.18}$$

$$f_{X^{(0)}} = \beta \textstyle\sum_j \tau_j, \tag{A.19}$$

and we minimize f by minimizing $X^{(0)}$ and β.

When $k_j > 0$ for any j, stationary points occur at the points $(X^{(0)*}, \beta^*)$ that satisfy

$$\beta^* = m(X^{(0)*}), \tag{A.20}$$

$$\beta^* = \textstyle\sum_j [\psi(X^{(0)*} + 1) - \psi(X^{(0)*} - k_j + 1)] / \sum_j \tau_j, \tag{A.21}$$

where

$$m(X^{(0)}) = \sum_j k_j / \sum_j [X^{(0)}\tau_j - k_j\tau_j + \sum_i h_j(e_{i\,j})]. \tag{A.22}$$

Because we are considering that some $k_j > 0$, $f_{\beta\,\beta} = \sum_j k_j/\beta^2 > 0$. Therefore, f is a convex function of β along the line $X^{(0)} = C$ for some $C \geq \max\{k_j\}$. Therefore, $m(X^{(0)})$ gives the value of β that minimizes f along the line $X^{(0)} = C$. Thus, our minimization problem is effectively a matter of finding the $X^{(0)}$ that minimizes

$$\begin{aligned}\tilde{f}(X^{(0)}) = \sum_{j=1}^{n} &[-\ \ln\Gamma(X^{(0)}+1) + \ \ln\Gamma(X^{(0)} - k_j + 1)] \\ &+ \sum_{j=1}^{n}[m(X^{(0)})\tau_j(X^{(0)} - k_j) - \textstyle\sum_i \ln Y_j(e_{i\,j})] \\ &+ \sum_{j=1}^{n}[-k_j\ \ln m(X^{(0)}) + m(X^{(0)})\textstyle\sum_i h_j(e_{i\,j})]\end{aligned} \tag{A.23}$$

on the interval $[\max\{k_j\}, \infty)$.

Taking the derivative of $\tilde{f}$ with respect to $X^{(0)}$, we have

$$\tilde{f}_{X^{(0)}} = \frac{\sum_j k_j \sum_j \tau_j}{\sum_j [\tau_j(X^{(0)} - k_j + U_j)]} + \sum_j [-\psi(X^{(0)}+1) + \psi(X^{(0)} - k_j + 1)], \tag{A.24}$$

where $U_j = \sum_i h_j(e_{i\,j})/\tau_j$. When $\max\{k_j\}$ is equal to one, f is at its minimum all along the parametric curve $(X^{(0)}, m(X^{(0)}))$.

As $X^{(0)}$ increases, the sign of $\tilde{f}_{X^{(0)}}$ may change only from negative to positive. After using Equation (A.13) to rewrite each term in the summation

over digamma functions in Equation (A.24), we can regroup terms to obtain

$$\begin{aligned}
\tilde{f}_{X^{(0)}} &= \sum_{J,k_J>0} \sum_{i=1}^{k_J} \left(\frac{\sum_j \tau_j}{\sum_j \tau_j (X^{(0)} - k_j + U_j)} - \frac{1}{X^{(0)} - i + 1} \right) \\
&= \sum_{J,k_J>0} \sum_{i=1}^{k_J} \frac{\sum_j \tau_j k_j - \sum_j U_j \tau_j - i \sum_j \tau_j + \sum_j \tau_j}{\sum_j \tau_j (X^{(0)} - k_j + U_j)(X^{(0)} - i + 1)} \\
&= \sum_{i=1}^{\max\{k_j\}} n p_i \frac{\sum_j \tau_j k_j - \sum_j U_j \tau_j - i \sum_j \tau_j + \sum_j \tau_j}{\sum_j \tau_j (X^{(0)} - k_j + U_j)(X^{(0)} - i + 1)}, \qquad \text{(A.25)}
\end{aligned}$$

where $p_i = \sum_{J,k_J>i} n^{-1}$ is the fraction of outbreaks in which k exceeds i.

The last expression in Equation (A.25) shares several properties with that of Equation (A.14): numerators decrease from left to right, denominators are all positive and decrease from left to right, and relative differences between successive denominators decrease as $X^{(0)}$ increases. It follows that the right-hand side of Equation (A.25), like that of Equation (A.14), can only change sign from negative to positive.

Equation (A.25), like Equation (A.14), tells us that $\tilde{f}_{X^{(0)}}$ can only be positive if the numerators add up to a positive number. The necessary condition for positive $\tilde{f}_{X^{(0)}}$ in the case of multiple outbreaks is then: $(\sum_j k_j) \sum_j (\tau_j k_j - U_j \tau_j) - (\sum_j \tau_j) \sum_j [(k_j - 1)k_j/2] > 0$.

The minimum feasible value for $X^{(0)}$ is $\max\{k_j\}$ in the multiple outbreak case. We see no simple and general expression for the condition for a positive $\tilde{f}_{X^{(0)}}$ at this boundary. But the lack of an analytic expression here does not change our general procedure for finding the minimum of $\tilde{f}_{X^{(0)}}$. We can numerically evaluate $\tilde{f}_{X^{(0)}}(\max\{k_j\})$ and then know that the minimum occurs there if the sign is non-negative. If the sign of $\tilde{f}_{X^{(0)}}(\max\{k_j\})$ is negative, then we can evaluate the condition for positive $\tilde{f}_{X^{(0)}}$ and know that $\tilde{f}_{X^{(0)}}$

always decreases towards an asymptote if the condition is not satisfied. If the condition for positive $\tilde{f}_{X^{(0)}}$ is satisfied, we can numerically find the root of $\tilde{f}_{X^{(0)}}$ to find the minimum of $\tilde{f}$.

Consistency

In this section, we show that the estimates $(\hat{X}^{(0)}, \hat{\beta})$ are consistent. That is, we show that $(\hat{X}^{(0)}, \hat{\beta})$ converge in probability to the true values as the number of outbreaks n goes to ∞. To be consistent with common statistical notation, we denote the true values of $(X^{(0)}, \beta)$ with subscript zeros. Thus they are written $(X_0^{(0)}, \beta_0)$, where the subscript does not indicate type 0 as it would in the main text. We consider the case in which $\beta_0 > 0$, $X_0^{(0)} > 1$, and β_0 and $X_0^{(0)}$ are both finite.

As $n \to \infty$, an outbreak that infects the entire susceptible population will occur almost surely (i.e., with probability 1). In this limit, therefore, $X_0^{(0)}$ is the minimum feasible value for $\hat{X}^{(0)}$ and our estimation procedure will begin by evaluating $\tilde{f}_{X^{(0)}}$ at $X_0^{(0)}$. Our estimates will be consistent if $\tilde{f}_{X^{(0)}}(X_0^{(0)})$ converges to a non-negative value as $n \to \infty$ and $m(X_0^{(0)})$ is a consistent estimator of β_0.

Our expression for m, Equation (A.22), is the maximum-likelihood estimator for the rate parameter of independent exponential random variables that are right-censored. The standard result for the asymptotic consistency of maximum likelihood estimates applies [52].

Next, we will show that $\tilde{f}_{X^{(0)}}(X_0^{(0)})$, Equation (A.24), converges in probability to a positive value. Because $m(X_0^{(0)}) \xrightarrow{p} \beta_0$, application of the contin-

uous mapping theorem yields

$$\tilde{f}_{X^{(0)}}(X_0^{(0)})/n \xrightarrow{p} \beta_0\langle\tau_j\rangle + \langle -\psi(X_0^{(0)}+1) + \psi(X_0^{(0)}-k_j+1)\rangle, \qquad \text{(A.26)}$$

where $\langle\cdot\rangle$ denotes an average over outbreaks. Thus, our claim is equivalent to $\beta_0\langle\tau_j\rangle > \langle\psi(X_0^{(0)}+1) - \psi(X_0^{(0)}-k_j+1)\rangle$.

For notational convenience in the following argument, we introduce new notation here. Denote $\psi(X_0^{(0)}+1) - \psi(X_0^{(0)}-k_j+1) = 1/X_0^{(0)} + 1/(X_0^{(0)}-1) + \cdots + 1/(X_0^{(0)}-k_j+1)$ as $\langle Q_{(k_j)}\rangle$, the expected value of the k_jth order statistic from $X_0^{(0)}$ exponential variables with a rate of unity. For consistency, we let $\langle Q_{(0)}\rangle = 0$.

We introduce additional notation based on the Sellke construction of our outbreak model. Because we are interested in the final state of the model and not the dynamics, we consider the progress of the outbreak in terms of generations to further simplify matters. In the first generation, generation 0, the infective people realize their infectious periods. We define infection pressure of a generation t as $A^{(t)} = \beta_0 \sum_{i\le t} \sum_{0<j\le Y^{(i)}} I^{(i,j)}$, where $I^{(i,j)}$ is the length of the infectious period of the jth infective person in generation i. Susceptible people that have a threshold to infection that is less than the infection pressure will become infective in the next generation and contribute to $A^{(t+1)}$. People are only infective for one generation such that $Y^{(t+1)} + X^{(t+1)} = X^{(t)}$. Thus, $X^{(t+1)} < X^{(t)}$ until $t = T$, where $Y^{(T+1)} = 0$ and $X^{(T)} = X^{(T+1)}$, and we say that generation T is the final generation of the outbreak.

Using our newly introduced notation, the difference between $\beta_0\langle\tau_j\rangle$ and $\langle\psi(X_0^{(0)}+1) - \psi(X_0^{(0)}-k_j+1)\rangle$ can be written as

$$\mathrm{E}(A^{(T)} - \langle Q_{(X_0^{(0)}-X^{(T+1)})}\rangle), \qquad \text{(A.27)}$$

where $\mathrm{E}(\cdot)$ denotes an average over realizations of the model. To be precise,

$$\mathrm{E}(A^{(T)} - \langle Q_{(X_0^{(0)} - X^{(T+1)})} \rangle) = \int_{y=0}^{\infty} \sum_{i=0}^{X^{(0)}} \Pr(A^{(T)} = y, X^{(T+1)} = i)(y - \langle Q_{(X_0^{(0)} - i)} \rangle) \, \mathrm{d}y. \tag{A.28}$$

Now, according to our model

$$\mathrm{E}(A^{(t)} | A^{(t-1)}) = A^{(t-1)} + \int_{y=0}^{\infty} \Pr(\textstyle\sum_{0<j<Y^{(t)}} I^{(t,j)} = y) y \, \mathrm{d}y \tag{A.29}$$

and

$$\mathrm{E}(\langle Q_{(X^{(0)} - X^{(t+1)})} \rangle | X^{(t)}) = \sum_{i=0}^{X^{(t)}} \Pr(Y^{(t+1)} = i) \langle Q_{(X^{(0)} - X^{(t)} + i)} \rangle. \tag{A.30}$$

Subtracting Equation (A.30) from (A.29) yields

$$\mathrm{E}(A^{(t)} - \langle Q_{(X_0^{(0)} - X^{(t+1)})} \rangle | A^{(t-1)}, X^{(t)}) = A^{(t-1)} + \int_{y=0}^{\infty} \sum_{i=0}^{X^{(t)}} \Pr(\textstyle\sum_{0<j\leq Y^{(t)}} I^{(t,j)} = y, Y^{(t+1)} = i) \left(y - \langle Q_{(X_0^{(0)} - X^{(t)} + i)} \rangle \right) \mathrm{d}y. \tag{A.31}$$

Consider the case in which $\sum_{0<j\leq Y^{(t)}} I^{(t,j)} = c/\beta_0$ for some positive constant c. Then the probability that $A^{(t)} - A^{(t-1)}$ exceeds the threshold of i of the remaining susceptibles is

$$\Pr(Y^{(t+1)} = i) = \binom{X^{(t)}}{i} [1 - \exp(-c)]^i [\exp(-c)]^{X^{(t)} - i}. \tag{A.32}$$

From the model definition, we have $A^{(-1)} = 0$. Also, as $c \to 0$, $\Pr(Y^{(t+1)} = 0) \to 1$ for $t \geq 0$. Thus none of the initial susceptibles are ever infected and we have

$$\lim_{c\to 0} \mathrm{E}(A^{(t)} - \langle Q_{(X_0^{(0)} - X^{(t+1)})}\rangle) = 0. \tag{A.33}$$

For any $A^{(t-1)}$ and finite $X^{(t)}$, Equation (A.31) diverges as $c \to \infty$. Thus we have

$$\lim_{c\to \infty} \mathrm{E}(A^{(t)} - \langle Q_{(X_0^{(0)} - X^{(t+1)})}\rangle) = \infty. \tag{A.34}$$

Differentiating Equation (A.31) with respect to c yields

$$\frac{\mathrm{d}}{\mathrm{d}c} \mathrm{E}(A^{(t)} - \langle Q_{(X_0^{(0)} - X^{(t+1)})}\rangle | A^{(t-1)}, X^{(t)}) = (1 - \mathrm{e}^{-c})^{X^{(t)}}. \tag{A.35}$$

Equation (A.35) can be proven by induction. It can also be proven using the binomial theorem as follows. Expanding $(1-\mathrm{e}^{-c})^i$ and letting $a = X^{(t)}+\ell-i$ yields

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}c} \mathrm{E}(A^{(t)} - \langle Q_{(X_0^{(0)} - X^{(t+1)})}\rangle | A^{(t-1)}, X^{(t)}) = \\ \sum_{a=0}^{X^{(t)}} \binom{X^{(t)}}{a} \mathrm{e}^{-ac} \sum_{\ell=0}^{a} \binom{a}{\ell} (-1)^{\ell} (a\langle Q_{(X^{(0)}+\ell-a)}\rangle + 1 - ac),\end{aligned} \tag{A.36}$$

Using the integral representation of $\langle Q_{(k)}\rangle$ implied by Equation (A.12), we find

$$\sum_{\ell=0}^{a} \binom{a}{\ell} (-1)^{\ell} a \langle Q_{(X^{(0)}+\ell-a)}\rangle = (-1)^a.$$

It follows that

$$\frac{\mathrm{d}}{\mathrm{d}c} \mathrm{E}(A^{(t)} - \langle Q_{(X_0^{(0)} - X^{(t+1)})}\rangle | A^{(t-1)}, X^{(t)}) = \sum_{a=0}^{X^{(t)}} \binom{X^{(t)}}{a} (-\mathrm{e}^{-c})^a, \tag{A.37}$$

which is equivalent to Equation (A.35).

Equation (A.35) is positive for all positive c regardless of the values of $X^{(t)}$ and $A^{(t-1)}$. Therefore,

$$\frac{\mathrm{d}}{\mathrm{d}c} \mathrm{E}(A^{(t)} - \langle Q_{(X_0^{(0)} - X^{(t+1)})}\rangle) > 0. \tag{A.38}$$

Equation (A.38) holds when $\sum_{0<j\leq Y^{(t)}} I^{(t,j)}$ is randomly distributed among positive numbers. Equations (A.33), (A.34), and (A.38) together show that $\mathrm{E}(A^{(T)} - \langle Q_{(X_0^{(0)} - X^{(T+1)})}\rangle) > 0$. It follows that $\beta_0\langle\tau_j\rangle > \langle\psi(X_0^{(0)}+1) - \psi(X_0^{(0)} - k_j + 1)\rangle$, $\tilde{f}_{X^{(0)}}(X_0^{(0)})$ converges in probability to a positive value, and our estimates are consistent.

Bibliography

[1] Altmann, M., 1995. Susceptible-infected-removed epidemic models with dynamic partnerships. Journal of Mathematical Biology 33, 661–675.

[2] Anderson, R.M., May, R.M., 1992. Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, USA.

[3] Andersson, H., Britton, T., 2000. Stochastic Epidemic Models and Their Statistical Analysis. Springer. 1 edition.

[4] Auerbach, D.M., Darrow, W.W., Jaffe, H.W., Curran, J.W., 1984. Cluster of cases of the acquired immune deficiency syndrome. Patients linked by sexual contact. The American Journal of Medicine 76, 487–492. PMID: 6608269.

[5] Bansal, S., Grenfell, B.T., Meyers, L.A., 2007. When individual behaviour matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface 4, 879–891.

[6] Barthélemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A., 2004. Velocity and hierarchical spread of epidemic outbreaks in scale-free networks. Physical Review Letters 92, 178701.

[7] Barthélemy, M., Barrat, A., Pastor-Satorras, R., Vespignani, A., 2005. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. Journal of Theoretical Biology 235, 275–288.

[8] Becker, N., 1979. An estimation procedure for household disease data. Biometrika 66, 271–277.

[9] Becker, N.G., 1989. Analysis of Infectious Disease Data. Chapman and Hall/CRC.

[10] Becker, N.G., 1991. Analysis of infectious disease data from a sample of households. Lecture Notes-Monograph Series 18, 27–40.

[11] Bennett, S.N., Drummond, A.J., Kapan, D.D., Suchard, M.A., Muoz-Jordn, J.L., Pybus, O.G., Holmes, E.C., Gubler, D.J., 2010. Epidemic dynamics revealed in dengue evolution. Molecular Biology and Evolution 27, 811–818.

[12] Biek, R., Henderson, J.C., Waller, L.A., Rupprecht, C.E., Real, L.A., 2007. A high-resolution genetic signature of demographic and spatial expansion in epizootic rabies virus. Proceedings of the National Academy of Sciences of the United States of America 104, 7993–7998.

[13] Bolker, B., Skaug, H., 2011. R2admb: ADMB to R interface functions. R package version 0.7.5.1.

[14] Britton, T., Lindholm, M., 2010. Dynamic random networks in dynamic populations. Journal of Statistical Physics 139, 518–535.

[15] Britton, T., Lindholm, M., Turova, T., 2011. A dynamic network in a dynamic population: asymptotic properties. Journal of Applied Probability 48, 1163–1178.

[16] Burnham, K.P., Anderson, D.R., 2004. Multimodel inference. Sociological Methods & Research 33, 261 –304.

[17] Cannon, J.L., Lindesmith, L.C., Donaldson, E.F., Saxe, L., Baric, R.S., Vinjé, J., 2009. Herd immunity to GII.4 noroviruses is supported by outbreak patient sera. J Virol 83, 5363–5374.

[18] Centers for Disease Control and Prevention, 2011. Updated norovirus outbreak management and disease prevention guidelines. MMWR 60, 1–18.

[19] Chan, R.K.W., Tan, H.H., Chio, M.T.W., Sen, P., Ho, K.W., Wong, M.L., 2008. Sexually transmissible infection management practices among primary care physicians in Singapore. Sexual Health 5, 265–271.

[20] Craft, M.E., Volz, E., Packer, C., Meyers, L.A., 2009. Distinguishing epidemic waves from disease spillover in a wildlife population. Proceedings of the Royal Society B 276, 1777–1785.

[21] Csardi, G., Nepusz, T., 2006. The igraph software package for complex network research. InterJournal Complex Systems, 1695.

[22] Donnelly, C.A., Ghani, A.C., Leung, G.M., Hedley, A.J., Fraser, C., Riley, S., Abu-Raddad, L.J., Ho, L.M., Thach, T.Q., Chau, P., Chan, K.P., Lam, T.H., Tse, L.Y., Tsang, T., Liu, S.H., Kong, J.H.B., Lau, E.M.C., Ferguson, N.M., Anderson, R.M., 2003. Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. Lancet 361, 1761–1766.

[23] Drummond, A.J., Rambaut, A., 2007. BEAST: Bayesian evolutionary analysis by sampling trees. BioMed Central Evolutionary Biology 7, 214.

[24] Drummond, A.J., Rambaut, A., Shapiro, B., Pybus, O.G., 2005. Bayesian coalescent inference of past population dynamics from molecular sequences. Molecular Biology and Evolution 22, 1185–1192.

[25] Edwards, C.T.T., Holmes, E.C., Wilson, D.J., Viscidi, R.P., Abrams, E.J., Phillips, R.E., Drummond, A.J., 2006. Population genetic estimation of the loss of genetic diversity during horizontal transmission of HIV-1. BioMed Central Evolutionary Biology 6, 28.

[26] Erdős, P., Rényi, A., 1959. On random graphs. I. Publicationes Mathematicae 6, 290–297.

[27] Eubank, S., 2005. Network based models of infectious disease spread. Japanese Journal of Infectious Diseases 58, S9–13.

[28] Evans, M.R., Meldrum, R., Lane, W., Gardner, D., Ribeiro, C.D., Gallimore, C.I., Westmoreland, D., 2002. An outbreak of viral gastroenteritis following environmental contamination at a concert hall. Epidemiol and Infect 129, 355–360.

[29] Fahrmeir, L., 1985. Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models. The Annals of Statistics 13, 342–368.

[30] Felsenstein, J., 2003. Inferring Phylogenies. Sinauer Associates. 2 edition.

[31] Fenner, F., Henderson, D.A., Arita, I., Jeẑek, Z., Ladnyi, I.D., 1988. Smallpox and its eradication. volume 6 of *History of International Public Health*. World Health Organization, Geneva, Switzerland.

[32] Fournier, D.A., Skaug, H.J., Ancheta, J., Ianelli, J., Magnusson, A., Maunder, M.N., Nielsen, A., Sibert, J., 2011. AD Model Builder: using automatic differentiation for statistical inference of highly parameterized complex nonlinear models. Optimization Methods and Software 27, 233–249.

[33] Fraser, C., Donnelly, C.A., Cauchemez, S., Hanage, W.P., Kerkhove, M.D.V., Hollingsworth, T.D., Griffin, J., Baggaley, R.F., Jenkins, H.E., Lyons, E.J., Jombart, T., Hinsley, W.R., Grassly, N.C., Balloux, F., Ghani, A.C., Ferguson, N.M., Rambaut, A., Pybus, O.G., Lopez-Gatell, H., Alpuche-Aranda, C.M., Chapela, I.B., Zavala, E.P., Guevara, D.M.E., Checchi, F., Garcia, E., Hugonnet, S., Roth, C., Collaboration, W.H.O.R.P.A., 2009. Pandemic potential of a strain of influenza A (H1N1): early findings. Science 324, 1557–1561.

[34] Frost, S.D.W., Volz, E.M., 2010. Viral phylodynamics and the search for an 'effective number of infections'. Philosophical Transactions of the Royal Society B 365, 1879–1890.

[35] Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M., Rossi, F., 2009. GNU Scientific Library Reference Manual. Network Theory Ltd.. third edition.

[36] Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. Journal of Computational Physics 22, 403–404.

[37] Gillespie, D.T., 2007. Stochastic simulation of chemical kinetics. Annual Review of Physical Chemistry 58, 35–55.

[38] Glass, R.I., Parashar, U.D., Estes, M.K., 2009. Norovirus gastroenteritis. N Engl J Med 361, 1776–1785.

[39] Grenfell, B.T., Pybus, O.G., Gog, J.R., Wood, J.L.N., Daly, J.M., Mumford, J.A., Holmes, E.C., 2004. Unifying the epidemiological and evolutionary dynamics of pathogens. Science 303, 327–32.

[40] Hagstrom, G.I., Hang, D.H., Ofria, C., Torng, E., 2004. Using Avida to test the effects of natural selection on phylogenetic reconstruction methods. Artificial Life 10, 157–166.

[41] Halloran, M.E., Longini, I.M., Struchiner, C.J., 2009. Design and Analysis of Vaccine Studies. Springer.

[42] Hayakawa, Y., O'Neill, P.D., Upton, D., Yip, P.S., 2003. Bayesian inference for a stochastic epidemic model with uncertain numbers of susceptibles of several types. Aust N Z J Stat 45, 491–502.

[43] Heal, C., Muller, R., 2008. General practitioners' knowledge and attitudes to contact tracing for genital chlamydia trachomatis infection in North Queensland. Australian and New Zealand Journal of Public Health 32, 364–366.

[44] Heijne, J.C.M., Teunis, P., Morroy, G., Wijkmans, C., Oostveen, S., Duizer, E., Kretzschmar, M., Wallinga, J., 2009. Enhanced hygiene measures and norovirus transmission during an outbreak. Emerg Infect Dis 15, 24–30.

[45] Hens, N., Goeyvaerts, N., Aerts, M., Shkedy, Z., Damme, P.V., Beutels, P., 2009. Mining social mixing patterns for infectious disease models

based on a two-day population survey in Belgium. BMC Infectious Diseases 9, 5.

[46] Höhle, M., 2009. Additive-multiplicative regression models for spatio-temporal epidemics. Biom J 51, 961–978.

[47] Hohle, M., Jorgensen, E., O'Neill, P.D., 2005. Inference in disease transmission experiments by using stochastic epidemic models. Journal of the Royal Statistical Society: Series C (Applied Statistics) 54, 349–366.

[48] Holmes, E.C., Grenfell, B.T., 2009. Discovering the phylodynamics of RNA viruses. PLoS Computational Biology 5, e1000505.

[49] Huggins, R.M., Yip, P.S.F., Lau, E.H.Y., 2004. A note on the estimation of the initial number of susceptible individuals in the general epidemic model. Stat Probab Lett 67, 321–330.

[50] Hughes, G.J., Fearnhill, E., Dunn, D., Lycett, S.J., Rambaut, A., Brown, A.J.L., Collaboration, U.K.H.D.R., 2009. Molecular phylodynamics of the heterosexual HIV epidemic in the United Kingdom. PLoS Pathogens 5, e1000590.

[51] Hutcheon, J.A., Chiolero, A., Hanley, J.A., 2010. Random measurement error and regression dilution bias. BMJ 340, c2289–c2289.

[52] Kalbfleisch, J.D., Prentice, R.L., 2002. The Statistical Analysis of Failure Time Data. Wiley. 2 edition.

[53] Keeling, M.J., 2005. Models of foot-and-mouth disease. Proc Biol Sci 272, 1195–1202.

[54] Kingman, J.F.C., 1982. On the genealogy of large populations. Journal of Applied Probability 19, 27–43.

[55] Klovdahl, A.S., 1985. Social networks and the spread of infectious diseases: The AIDS example. Soc Sci Med 21, 1203–1216.

[56] Kypraios, T., 2009. A note on maximum likelihood estimation of the initial number of susceptibles in the general stochastic epidemic model. Stat Probab Lett 79, 1972–1976.

[57] Lau, E.H.Y., Yip, P.S.F., 2008. Estimating the basic reproductive number in the general epidemic model with an unknown initial number of susceptible individuals. Scandinavian Journal of Statistics 35, 650–663.

[58] L'Ecuyer, P., Simard, R., Chen, E., Kelton, W., 2002. An object-oriented random-number package with many long streams and substreams. Operations Research 50, 1073–5.

[59] Lemey, P., Salemi, M., Vandamme, A.M. (Eds.), 2009. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge University Press. 2nd edition.

[60] Liljeros, F., Edling, C.R., Amaral, L.A.N., 2003. Sexual networks: implications for the transmission of sexually transmitted infections. Microbes and Infection 5, 189–196.

[61] Lively, C.M., 2010. An epidemiological model of host-parasite coevolution and sex. J Evol Biol 23, 1490–1497.

[62] Lopman, B., Armstrong, B., Atchison, C., Gray, J.J., 2009. Host, weather and virological factors drive norovirus epidemiology: time-series analysis

of laboratory surveillance data in England and Wales. PLoS One 4, e6671.

[63] Lopman, B.A., Hall, A.J., Curns, A.T., Parashar, U.D., 2011. Increasing rates of gastroenteritis hospital discharges in US adults and the contribution of norovirus, 1996-2007. Clin Infect Dis 52, 466–474.

[64] Lopman, B.A., Reacher, M.H., Vipond, I.B., Hill, D., Perry, C., Halladay, T., Brown, D.W., Edmunds, W.J., Sarangi, J., 2004a. Epidemiology and cost of nosocomial gastroenteritis, Avon, England, 2002-2003. Emerg Infect Dis 10, 1827–1834.

[65] Lopman, B.A., Reacher, M.H., Vipond, I.B., Sarangi, J., Brown, D.W.G., 2004b. Clinical manifestation of norovirus gastroenteritis in health care settings. Clin Infect Dis 39, 318–324.

[66] Lu, X., Bengtsson, L., Britton, T., Camitz, M., Kim, B.J., Thorson, A., Liljeros, F., 2012. The sensitivity of respondent-driven sampling. Journal of the Royal Statistical Society: Series A (Statistics in Society) 175, 191216.

[67] McCarthy, M., Haddow, L.J., Furner, V., Mindel, A., 2007. Contact tracing for sexually transmitted infections in New South Wales, Australia. Sexual Health 4, 21–25.

[68] McCreesh, N., Frost, S.D.W., Seeley, J., Katongole, J., Tarsh, M.N., Ndunguse, R., Jichi, F., Lunel, N.L., Maher, D., Johnston, L.G., Sonnenberg, P., Copas, A.J., Hayes, R.J., White, R.G., 2012. Evaluation of respondent-driven sampling. Epidemiology (Cambridge, Mass.) 23, 138–147. PMID: 22157309.

[69] Meyers, L.A., Pourbohloul, B., Newman, M.E.J., Skowronski, D.M., Brunham, R.C., 2005. Network theory and SARS: predicting outbreak diversity. Journal of Theoretical Biology 232, 71–81.

[70] Miller, J.C., 2009. Spread of infectious disease through clustered populations. Journal of the Royal Society Interface 6, 1121–1134.

[71] Minin, V.N., Bloomquist, E.W., Suchard, M.A., 2008. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. Molecular Biology and Evolution 25, 1459–1471.

[72] Morris, M., Kretzschmar, M., 2000. A microsimulation study of the effect of concurrent partnerships on the spread of HIV in Uganda. Mathematical Population Studies: An International Journal of Mathematical Demography 8, 109–133.

[73] Mossong, J., Hens, N., Jit, M., Beutels, P., Auranen, K., Mikolajczyk, R., Massari, M., Salmaso, S., Tomba, G.S., Wallinga, J., Heijne, J., Sadkowska-Todys, M., Rosinska, M., Edmunds, W.J., 2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Medicine 5, e74.

[74] Nakano, T., Lu, L., He, Y., Fu, Y., Robertson, B.H., Pybus, O.G., 2006. Population genetic history of hepatitis C virus 1b infection in China. Journal of General Virology 87, 73–82.

[75] Newman, M.E.J., 2002. Assortative mixing in networks. Physical Review Letters 89, 208701.

[76] O'Fallon, B.D., Seger, J., Adler, F.R., 2010. A continuous-state coalescent and the impact of weak selection on the structure of gene genealogies. Molecular Biology and Evolution 27, 1162–1172.

[77] O'Neill, P.D., Marks, P.J., 2005. Bayesian model choice and infection route modelling in an outbreak of Norovirus. Stat Med 24, 2011–2024.

[78] Potterat, J.J., Phillips-Plummer, L., Muth, S.Q., Rothenberg, R.B., Woodhouse, D.E., Maldonado-Long, T.S., Zimmerman, H.P., Muth, J.B., 2002. Risk network structure in the early epidemic phase of HIV transmission in Colorado Springs. Sexually Transmitted Infections 78 Suppl 1, i159–i163.

[79] R Development Core Team, 2010. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-900051-07-0.

[80] Rambaut, A., Drummond, A.J., 2009. Tracer v1.5. Available from http://beast.bio.ed.ac.uk/Tracer.

[81] Rambaut, A., Holmes, E.C., 2009. The early molecular epidemiology of the swine-origin A/H1N1 human influenza pandemic. PLoS Currents Influenza , RRN1003.

[82] Rida, W.N., 1991. Asymptotic properties of some estimators for the infection rate in the general stochastic epidemic model. Journal of the Royal Statistical Society. Series B (Methodological) 53, 269–283.

[83] Rosenthal, N.A., Lee, L.E., Vermeulen, B.A.J., Hedberg, K., Keene, W.E., Widdowson, M., Cieslak, P.R., Vinjé, J., 2011. Epidemiologi-

cal and genetic characteristics of norovirus outbreaks in long-term care facilities, 2003-2006. Epidemiol and Infect 139, 286–294.

[84] Rothenberg, R., Dan My Hoang, T., Muth, S.Q., Crosby, R., 2007. The Atlanta Urban Adolescent Network Study: a network view of STD prevalence. Sexually Transmitted Diseases 34, 525–531. PMID: 17297380.

[85] Rothenberg, R.B., Long, D.M., Sterk, C.E., Pach, A., Potterat, J.J., Muth, S., Baldwin, J.A., Trotter, R.T., 2000. The Atlanta Urban Networks Study: a blueprint for endemic transmission. AIDS 14, 2191–2200.

[86] Rothenberg, R.B., McElroy, P.D., Wilce, M.A., Muth, S.Q., 2003. Contact tracing: comparing the approaches for sexually transmitted diseases and tuberculosis. International Journal of Tuberculosis and Lung Disease 7, S342–S348.

[87] Ruan, Y.J., Wei, C.L., Ee, A.L., Vega, V.B., Thoreau, H., Su, S.T.Y., Chia, J., Ng, P., Chiu, K.P., Lim, L., Zhang, T., Peng, C.K., Lin, E.O.L., Lee, N.M., Yee, S.L., Ng, L.F.P., Chee, R.E., Stanton, L.W., Long, P.M., Liu, E.T., 2003. Comparative full-length genome sequence analysis of 14 SARS coronavirus isolates and common mutations associated with putative origins of infection. Lancet 361, 1779–1785.

[88] Salkeld, D.J., Salathé, M., Stapp, P., Jones, J.H., 2010. Plague outbreaks in prairie dog populations explained by percolation thresholds of alternate host abundance. Proceedings of the National Academy of Sciences of the United States of America 107, 14247–14250.

[89] Sander, L.M., Warren, C., Sokolov, I., Simon, C., Koopman, J., 2002.

Percolation on heterogeneous networks as a model for epidemics. Mathematical Biosciences 180, 293–305.

[90] Scallan, E., Hoekstra, R.M., Angulo, F.J., Tauxe, R.V., Widdowson, M., Roy, S.L., Jones, J.L., Griffin, P.M., 2011. Foodborne illness acquired in the United States—major pathogens. Emerg Infect Dis 17, 7–15.

[91] Sellke, T., 1983. On the asymptotic distribution of the size of a stochastic epidemic. J App Prob 20, 390–394.

[92] Shao, Q.X., 1999. Some properties of an estimator for the basic reproduction number of the general epidemic model. Mathematical Biosciences 159, 79–96.

[93] Siebenga, J.J., Lemey, P., Pond, S.L.K., Rambaut, A., Vennema, H., Koopmans, M., 2010. Phylodynamic reconstruction reveals norovirus GII.4 epidemic expansions and their molecular determinants. PLoS Pathogens 6, e1000884.

[94] Smith, G.J.D., Vijaykrishna, D., Bahl, J., Lycett, S.J., Worobey, M., Pybus, O.G., Ma, S.K., Cheung, C.L., Raghwani, J., Bhatt, S., Peiris, J.S.M., Guan, Y., Rambaut, A., 2009. Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza A epidemic. Nature 459, 1122–1125.

[95] St. Lawrence, J.S., Montaño, D.E., Kasprzyk, D., Phillips, W.R., Armstrong, K., Leichliter, J.S., 2002. STD screening, testing, case reporting, and clinical and partner notification practices: a national survey of US physicians. American Journal of Public Health 92, 1784–1788.

[96] Stack, J.C., Welch, J.D., Ferrari, M.J., Shapiro, B.U., Grenfell, B.T., 2010. Protocols for sampling viral sequences to study epidemic dynamics. Journal of the Royal Society Interface .

[97] Stokes, T., Schober, P., 1999. A survey of contact tracing practice for sexually transmitted diseases in GUM clinics in England and Wales. International Journal of STD & AIDS 10, 17–21.

[98] Sukhrie, F.H.A., Beersma, M.F.C., Wong, A., van der Veer, B., Vennema, H., Bogerman, J., Koopmans, M., 2011. Using molecular epidemiology to trace transmission of nosocomial norovirus infection. J Clin Microbiol 49, 602–606.

[99] Thornley, C.N., Emslie, N.A., Sprott, T.W., Greening, G.E., Rapana, J.P., 2011. Recurring norovirus transmission on an airplane. Clin Infect Dis 53, 515–520.

[100] Tony Vignaux, Klaus Muller, and Bob Helmbold, 2012. SimPy Manual. Available at http://simpy.sourceforge.net.

[101] UNAIDS, 2012. Global report: UNAIDS report on the global AIDS epidemic 2012. Joint United Nations Programme on HIV/AIDS (UNAIDS).

[102] Viger, F., Latapy, M., 2005. Computing and Combinatorics. Springer Berlin / Heidelberg. chapter Efficient and Simple Generation of Random Simple Connected Graphs with Prescribed Degree Sequence. pp. 440–449.

[103] Volz, E., 2008. SIR dynamics in random networks with heterogeneous connectivity. Journal of Mathematical Biology 56, 293–310.

[104] Volz, E., Heckathorn, D.D., 2008. Probability based estimation theory for respondent driven sampling. Journal of Official Statistics 24, 79–97.

[105] Volz, E., Meyers, L.A., 2007. Susceptible-infected-recovered epidemics in dynamic contact networks. Proceedings of the Royal Society B 274, 2925–2933.

[106] Volz, E., Meyers, L.A., 2009. Epidemic thresholds in dynamic contact networks. Journal of the Royal Society Interface 6, 233–241.

[107] Volz, E.M., Pond, S.L.K., Ward, M.J., Brown, A.J.L., Frost, S.D.W., 2009. Phylodynamics of infectious disease epidemics. Genetics 183, 1421–1430.

[108] Wakeley, J., Sargsyan, O., 2009. Extensions of the coalescent effective population size. Genetics 181, 341–345.

[109] Wallinga, J., Teunis, P., Kretzschmar, M., 2006. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. American Journal of Epidemiology 164, 936–944.

[110] Welch, D., Nicholls, G.K., Rodrigo, A., Solomon, W., 2005. Integrating genealogy and epidemiology: the ancestral infection and selection graph as a model for reconstructing host virus histories. Theoretical Population Biology 68, 65–75.

[111] Wickham, H., 2009. ggplot2: elegant graphics for data analysis. Springer New York.

[112] Wikswo, M.E., Cortes, J., Hall, A.J., Vaughan, G., Howard, C., Gregoricus, N., Cramer, E.H., 2011. Disease transmission and passenger behaviors during a high morbidity Norovirus outbreak on a cruise ship, January 2009. Clin Infect Dis 52, 1116–1122.

[113] William, D.C., 1979. Sexually transmitted diseases in gay men: an insider's view. Sexually Transmitted Diseases 6, 278–280. PMID: 583369.

[114] Woodhouse, D.E., Rothenberg, R.B., Potterat, J.J., Darrow, W.W., Muth, S.Q., Klovdahl, A.S., Zimmerman, H.P., Rogers, H.L., Maldonado, T.S., Muth, J.B., 1994. Mapping a social network of heterosexuals at high risk for HIV infection. AIDS (London, England) 8, 1331–1336. PMID: 7802989.

[115] Yang, Z., 2006. Computational Molecular Evolution (Oxford Series in Ecology and Evolution). Oxford University Press, USA.

[116] Zelner, J.L., King, A.A., Moe, C.L., Eisenberg, J.N.S., 2010. How infections propagate after point-source outbreaks: an analysis of secondary norovirus transmission. Epidemiology 21, 711–718.