Integrating Social and Biological Processes of Infectious Disease Transmission at Three Levels: Household, Community and Region

by

Jonathan L. Zelner

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Public Policy and Sociology)
in The University of Michigan
2011

Doctoral Committee:

Associate Professor Joseph N. Eisenberg, Co-Chair
Professor James S. House, Co-Chair
Professor Ana V. Diez-Roux
Professor Jeffrey D. Morenoff
Assistant Professor Elizabeth E. Bruch

ACKNOWLEDGMENTS

Thanks to Sarah, for a thousand productive conversations, a happy home, and so very much more. Thanks to Dad for the understanding that the road to finished projects is paved by endless tinkering. Thanks to Judy, for her support and patience. Thanks to Mom for inspiration and strongheadedness. Thanks to Dan Katz and Mike Bommarito for their friendship, a million ideas, endless conversations and technical support galore. Thanks to Jessica Wyse for her friendship and support from the beginning to the end of graduate school. Thanks to Manze, for making it fun to take my work home with me.

Thanks to Judy Porter for planting a seed that has turned into a career. Thanks to Rick Riolo for teaching me that most programming bugs occur somewhere between the screen and the chair. Thanks to Aaron King, Carl Simon & Scott Page for much help and advice through all of graduate school. Thanks to Howard Oishi & Mita Gibson for helping keep it all together. Thanks to the members of my committee: Elizabeth Bruch, Ana Diez-Roux and Jeff Morenoff for their help and advice through the whole dissertation process. Thanks to Jim House for his advice, ideas and encouraging me to explore.

Finally, thanks to Joe Eisenberg, for endless hours of help and encouragement over many hundreds of slices of Silvio's pizza.

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE

LIST OF TABLES

TABLE

LIST OF APPENDICES

APPENDIX

Chapter 1

Introduction

Social processes often underlie the transmission of infectious diseases, from the micro-level structure of direct contacts to the macro, community-level networks that impact the maintenance of community water sanitation infrastructure and public health services. Despite their importance, these social processes are often excluded from models of infectious disease transmission. For example, some models assume completely homogeneous contact patterns across individuals and age groups (2) whereas others aggregate behavior into a single construct with biological processes, making the distinct influence of social structure and behavior on disease dynamics unidentifiable e.g. (3).

The social epidemiology of infectious diseases

In a recent paper, Cohen et al. (4) assessed the evolution of social epidemiology's infectious disease subfield through a systematic review of academic articles matching the keyword 'social epidemiology' published between 1966 to 2005. The authors found that the number of studies with an infectious disease outcome in social epidemiology published each year increased at a much slower rate than in other fields, such as cancer, cardiovascular disease and psychiatry. Furthermore, most of the increase in articles about the social epidemiology of infectious diseases focused on sexually transmitted diseases (STDs) and HIV and were therefore not necessarily applicable to a broad range of

pathogens and modes of transmission. One explanation for this strong focus on STDs and HIV is that the social and behavioral risk factors associated with infection fit well with definitions of direct contact, e.g., sexual contact and needle sharing. Modes of transmission for other types of pathogens (such as influenza, those that cause gastrointestinal disease, and vector borne illnesses), may appear less amenable to analyses that incorporate social factors. Discordant understandings about what epidemiologist think constitutes a contact and what sociologists think of as a social relationship may have created a longstanding perception that the broader universe of communicable diseases is not amenable to analysis that accounts for social factors.

The sociology of contact, exposure and infection

The social epidemiological literature on infectious diseases shows that a sociological perspective enhances understanding of infectious disease outcomes. However, this literature lacks a framework in which social factors are understood as integral, rather than incidental, to the transmission of many pathogens. Such a framework requires a structured mode for integrating social factors, such as socioeconomic status and individual behavior with the population ecology of infectious diseases. Although existing studies in social epidemiology provide compelling and instructive explanations for some cases of some infectious diseases, studies have not typically shown how social factors impact systemic, population-wide risk. Here, I review some of the existing studies in the social epidemiological literature on infectious diseases, drawing attention to how social factors at the individual, community and macro-level have previously been discussed.

Authors such as Dubos (5) recognized the importance of social stratification and inequality in conditioning individual *susceptibility* to pathogens, such as tuberculosis (TB). As one example of social inequality's impact on individual susceptibility to pathogens, Dubos cited the case of a lacquer sprayer whose job obligated him to work with harsh chemicals and inhale abrasive particles in damp, poorly ventilated spaces. Dubos argued that social factors, as such socioeconomic status, race, and education mediated exposure to these environmental conditions, and therefore increased the lacquer sprayer's susceptibility to infection upon exposure to the TB bacillus.

Additional work on TB by Klovdahl et al (6) demonstrated the role that social networks and geographic place play in explaining patterns of *exposure* to TB that eventually lead to infection. The authors showed that several TB case clusters that initially appeared to be separate events were actually part of a single larger outbreak in which TB was transmitted via the environment in a number of locations in the community, including a number of gay bars. Public health practice regarding TB contact tracing had previously focused on direct transmission and close contact, particularly in homes and schools. By following the social linkages between cases, Klovdahl et al. were able to illuminate the previously underappreciated route of passive, environmental exposure to TB in public places.

Finnaly, Wallace & Wallace's work on HIV and TB in New York City demonstrates the impact of social relationships and social inequality on risk for these illnesses (7). The authors conducted a series of studies that comprise an extensive investigation of linked HIV and TB epidemics in New York City, e.g (8, 9). In their book, *A Plague on Your Houses* (7), Wallace & Wallace explained the rise of HIV and

TB infection in New York City in the 1980s and 1990s as a partial consequence of the city's approach to housing policy and budget management. During this time, the city followed a policy termed 'benign neglect,' in which the city, citing severe budget constraints, re-deployed fire units from poorer neighborhoods to wealthier ones. The authors argued that this effectively consigned a large number of affordable housing units, particularly in the borough of the Bronx, to destruction by fire. Wallace & Wallace contend that the subsequent loss of affordable housing caused significant within-city migration and doubling-up with friends and relatives, resulting in overcrowded housing conditions and a rise in TB incidence. They also claimed that this migration resulted in the destruction of social and financial support networks, leaving already at-risk individuals even more likely to engage in practices, such as unsafe sex and injection drug use, that increased their risk for HIV, hepatitis, and other infections. The authors also contended that this social instability decreased compliance with TB therapy, which is known to accelerate the emergence of multi-drug resistant strains, worsening the TB epidemic and making future cases harder to control.

Wallace & Wallace's study contributes to understanding of the role of social networks in the spread of infectious diseases through its attention to the idea of socio-biological *syndemics*, or interacting, related epidemics. In the case of housing destruction in the Bronx, the authors demonstrated that the destruction of housing constituted a social epidemic in which apartment buildings adjacent to burned-out housing became more susceptible to fire. This accelerating housing destruction and the disruption of social networks interacted with the biological processes of TB and HIV transmission to amplify the severity of existing epidemics.

Wallace & Wallace also showed that social processes impact the meaning and function of social network linkages on two levels: First, transmission probabilities upon exposure are mediated by social conditions, such as living in adequate versus overcrowded housing. Second, community-level social networks may reflect organization and cohesion in ways that can buffer their impact on person-to-person transmission, by fostering or distrupting, e.g. safe sex practices and harm-reduction practices among injection drug users (10, 11). Finally, Wallace and Wallace brought these processes together in a unified conceptual model that both explains variation in HIV risk across social groups while also showing how these processes impact city-wide and regional risk. It is this conceptual integration that makes their work compelling.

Modeling contact processes

The aforementioned literature demonstrates that social conditions and social factors at the levels of individual, community, and society impact exposure to pathogens and transmission of infectious diseases. Although this body of research makes clear that social factors and conditions can impact understanding of infectious disease patterns and outbreaks, theoretical and empirical models of infectious disease risk including these factors remain underdeveloped. In this section, I highlight the ability of current modeling techniques and approaches to incorporate social factors and conditions and point to methodological gaps in network modeling approaches.

At present, most models of disease transmission attend to the impact of social relationships that correspond to contacts that allow for disease transmission, e.g., in workplaces, schools etc. This approach, where types of social relationships are mapped to

a specific intensity of contact, has spawned a network-oriented subfield of infectious disease epidemiology. The strength of the network-oriented approach to understanding disease transmission is its ability to represent the concrete contacts, and fine-grained individual-level variability in contact, rather than relying on assumptions of homogeneous mixing when such assumptions are untenable. The network-oriented approach is also able to incorporate heterogeneities, in host biology as well as behavior, across individuals that impact transmission dynamics. Network epidemiology has been particularly effective for understanding the social epidemiology of sexually transmitted diseases, where the correspondence between sexual behavior, relationships, and contact may be more intuitive than for, e.g. diarrheal disease. Using this approach, the population can also be partitioned into risk groups on the basis of rates of engaging in different risk and protective practices, which allows for more nuanced understanding of infectious disease outbreak dynamics.

HIV research exemplifies the positive impact of a socially and behaviorally minded approach to modeling approach in capturing the nuanced the relationship between behavioral and biological mechanisms can be for understanding transmission dynamics and risk. Koopman et al. (12) argued that being in a high-risk group for HIV infection is often an episodic condition that individuals cycle in and out of, rather than a chronic state in which individuals permanently reside. This is a subtle but important contrast to the core-periphery notion implied by many studies of sexual contact networks (13) that imply fixed contacts over the duration of an outbreak. Koopman et al. show that the period immediately after infection with HIV and before seroconversion, when routine testing does not detect antibodies, is likely to co-occur with these periods of intense high-

risk activity. This temporal correlation between individual immune response and social behavior implies that testing-based interventions alone are unlikely to be effective in reducing the number of new HIV infections.

At first, other types of pathogens appear to be less amenable to an analysis in which contacts correspond to concrete relationships or practices. Take, for instance, the gastrointestinal pathogen norovirus (NoV), which is the focus of the first and second papers in this dissertation. Transmission of gastrointestinal pathogens, which typically occurs via the fecal-oral route, occurs in a variety of contexts – homes, schools, restaurants – in which the likelihood of transmission from an infected individual to a susceptible one is partially dependent on the nature of the social relationship between the individuals. Unlike individuals within sexual networks, individuals involved in NoV transmission are more likely to be exposed via passive modes of contact (14), such as transmission via a shared environment rather than direct contact. Nonetheless, as Klovdahl et al. (6) showed, social factors shape even these passive contacts and cannot help but have important ramifications for the dynamics of infectious disease outbreaks. Consequently, the models we use to understand risk from pathogens, such as NoV need to incorporate social relational dynamics in order to develop a meaningful understanding of risk.

This is the overarching goal of the three papers in this dissertation. In the first paper, we address some of the methodological challenges involved in estimating household transmission rates for NoV using partially observed outbreak data. In the second paper, we integrate these findings about household transmission dynamics into a broader, community-level model of transmission that accounts for the roles of both social

and biological heterogeneity in transmission. We operationalize heterogeneity in behavior using data on variation in time-use by age (15) and individual-level variation in the propensity of individuals to have relationships (16). In the third paper, we integrate ideas about transmission via direct contacts and the social mediation of transmission via the environment to understand gastrointestinal illness risk in 19 villages in rural, northern coastal Ecuador.

Modeling population-level susceptibility

Although the network-based approach to infectious disease outbreaks has primarily focused on individuals, the study of point-source outbreaks involving the exposure of many individuals to a pathogen at one time is also particularly amenable to a structural, sociological perspective. For example, work by Moehle-Boetani et al. (17) illustrated a context in which socioeconomic status may mediate the risk of point-source outbreaks. The authors found that the best predictor of *Shigella Sonnei* transmission, a gastrointestinal pathogen, within day care centers is the presence of food-handlers who also change diapers. This dual role more likely to occur in smaller daycare centers, often within homes and outside of the view of public health authorities. To the extent that such conditions are the result of policies that make regulated day care unaffordable, this transmission may find its ultimate cause in the networks of social and political relationships governing such policy, rather than exposure to infectious individuals and food-handling practices.

From this perspective, the role of social relationships relates to the effect they have on factors that impact the susceptibility of entire populations, such as exposure via

shared environments and infrastructural sources, such as water supply. In the third paper, we show how social factors relating to social cohesion, e.g. attachment to local social networks and participation in community organizations, are upstream risk factors for diarrheal illness in a group of villages in rural Ecuador. Because the social organization of communities is critical to their ability to build and maintain water sanitation infrastructure and promote effective hygiene practices, the presence of pathogens is by no means a sufficient condition for observing outbreaks. This is analogous to Dubos' example of the lacquer sprayer who contracts TB. While exposure to the TB bacillus is a necessary condition for infection, it is not sufficient, because not all exposures result in disease. In both examples, social structure mediates susceptibility in a broad sense. In the case of the lacquer sprayer, this susceptibility is at the individual level of likelihood of infection upon exposure. At the community level, susceptibility can be conceptualized in terms of the likelihood of population-level exposure.

***

Taken together, the papers in this dissertation represent an attempt to integrate social factors that operate at the individual, community, and regional levels with models of infectious disease risk. It is important to attend to the space between sociology to infectious disease epidemiology because the respective areas have much to contribute to each other, but have historically had minimal interaction. It is my sincere hope that this thesis represents a meaningful step in this direction.

References

1. Atmar R, Estes M. The epidemiologic and clinical importance of norovirus infection. Gastroenterology Clinics of North America 2006;35:275-90.
2. Anderson R, May R. Infectious Diseases of Humans: Dynamics and Control. 1992.
3. Lloyd-Smith J, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.
4. Cohen JM, Wilson ML, Aiello A. Analysis of social epidemiology research on infectious diseases: historical patterns and future opportunities. Journal of Epidemiology and Community Health 2007;61:1021-7.
5. Dubos R. Mirage of Health. 1959.
6. Klovdahl AS, Graviss EA, Yaganehdoost A, et al. Networks and tuberculosis: an undetected community outbreak involving public places. Social Science and Medicine 2001;52:681-94.
7. Wallace D, Wallace R. A Plague on Your Houses: How New York Was Burned Down and National Public Health Crumbled. 1998.
8. Wallace R. A Synergism of Plagues: "Planned Shrinkage," Contagious Housing Destruction, and AIDS in the Bronx. Environmental Research 1988;47:1-33.
9. Wallace R. AIDS in the HAART era: New York's heterogeneous geography. Social Science & Medicine 2003;56:1155-71.
10. Pilowsky DJ, Hoover D, Hadden B, et al. Impact of social network characteristics on high-risk sexual behaviors among non-injection drug users. Subst Use Misuse 2007;42:1629-49.

11. De P, Cox J, Boivin J-F, et al. The importance of social networks in their association to drug equipment sharing among injection drug users: a review. Addiction 2007;102:1730-9.

12. Koopman JS, Jacquez J, Welch G, et al. The role of early HIV infection in the spread of HIV through populations. JAIDS 1997;14:249-58.

13. Borgatti S, Everett M. Models of core/periphery structures. Social Networks 2000.

14. Li S, Eisenberg JNS, Spicknall IH, et al. Dynamics and control of infections transmitted from person to person through the environment. American Journal of Epidemiology 2009;170:257-65.

15. Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. Plos Med 2008.

16. Bansal S, Grenfell BT, Meyers LA. When individual behavior matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface 2008;4:879-91.

17. Mohle-Boetani JC, Stapleton M, Finger R, et al. Communitywide Shigellosis: Control of an Outbreak and Risk Factors in Child Day-Care Centers. American Journal of Public Health 1995;85:812-6.

Chapter 2

How infections propagate after point-source outbreaks: An analysis of secondary Norovirus transmission

Abstract

Background: Secondary transmission after point-source outbreaks is an integral feature of the epidemiology of gastrointestinal pathogens such as norovirus. The household is an important site of these secondary cases. It can become the source of further community transmission as well as new point-source outbreaks. Consequently, time-series data from exposed households provide information for risk assessment and intervention.

Methods: Analysis of these data requires models that can address: 1) dependencies in infection transmission, 2) random variability resulting from households with few members, and 3) unobserved state variables important to transmission. We use Monte Carlo maximum likelihood via data augmentation for obtaining estimates of the transmission rate and infectious period from household outbreaks with the three above features.

Results: We apply this parameter estimation technique to 153 infection sequences within households from a norovirus outbreak in Sweden and obtain maximum likelihood estimates of the daily rate of transmission ($\hat{\beta}$ = 0.14, 95% confidence interval [CI] =

0.08–0.24) and average infectious period ($1/\hat{\gamma}$ = 1.17 days, 95% CI =1.00–1.88). We also demonstrate the robustness of the estimates to missing household sizes and asymptomatic infections.

Conclusions: Maximum likelihood techniques such as these can be used to estimate transmission parameters under conditions of unobserved states and missing household size data, and to aid in the understanding of secondary risks associated with point-source outbreaks.

Introduction

Norovirus is a highly-infectious gastrointestinal pathogen that affects all age groups (1). Investigations of primary point-source outbreaks, therefore, often focus on secondary cases (2,3). Households constitute a particularly important site of these secondary cases, as living in close proximity facilitates a higher effective rate of contact, particularly for diseases where the fecal-oral route is important to transmission. This household transmission contributes to overall disease burden, and individuals infected at the household level may generate infections in the community that result in new point-source outbreaks that infect many people at one time.

From 1997 through 2002, norovirus was responsible for 93% of non-bacterial gastroenteritis outbreaks in the United States (4). The high incidence of norovirus is attributable both to its low infectious dose (1) and its ability to survive in the environment (5). As a leading cause of gastroenteritis worldwide (6), norovirus is an important concern for local public health departments as well the U.S Environmental Protection

Agency (EPA). It is important, therefore, to develop effective intervention and control strategies for norovirus and similar pathogens. These require both reliable estimates of household transmission parameters and effective analytic tools for obtaining these estimates.

Although there have been studies of community norovirus outbreaks (7), there are no studies that quantify transmission dynamics in the community using a dynamic model. One of the difficulties of these studies is that we often observe only the time of symptom onset for infectious cases. Unobserved events typically include infection and recovery and the times at which these occur. Properly describing the transmission dynamics in household systems necessitates the use of mechanistic models that account for unobserved state variables (eg,the number of infectious and susceptible individuals at any given time), and the more pronounced random variability in outbreaks in small populations.

In this paper, we develop tools to address these challenges and analyze household data collected subsequent to a norovirus outbreak. Götz et al. (8) followed a series of 153 households exposed to norovirus after a 1999 point-source, food-borne outbreak within a network of daycare centers in Stockholm, Sweden. For each of these households, one person (the household index case) was infectious and symptomatic due to the point-source outbreak, and the time of symptom onset for all subsequent cases was recorded. We denote each of these case sequences as a time series.

We analyze these outbreak data using a dynamic model, and obtain maximum likelihood estimates of the household transmission parameter, β, and the average duration

of infectiousness, $1/\gamma$, where γ is the mean daily rate of recovery from infectiousness. We find that the observation of multiple household time-series may provide enough information to mitigate the absence of observed infection times, infectious periods and household sizes.

Methods

Data

Illness data were obtained from a published study of a food-borne norovirus outbreak in 30 daycare centers in Stockholm, Sweden in 1999 (8). The origin of this outbreak was a single food-service worker who shedded norovirus while preparing lunches that were distributed from a central location to 30 daycare centers throughout Stockholm. At the time of the outbreak this worker was infectious but had no overt symptoms.

Among 775 subjects surveyed after the outbreak, 195 cases of gastroenteritis were identified, 176 as norovirus. Among those subjects with norovirus infections, 23 lived alone, 49 lived in households where transmission occurred, and 104 lived with one or more persons but with no observed transmission. Nineteen subjects were excluded because they lived in households with pre-existing cases of gastroenteritis at the time of the outbreak. The primary dataset used in this analysis consists of time series from the 153 exposed households with two or more members.

Data were collected retrospectively for the nine days following the onset of symptoms in index cases. The data consist of the times that cases became symptomatic, reported to the nearest twelve hours and normalized (with the onset of symptoms in the

index case set to time zero). Stool samples were collected from five symptomatic individuals, and the presence of norovirus was confirmed via electron microscopy. Remaining cases were diagnosed based on a norovirus screening interview and a confirmed exposure to a household member infected at the point-source event. Figure 1 provides a visual depiction of the household time-series data for exposed households with secondary cases (modified from the paper by Götz and colleagues (8), Figure 5). When describing household transmission dynamics, we assume that the onset of symptoms corresponds to the beginning of the infectious period. This is supported by a controlled norovirus dosing trial in which early shedding in the absence of symptoms occurred primarily in persons who never became symptomatic (9). Our model also allows the infectious period to be longer than the symptomatic period, which is typical of norovirus infections (9, 10).

In addition, we estimate the distribution of the incubation period, using data reported for the Stockholm outbreak (8) on the time lag between the point-source event and the onset of symptoms in the 153 household index cases. A gamma distribution with mean $1/\varepsilon$, and shape parameter ε_s was fit to these incubation time data by maximum likelihood ($1/\hat{\varepsilon}$ = 1.7 days; $\hat{\varepsilon}_s$ = 3.73 [SE = 0.048]) (Figure 2). In order to fit the assumptions of the compartmental transmission model described in the following section, we round the estimated shape parameter to the nearest integer. However, our estimation approach is robust to models with arbitrarily-distributed infectious periods.

When estimating parameters of the infection-process model, we characterize the infectious period as gamma distributed with an unknown mean and shape parameter. Household sizes were not reported in the original outbreak dataset. To address this

missing data issue, census data on the distribution of Swedish household sizes during the study period were incorporated into our analysis.

Because the Stockholm outbreak data include only the time of symptom onset, we are unable to directly estimate the rate at which asymptomatic infections were created. Accounting for asymptomatic infections is important, as they have been estimated to comprise from 12% to 50% of norovirus cases (11,12,13,14). Additional analysis was conducted to assess the impact of increasing levels of asymptomatic infection on our results.

Model

We treat the household infection process as a continuous-time Markov chain, where persons can be in one of four states: susceptible (S), exposed/incubating (E), symptomatic/infectious (I) and recovered (R) (see Figure 3). The daily transmission rate, β, is defined as rate of contact at time t multiplied by the probability that contact between a susceptible and an infected person results in transmission. We account for the baseline risk of community and environmental infection through the parameter α, which is measured in terms of the daily risk of infection per susceptible. The incubation and infectious periods are assumed to follow gamma distributions, where each is defined by a mean duration ($1/\varepsilon$, $1/\gamma$) and shape parameter (ε_s, γ_s). The shape parameters for the distributions of the incubation and infectious periods can be thought of as the number of stages that persons pass through before they are either infectious or recovered, respectively. These stages are represented by the first-order compartments in Figure 3.

At any given time, t, the hazard, ω_t, to each susceptible in a household is defined by the force of infection,

$$\omega_t = \beta I_t + \alpha$$

Equation 2.1

where I_t denotes the total number of infectious persons in a household at time t.

Consistent with a Poisson process, we assume that these waiting times are exponentially distributed with mean $1/\omega_t$. Under these assumptions, the probability of observing one or more infections over this interval Δt is the exponential cumulative distribution function.

$$p_{Infection}(t, t+\Delta t) = 1 - \exp\left(-\omega_t S_t \Delta t\right)$$

Equation 2.2

The classic model for infectious disease dynamics is the flow of hosts among various compartments defined on susceptible, exposed but not infectious, infectious, and recovered [SEIR]. To generate sample data for evaluating the statistical method described in the next section, we use the force of infection (Eq. 1), gamma-distributed incubation and infectious periods, and household sizes drawn from the census distribution in a stochastic SEIR simulation model. Implementation details are available in the supplementary materials.

Data Model

First, we define a likelihood function for an infection time series when all 4 individual states (susceptible, exposed/incubating, infectious, and recovered) are

observable, and only the transmission parameters β and α are unknown. We then outline a data augmentation method (10) that allows us to extend this likelihood function to the case in which some states are unobserved (Figure 2.4).

Likelihood

The household time series is described as a series of system states, $q_{ij} = \{S_{ij}, E_{ij}, I_{ij}, R_{ij}\}$, for each household, *i*, and state, *j*, where N_Q is the number of distinct system states in a household time series and $Q_i = \{q_{i,0} \ldots q_{i,N_Q}\}$ is the entire set of states in a household in chronological order (see Figure 4). Beginning times for each system state are denoted t_{ij}. Three state transitions are possible: infection, onset of symptoms (and infectiousness), and recovery. The states of the system immediately before the occurrence of infection events, where infection is defined as a transition into E, are indexed by *k* and denoted as $v_{ik} \in V_i$, where $V_i \subset Q_i$. The number of infections in a household observation is N_K.

With this notation, $q_{i,0}$ corresponds to the state of household *i* immediately after the onset of symptoms in the index case, and $v_{i,0}$ corresponds to the state of the household immediately before the first household infection.

Assuming that the times of infection, symptom onset, and recovery are known, we can formulate the household likelihood function as the product of two terms: 1) the likelihood of observing no new cases during the Δt between all state transitions (ℓ_a) and 2) the likelihood of infection at the time when new infection events are observed (ℓ_b).

The expected number of new infections for a given household, *i,* at state *j*, is given by:

$$\lambda(S_{ij}, I_{ij}, \beta, \alpha) = S_{ij}\left(\beta I_{ij} + \alpha\right)$$

Equation 2.3

The first term, ℓ_a, is the probability of observing no infections over all of the time intervals between states:

$$\ell_{i,a} = \prod_{j=0}^{N_Q - 1} \exp\left(-\lambda(S_{ij}, I_{ij}, \beta, \alpha)(t_{j+1} - t_j)\right)$$

Equation 2.4

The second, ℓ_b, describes the joint likelihood of all observed infection events, i.e., the product of all instantaneous infection probabilities at times when infection events are observed:

$$\ell_{i,b} = \prod_{k=1}^{N_K} \lambda(S_{ik}, I_{ik}, \beta, \alpha)$$

Equation 2.5

Based on these definitions, the likelihood of the data for household *i*, given β and α, is:

$$\ell_i = \ell_{i,a} \times \ell_{i,b}$$

Equation 2.6

The product of the likelihoods for all observed households is taken to be the likelihood of the entire observed outbreak, *O:*

$$\ell_O = \prod_{i \in H} \ell_i$$

Equation 2.7

Data Augmentation

The observed data consist of the times of symptom onset in new cases, represented by increments to the household infectious-state variable I_i, and, by consequence, decrements to the state variable E_i. We do not observe infection events for household cases; this is represented by an increment to the household incubating state E_i and a decrement in the number of susceptibles S_i. We also do not observe recovery from infectiousness, represented by an increment to the household immune state R_i (and decrement in I_i). Because all states are necessary to characterize the transmission dynamics of the system, but only transitions into state *I* are observed, a method is needed to evaluate the likelihood. To address this missing-data problem, we generate an augmented household time series by sampling from our incubation and infectious period distributions (mean, shape = $1/\varepsilon, \varepsilon_s$ and $1/\gamma, \gamma_s$, respectively) for each case, as described by Cooper et al.[15] We account for right-censoring by following the convention that all recovery times greater than the observation period, t_f, are truncated to be equal to t_f. This returns the correct likelihood of the data when sampled recovery times are outside

the observation window. In this way, we create an outbreak realization with all states accounted for. Using this augmented dataset, we can calculate the likelihood. We repeat this process many times, re-sampling new times from the distributions and calculating a new likelihood each time. The mean of this set of sampled likelihoods approximates the true likelihood of the household time series. This procedure is equivalent to Monte Carlo numerical integration with importance sampling (16) and is depicted visually in Figure 4.

(See papers by Rampey et al (17) and Rhodes (18) for alternative approaches to estimating transmission parameters with this type of data.)

We obtain a likelihood estimate for an entire outbreak by augmenting all households 10^4 times and estimating their joint likelihood (Eq. 7). Because we are sampling incubation and infectious periods proportionally from their joint distribution, the expectation of this set of likelihoods approximates the likelihood of the data, given the parameters vector $\theta = \{\alpha, \beta, 1/\varepsilon, \varepsilon_s, 1/\gamma, \gamma_s\}$.

In the Stockholm outbreak dataset, the number of people in each household is unobserved. We account for these missing data with household size data obtained from a national census (19) and combine this with information from the household observations; the number of household members must be equal to or greater than the number of observed cases. We combine the census distribution with this lower bound on size for each household to construct a conditional distribution of sizes for each household. When an augmented household time series is generated, a size is sampled from this distribution, allowing us to incorporate and bound our uncertainty regarding household sizes when estimating the likelihood. In the following section we will demonstrate that this does not have a significant negative impact on our results. For details on the implementation of the data augmentation procedure, see the supplementary materials.

The Table lists the two parameterizations used in the analysis. Parameter set 1 uses case and incubation-period data from the Stockholm outbreak. We estimate the transmission parameter, β, as well as the mean, $1/\gamma$, and shape parameter γ_s of the distribution of the infectious period. We constrain our parameter search to values of $1/\gamma$

> 1 day, as durations of symptomatic shedding less than 1 day are biologically implausible (10,11). Parameter set 2 consists of the population parameter values of a single 153-household outbreak realization from the stochastic model, with household sizes drawn from the census distribution. With these simulated data, we estimate β and $1/\gamma$ under two conditions: known household sizes and unknown household sizes.

Results

Figure 5 contains the maximum likelihood estimates and confidence intervals of both the main transmission parameter ($\hat{\beta}$ = 0.14 [95% confidence interval (CI) = 0.08–0.24]; Fig. 5A) and average duration of infectiousness ($1/\hat{\gamma}$=1.17 days [1.00–1.88]; Fig. 5B) for the Stockholm outbreak. We also estimated the shape parameter for the duration of infectiousness (γ_s = 1.0 [1.0–2.0]; not pictured). Figure 6 is a contour plot showing a two-dimensional likelihood profile with respect to β and 1/γ. Each cell contains the likelihood corresponding to the optimized value of γ_S for each $(\beta,1/\gamma)$pair. We also estimate the parameters when α = 0.01 and obtain similar results ($\hat{\beta}$ = 0.13 [0.07–0.22]; $1/\hat{\gamma}$ = 1.0 days [1.0–1.33]; γ_s = 1.0 [1.0–2.0]; not pictured). Thus there is likely some bias in our estimated beta due to environmental infection, but this bias is small.

To examine the impact of unknown household sizes, we created a simulated dataset with parameters β = 0.14 (transmission rate), α = 0.001 (background transmission rate), $1/\varepsilon$ = 1.5 days, ε_S = 4.0, (incubation period), $1/\gamma$ = 1.17 days, γ_S= 1.0 (duration of infectiousness) (See Table, Parameter Set 2). We then estimated two of these parameters, the transmission rate and average duration of infectiousness, under two conditions: 1) where actual household sizes are explicitly included in the estimation (dashed line:

$\hat{\beta}_{knownHH}$ = 0.139 [95% CI = 0.087–0.273], $1/\hat{\gamma}_{knownHH}$ = 1.21 days [0.625–1.88], Fig. 7A); and (2) where household sizes are drawn from the census distribution (solid line: $\hat{\beta}_{unknownHH}$ = 0.133 [0.079–0.259] $1/\hat{\gamma}_{unknownHH}$ = 1.21 days [0.63–1.88], Fig. 7B).

Asymptomatic Infection

To understand the impact of unobserved asymptomatic infections, we performed a simulation-based sensitivity analysis that allows us to predict the value of the transmission parameter, β, for varying proportions of asymptomatic infections, τ.

We find that, starting from our maximum likelihood estimate of β = 0.14 when τ = 0, the predicted value of β increases linearly by approximately 0.035 units for each 10% increase in τ (Figure 2.8). For further details on the design and implementation of this analysis, see the supplementary materials.

Discussion

Using a collection of household-exposure and illness-onset time series, we have obtained estimates (and their confidence intervals) for the household person-to-person infection rate and average infectious period for norovirus. We also predict the value of the transmission parameter β as a function of the proportion of asymptomatic infections. We obtained these estimates despite the absence of potentially important data, including infection times, recovery times, and household sizes. The inclusion of census data with household-specific lower bounds (due to the number of observed cases) allowed us to obtain an accurate estimate of household force of infection in the absence of directly observed household sizes.

Although the pattern of contact in households tends to fit the standard mass-action assumption in susceptible-infected-removed models (20), their typically small sizes require careful consideration of the influence of random variability on results, obviating the use of deterministic models (21,22). This is a topic that has received considerable attention, and there is an extensive literature on techniques for fitting stochastic models to outbreak data (18,23,24) in a variety of settings (e.g., communities (25), schools (26), and households (27)). Using household-level infection data at the end of an outbreak, Longini and Koopman (24) generated estimates of household and community parameters for the distribution of final household outbreak sizes. However, because their method was developed to explain final-size data from public health reports and does not use temporal information, on the method provides only limited insights regarding the interaction between infectivity and the durations of the incubation and recovery periods in outbreak time-series.

Hohle, Jorgenson and O'Neill (28) present a technique that could be useful with household time-series data. They use Bayesian inference to estimate transmission parameters in spatially heterogeneous SEIR models, and innovate on previous Markov-chain-Monte-Carlo-based techniques by allowing variability in the incubation period. Two significant drawbacks of Bayesian approaches are that: 1) even when care is taken to use non-informative prior distributions, these priors can condition estimates (29) and 2) the results can be difficult to interpret, particularly with respect to reproducibility (30). We have presented an alternative, frequentist approach that produces maximum likelihood parameter estimates and allows a straightforward exploration of the likelihood surface.

Community transmission is undoubtedly more complicated than our representation. Fixing the community transmission parameter, α, to a value two orders of magnitude smaller than the household transmission parameter, β, makes the strong assumption that the within-household transmission process is dominant. We show that our results are not very sensitive to this assumption, and we argue that the assumption is reasonable with respect to our data because all households in the Stockholm dataset had a known source of exposure - an index case infected by the point-source outbreak - and all secondary cases identified in households occurred in a plausible temporal sequence. A better estimate of the rate of community transmission requires focused attention on the mechanisms behind this process, which is outside of the scope of both our dataset and this paper. This is an important focus for future research. In addition, the data used in this analysis come from only 9 days of observation, resulting in right-censoring. While our inferences for the transmission rate and effective duration of infectiousness in the course of a household outbreak are valid, they are not generalizable to community or regional scales.

Reliable transmission parameter estimates are critical to risk assessments and exploratory modeling for public health policy. The impact of interventions on norovirus prevalence and persistence can be better assessed in a model such as ours that includes realistic feedback in the transmission process and empirically-derived transmission parameters.

Although the analysis presented here focuses on the transmission of an infectious pathogen in a specific epidemiologic and social context, the methods employed are relevant to other problems in epidemiology and medicine, in which unobserved variables

strongly affect outcomes. We have focused on unobserved within-host disease states and household sizes, but other important variables, including contact structures and environmental reservoirs, are often difficult to observe or missing from otherwise-useful public-health surveillance data.

For example, social and economic factors are likely to increase within-household transmission of pathogens such as tuberculosis and shigellosis (31), by increasing host susceptibility to physical and social stress via mechanisms such as allostatic load and household overcrowding (32). Administrative records often include important information on the timing, geographic distribution, and infectious contacts of cases (33), but because of their focus on immediate control, often lack direct observations of contacts that do not result in infections. Consequently, we lack information on how those who become ill and those who escape infection differ in contact patterns and other factors important in transmission. Our work suggests that case-data missing such information can be combined with reasonable, empirically grounded models of contact structures to yield important and useful insights even in the absence of a full dataset. The next step is to apply this approach to different pathogens in more complicated social settings.

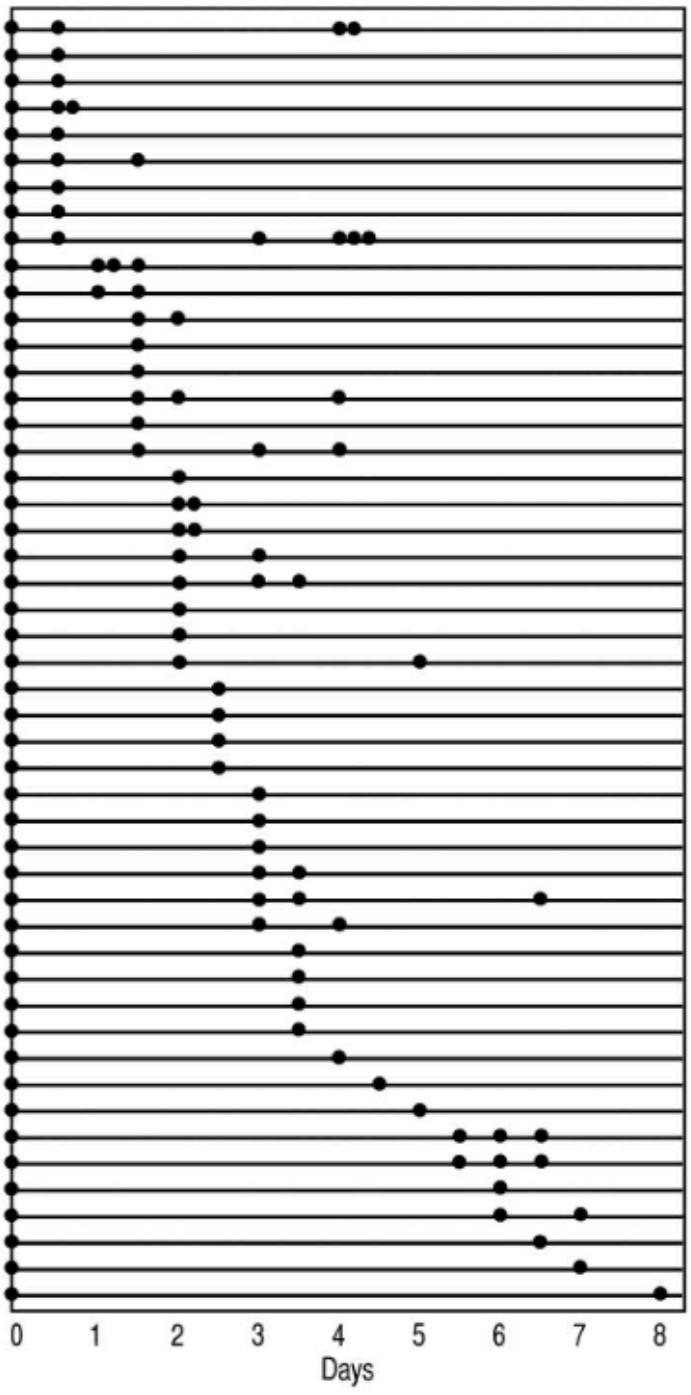

Figure 2.1. Time series for 49 households with secondary infections from Gotz et al data. Time of symptom onset, to nearest 12 hours, is denoted by a filled circle.

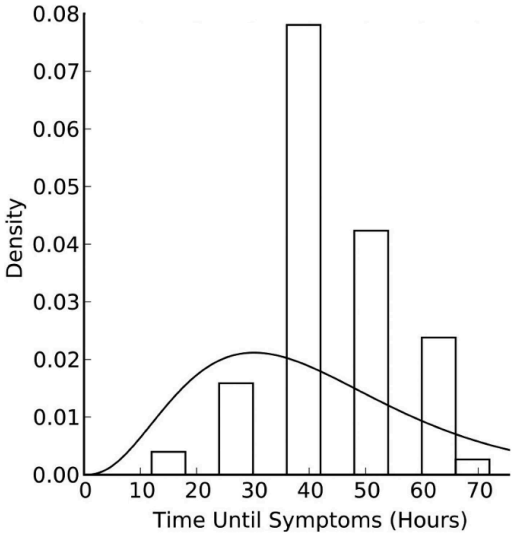

Figure 2.2. Histogram and ML gamma distribution of incubation times from Gotz et al data.

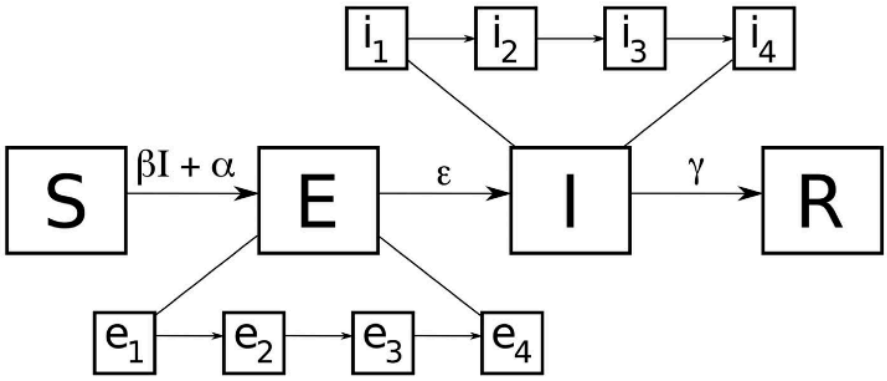

Figure 2.3. Flow diagram showing first and second order compartments in SEIR transmission model. The density-dependent infection rate is β and α is the rate of community transmission. The rate of transition from incubation to symptoms (and infectiousness) and from infectiousness to recovery are ε and Y, respectively.

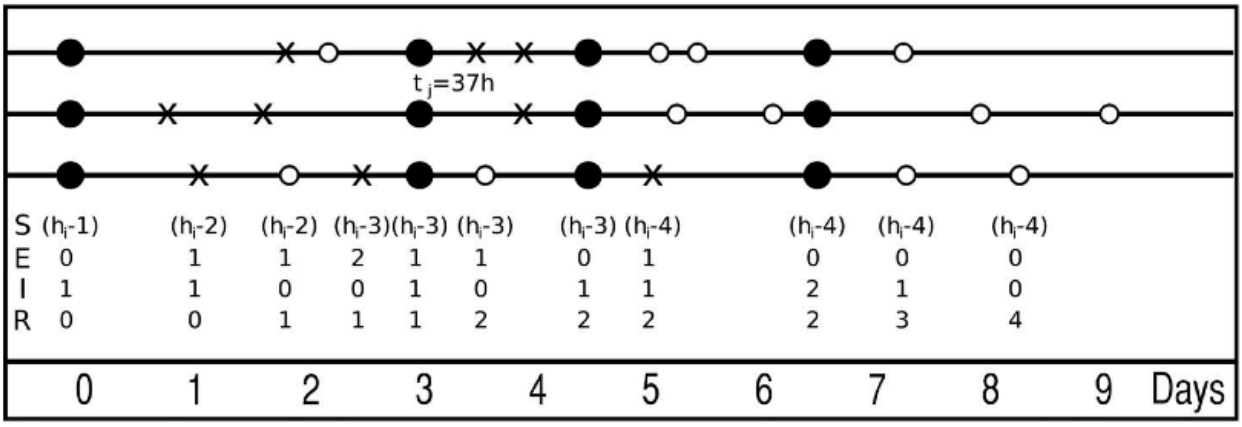

Figure 2.4 Three hypothetical infection histories. The only observed state transition is the onset of symptoms (denoted by a filled circle). Each of the 3 example histories illustrate different possibilities for the 2 unobserved state transitions, infection (denoted by x) and recovery (denoted by an open circle). Values, qij, under the bottom series are the complete state of the system in household h at state i, where S, E, I, R, are the number of individuals in the susceptible, incubation, infectious and recovered states, respectively; hj is the number of individuals in household j.

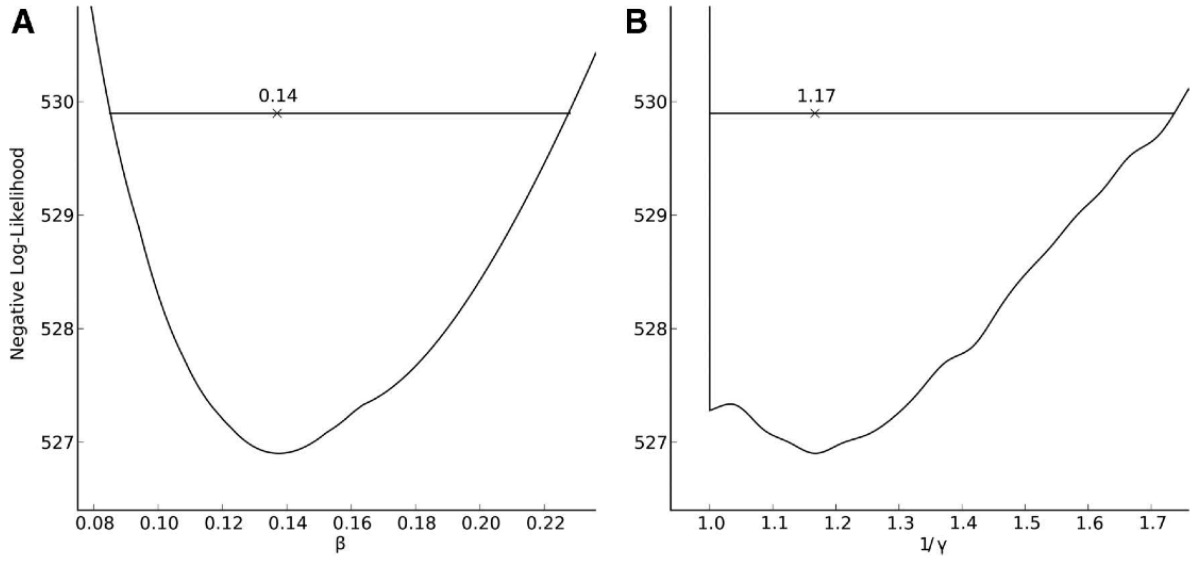

Figure 2.5. Profile likelihood plot of Stockholm outbreak data. Transmission rate (β) and mean infectious period (1/Y) are on the x-axis in panels A and B, respectively. On the y-axis is negative log-likelihood values for a given β or 1/Y when it is held fixed and the other parameters of interest are optimized. "x" denotes the location of the maximum likelihood estimates and the horizontal bar shows the width of the 95% CI.

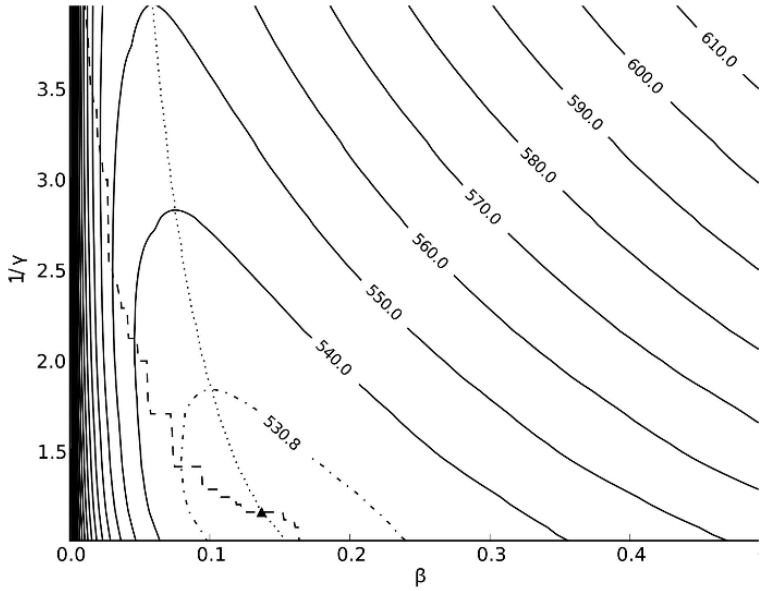

Figure 2.6. Two-dimensional likelihood profile for Stockholm outbreak data. A filled triangle denotes the location of the maximum likelihood estimates. Solid contours bound regions of lesser or equal negative-log-likelihood (NLL) than the contour label. The dash-dotted ellipsoid bounds the 95% confidence region. The dashed line represents the relationship between each value of the transmission rate (β) and the corresponding maximum likelihood estimate of the value of infectiousness period (1/Y) when β is held fixed. The dotted line represents this relationship in reverse, with points along the x-axis corresponding to maximum likelihood values of β for each Y.

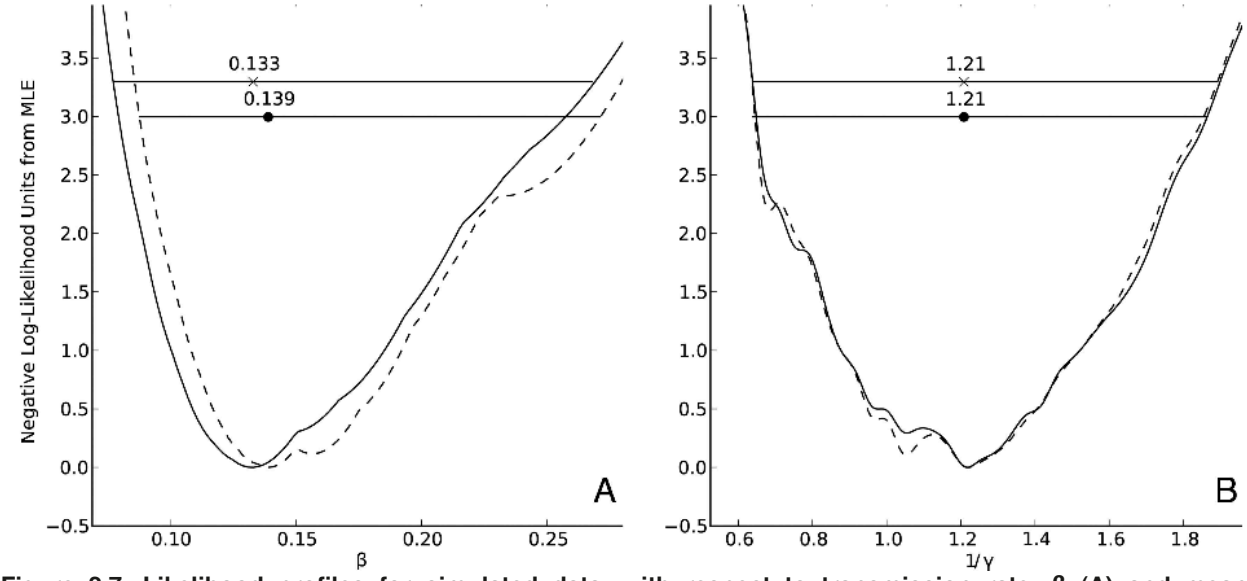

Figure 2.7. Likelihood profiles for simulated data, with respect to transmission rate, β (A) and mean infectious period (1/Y) (B). The dashed line is a profile where household sizes are known (location of the maximum likelihood estimates is denoted by F) and the dash-dotted line estimates the parameters in the case where household sizes are uncertain (MLE: "x"), and the horizontal bars span the 95% CI for both cases.

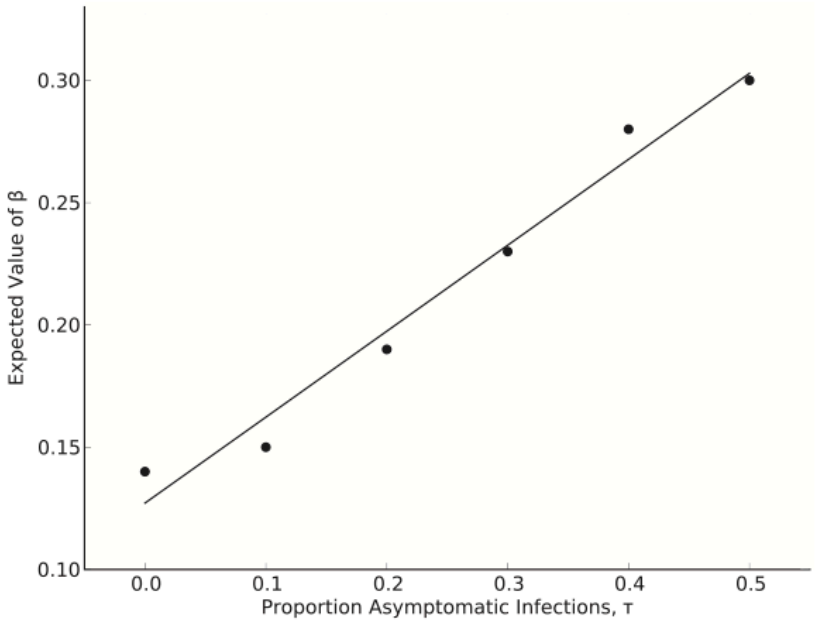

Figure 2.8. Expected household transmission rate, β, by increasing proportion of asymptomatic infections, τ. Note that the expected value of β when τ = 0 is 0.14.

Parameter	Description	Units	Parameter Set	
			1	2
β	Within-household infectivity	Infections/day	EST	0.14
α	Community infectivity	Infections/day	0.001	0.001
$1/\epsilon$	Average incubation period	Days	1.7	1.7
ε_s	Incubation period shape		4	4
$1/\gamma$	Average infectious period	Days	EST	1.17
γ_s	Infectious period shape		EST	1
h	Household size	Persons	Unknown	Known

Table 2.1. Household, Pathogen, and Transmission Parameter Sets. EST indicated parameters to be estimated.

References

1. Teunis P, Moe CL, Liu P, et al. Norwalk virus: How infectious is it? Journal of Medical Virology 2008;80:1468-76.

2. Widdowson MA, Glass R, Monroe S, et al. Probable Transmission of Norovirus on an Airplane. JAMA 2005;293:1859-60.

3. Tsang O, Wong A, Chow C, et al. Clinical characteristics of nosocomial norovirus outbreaks in Hong Kong. Journal of Hospital Infection 2008;69:135-40.

4. Atmar RL, Estes MK. The epidemiologic and clinical importance of norovirus infection. Gastroenterology Clinics of North America 2006;35:275-90.

5. Patel M, Hall A, Vinje J, et al. Noroviruses: A comprehensive review. Journal of Clinical Virology 2009;44:1-8.

6. Caul EO. Viral gastroenteritis: small round structured viruses, calciviruses and astroviruses. Part I. The clinical and diagnostic perspective. Journal of Clinical Pathology 1996;49:874-80.

7. Lopman BA, Adak GK, Reacher MH, et al. Two Epidemiologic Patterns of Norovirus Outbreaks: Surveillance in England and Wales, 1992-2000. Emerging Infectious Diseases 2003.

8. Götz H, Ekdahl K, Lindbäck J, et al. Clinical Spectrum and Transmission Characteristics of Infection with Norwalk-Like Virus: Findings from a Large Community Outbreak in Sweden. Clinical Infectious Diseases 2001;33:622-8.

9. Atmar RL, Opekun AR, Gilger MA, et al. Norwalk virus shedding after experimental human infection. Emerging Infectious Diseases 2008;14:1553-7.

10. Kirkwood C, Streitberg R. Calicivirus shedding in children after recovery from diarrhoeal disease. Journal of Clinical Virology 2008:3.

11. Gallimore CI, Cubitt D, du Pleiss N, et al. Asymptomatic and Symptomatic Excretion of Noroviruses during a Hospital Outbreak of Gastroenteritis. Journal of Clinical Microbiology 2004;42:2271-4.

12. Ozawa K, Oka T, Takeda N, et al. Norovirus Infections in Symptomatic and Asymptomatic Food Handlers in Japan. Journal of Clinical Microbiology 2007;45:3996-4005.

13. Goller J, Dimitriadis A, Tan A, et al. Long-term features of norovirus gastroenteritis in the elderly. Journal of Hospital Infection 2004;58:286-91.

14. Parashar UD, Dow L, Fankhauser RL, et al. An outbreak of viral gastroenteritis associated with consumption of sandwiches: implications for the control of transmission by food handlers. Epidemiology and Infection 1998;121:615-21.

15. Cooper B, Medley G, Bradley S, et al. An Augmented Data Method for the Analysis of Nosocomial Infection Data. American Journal of Epidemiology 2008;168:548-57.

16. Robert CP, Casella G. Monte Carlo Statistical Methods. New York: Springer, 2004.

17. Rampey AH, Longini IM, Haber M, et al. A discrete-time model for the statistical analysis of infectious disease incidence data. Biometrics 1992;48:117-28.

18. Rhodes PH. Counting Process Models for Infectious Disease Data: Distinguishing Exposure to Infection from Susceptibility. Journal of the Royal Statistical Society B: Methodology 1996;58:751-61.

19. United Nations Population Survey, United Nations (New York), 2008. (http://www.un.org/esa/population/unpop.htm).

20. Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford Science Publications, 1992.

21. Koopman JS. Modeling infection transmission. Annual Review of Public Health 2004;25:303-26.

22. Matthews L, Woolhouse M. New approaches to quantifying the spread of infection. Nature Reviews 2005;3:529-36.

23. Ionides EL, Breto C, King A. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences 2006;103:18438-43.

24. Longini IM, Koopman JS, Monto A, et al. Estimating household and community transmission parameters for influenza. American Journal of Epidemiology 1982;115:736-51.

25. Ionides E. Inference and filtering for partially observed diffusion processes via sequential Monte Carlo. Submitted to Journal of Computational and Graphical … 2004.

26. O'Neill PD, Marks PJ. Bayesian model choice and infection route modelling in an outbreak of Norovirus. Statistics in Medicine 2005;24:2011-24.

27. Longini IM, Koopman JS. Household and community transmission parameters from final distributions of infections in households. Biometrics 1982;38:115-26.

28. Hohle M, Jorgenson E, O'Neill PD. Infeference in disease transmission experiments by using stochastic epidemic models. Applied Statistics 2005;54:349-66.

29. Press SJ. Subjective and Objective Bayesian Statistics. New York: Wiley, 2003.

30. Lele S, Dennis B, Lutscher F. Data cloning: easy maximum likelihood estimation for complex ecological models using Bayesian Markov chain Monte Carlo methods. Ecol Letters 2007;10:551-63.

31. Wallace R. A Synergism of Plagues: "Planned Shrinkage," Contagious Housing Destruction, and AIDS in the Bronx. Environmental Research 1988;47:1-33.

32. House JS, Landis KR, Umberson D. Social Relationships and Health. Science 1988;241:540-5.

33. Jones RC, Liberatore M, Fernandez JR, et al. Use of a prospective space-time scan statistic to priorities shigellosis case investigations in an urban jurisdiction. Public Health Reports 2006;121:131-9.q

APPENDIX A

Chapter 2 Supplementary Materials

1. Stochastic SEIR Transmission Model Implementation

A sample outbreak is initialized by creating 153 households, with sizes h_i, drawn from the census distribution of household sizes. The initial household state is set to $q_{i,0} = \{(h_i - 1),0,1,0\}$, indicating that only the index case is symptomatic, all other household members being susceptible. The transmission model is summarized in the algorithm below (Figure 2.9), where S, E, I and R are the number of individuals in each state and the model is initialized at t=0:

```
If E + I > 0:
    For s in S:
        Draw x from Uniform(0,1]
        If x <= 1 − exp(−(βI + α)dt):
            S = S − 1
            E = E + 1
            Draw symptom onset time from
Gamma(1/ε, ε_S)
```

Figure 2.9

The model is stepped forward in hourly increments ($dt = 1/24$), which gives a reasonable approximation of a continuous time infection process. Rates are expressed in terms of days but scaled to the appropriate time step.

The incubation and infectious periods are conceptualized as a sequence of e_s and i_s second-order compartments, with the probability of transition between these compartments for each individual equal to $(\varepsilon \cdot \varepsilon_s)dt$ and $(\gamma \cdot \gamma_s)dt$. This process yields $E \rightarrow I$ and $I \rightarrow R$ transition rates that are gamma distributed with means e, g and shape parameters e_s, g_s, respectively. Transmission rates are also scaled in terms of dt (see Equation 1).

2. Asymptomatic Infections

To assess the effect of unobserved asymptomatic infections, we implemented the stochastic SEIR model outlined above, with an additional parameter, τ, that controls the proportion of new infections that are asymptomatic:

```
If E + I > 0:
    For s in S:
        Draw x from Uniform(0,1]
        If x <= 1 – exp(–(βI + α)dt):
            Draw y from Uniform(0,1]
            If y <= τ:
                S = S – 1
                R = R + 1
            Else:
                S = S – 1
                E = E + 1
```

Figure 2.10

Asymptomatic infections are, in this simplified model, immediately moved to the immune class. This is because they are significantly less infectious than symptomatic infections, e.g., (10), and can be expected to generate cases on a longer timescale than our window of observation (9 days). Although they are unlikely to contribute significantly to observed within-household transmission dynamics, we expect that they are important to the community-level persistence of norovirus and, as such, need to be accounted for in the estimate of rate of transmission. In this context, then, asymptomatic cases can be thought of as censored data that bias our estimate of the force of infection.

When simulating outbreaks, we fix the background infection rate and the distribution of the incubation and infectious periods,(a = 0.001, $1/e$= 1.7 days, e_S = 4.0, $1/g$= 1.14 days, g_S = 1.0) and allow the transmission parameter, β, and proportion of asymptomatic infections, τ, to vary. We then sample all 126 parameter combinations from β = {.10, .11, … ,0.30} and τ = {0, 0.1, 0.2, 0.3, 0.4, 0.5}. We draw 20 stochastic realizations of each parameter set and estimate the mean ML value of β (i.e., average over the 20 runs) for each (τ,β) combination, as though τ = 0. This gives a predicted value of β for each level of τ. Starting from our ML estimate of 0.14 for β when τ = 0, the predicted value of β increases linearly by 0.035 units for each 10% for increase in τ (Figure 8).

We test the sensitivity of these results to the assumption that asymptomatic individuals do not contribute to household transmission by allowing asymptomatic infections to be 10% as infectious as symptomatic ones. We find broadly similar results,

with the predicted value of β increasing linearly by 0.028 units for each 10% increase in τ (Figure 2.11).

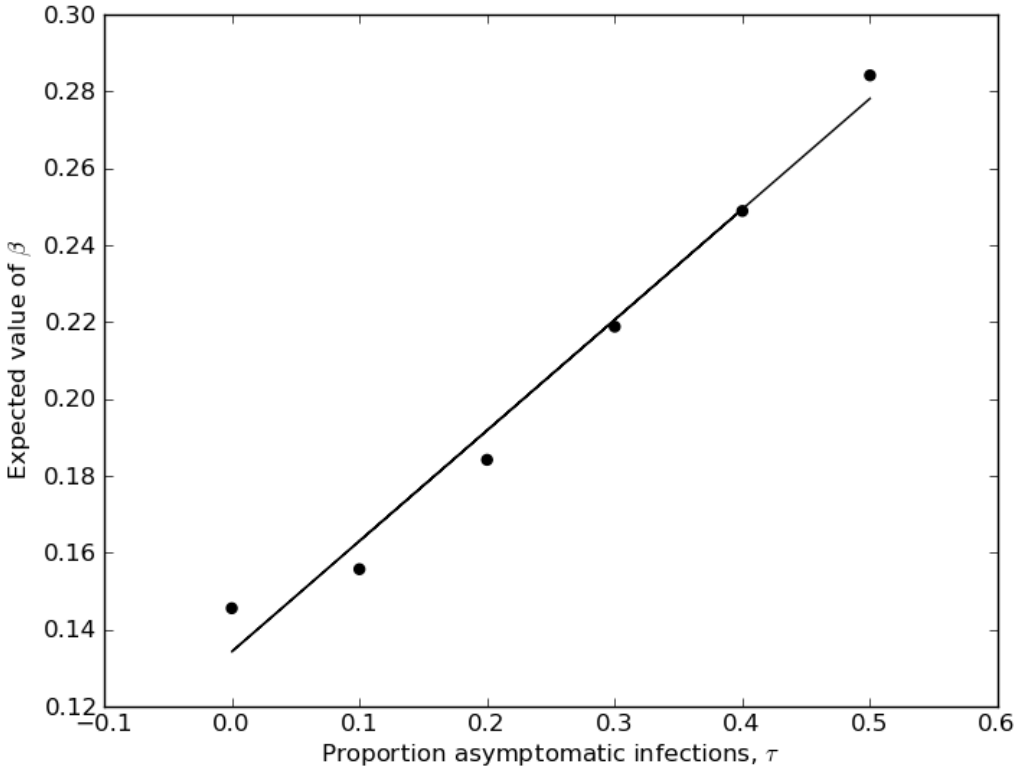

Figure 2.11. Relationship of proportion asymptomatic to expected value of β when asymptomatic infections are 10% as infectious as symptomatic infections.

3. Missing Household Sizes

Since all households in our dataset consist of two or more people, the minimum household size, *h*, is 2. We start with the empirical distribution of household sizes from a 1990 census of household sizes in Sweden (see Table 2.2), denoted as *C,* where *C(h)* is the probability of observing a household of size *h* in the total population .

If the minimum possible number of individuals, i.e., the number of infections observed in a household, $h_{\min}$, is less than or equal to 2, the entire empirical distribution is used to sample a household size. If $h_{\min} = 3$, the number of cases observed is set as the minimum household size, with values smaller than $h_{\min}$ assigned a density of zero. We assume that the case data provide no additional information on the distribution of the remaining household sizes, so the remaining sizes on the interval $h_{\min} \le h \le 10$ are assigned a uniform density.

This information is combined with the census data in the top row of Table 2.2 for each size to generate a distribution from which we can sample household sizes for h ³ $h_{\min}$:

$$P(h \mid C, h_{\min}) = \frac{C(h)}{\sum_{h \ge h_{\min}}^{10} C(h)}$$

Equation 8

In order to sample random variates from this distribution, we compute the conditional CDF of the household size distribution and draw a random number on the interval (0,1], and select the smallest value of *h* where the CDF is less than equal to the random number.

The second row of Table 2.2 shows the probability distribution resulting from this sampling procedure. We find that the expected household size increases slightly from 3.73 to 3.87 individuals, with most of this change accounted for by a decrease in the density of households of size 2 to slightly larger ones.

	# Household Members								
	2	3	4	5	6	7	8	9	10
Census Density	0.325	0.193	0.248	0.108	0.027	0.041	0.024	0.017	0.017
Sampled Density	0.283	0.192	0.265	0.115	0.031	0.047	0.027	0.018	0.019

Table 2.2 Empirical Probability Distribution of Household Sizes

4. Model Validation

In order to validate the SEIR model used for simulation and parameter estimation, we performed additional simulation analysis using a Gillespie[1] algorithm-based implementation of the model described in Figure 2.9, which is an exact, continuous-time simulation of the transmission model.

In each simulation, there are 153 households, the sizes of which are drawn from *C*, the empirical distribution of household sizes. At *t=0*, each household has a single index case. Model parameters are the same as those obtained from our statistical analysis

(β = 0.14, $1/\gamma$=1.17 days, γ_s = 1.0). For each of 1000 simulations, we record the number of households with no secondary cases, i.e., where there is stochastic die-out, and the average number of cases in households with secondary cases.

We find that our simulation results are in good agreement with the Stockholm data for both outbreak size (Simulated mean = 1.9 cases, SD = .2, vs. 1.6 for Stockholm data; Figure 2.12) and the number of simulated households in which there are no secondary cases (Simulated mean = 110.5 households, SD = 5.5 vs. 104 households for Stockholm data; Figure 2.13).

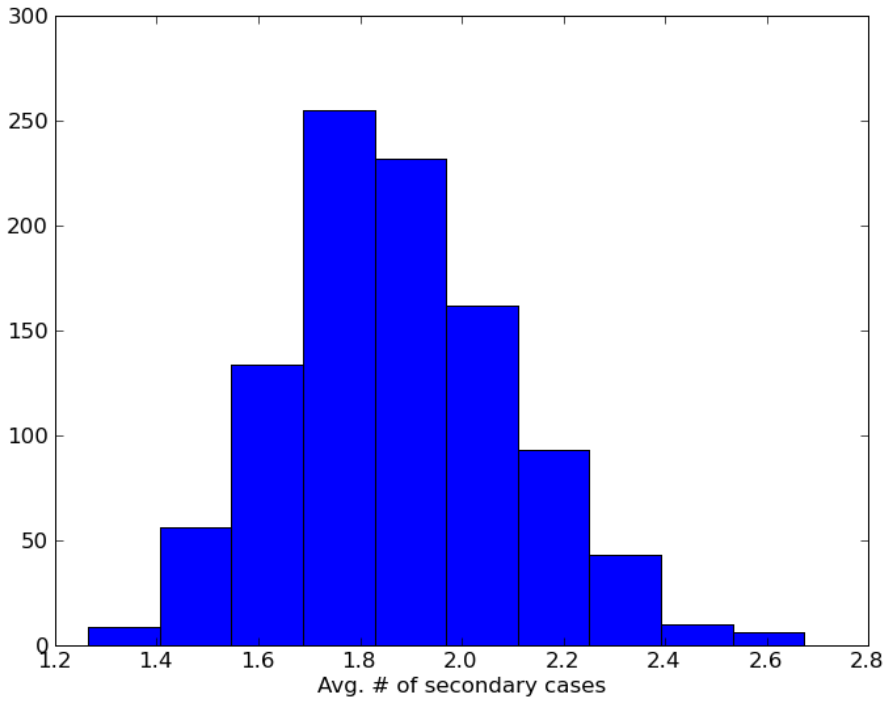

Figure 2.12. Histogram of average number of secondary cases in simulated household outbreaks.

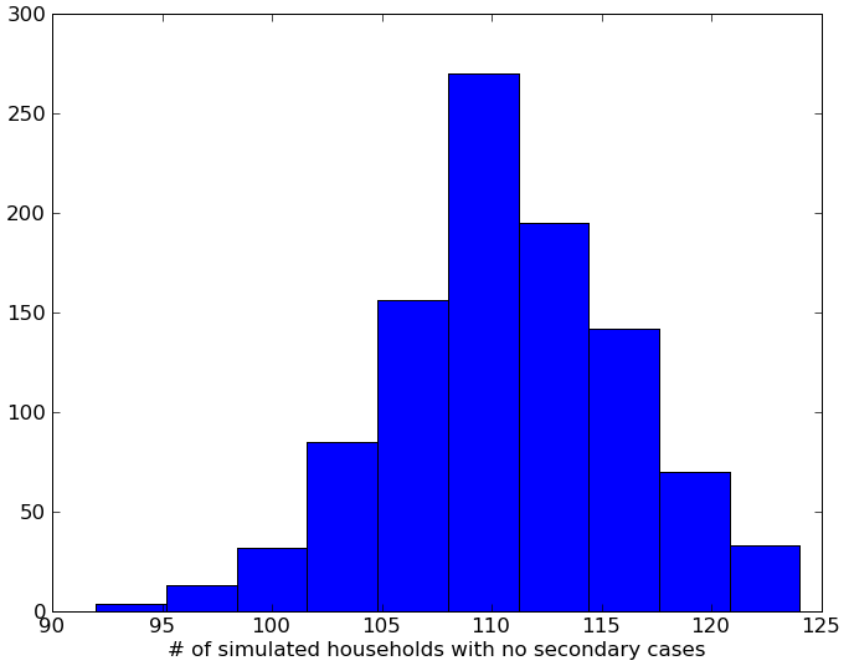

Figure 2.13. Histogram of number of households with no secondary cases.

5. Computational Details

Data augmentation software was implemented in C++ and Python 2.6 using *Boost.Python* and the *Numpy* and *Scipy* numerical and scientific computing libraries. Plots were generated with *Matplotlib* 0.98 graphing and plotting tools for Python. All diagrams were created in *Inkscape* 0.47.

All results presented here come from 10^4 independent samples for each parameter combination.

References

1.Gillespie, D.T. (1976). "A General Method for Numerically Simulating the Stochastic Time Evolution of Coupled Chemical Reactions". *Journal of Computational Physics* **22** (4): 403–434

Chapter 3

Modeling the effects of social behavior on community-level Norovirus outbreaks

Abstract

Norovirus (NoV) is a highly infectious gastrointestinal pathogen affecting all age groups. NoV epidemiology is characterized by small, explosive events that are often linked by common individuals and shared environments. This explosive, variable outbreak behavior is the product of variation in the number and type of contacts individuals have, denoted as *social variability*, and heterogeneity in individual infection histories, which we term *biological variability*. In this paper, we use a dynamic, network-based model to simulate community-level norovirus outbreaks. We examine the effects of social behaviors and attributes that lead to greater heterogeneity in contact patterns, such as age-structured mixing patterns and exponentially distributed variation in the propensity of individuals to initiate relationships. In addition, we examine the role of asymptomatically infectious food servers, who have the ability to infect many people at one time, in community norovirus outbreaks. We also examine the effect of behaviors that dampen heterogeneity in contact, such as the clustering of relationships within neighborhoods and amongst smaller groups of individuals. We find that adding age-structured contacts to our community outbreak model increases the average size of simulated outbreaks by 65%,

and that including variation in individual sociability amplifies outbreak size by 56%. We also find that strong neighborhood-based assortativity decreases outbreak size by an average of 17%. We use mediation analysis to understand how these social behaviors operate on outbreak risk through structural features of the social networks used in simulations. Finally, we shed light on the explosive character of norovirus outbreaks through an examination of drivers of heterogeneity in the distribution of NoV outbreak sizes.

Introduction

Norovirus (NoV) epidemiology is characterized by sporadic outbreaks that are epidemiologically linked (1). This pattern of explosive but locally limited transmission is driven by variability in the ability infectious individuals have to generate new infections. The mechanisms driving this variability can be categorized as social, i.e., variability in the number on contacts, or biological, i.e., variability in the duration of individual-level shedding. Most models of infectious disease transmission either ignore heterogeneity in overall transmission, e.g.(2, 3) or address heterogeneity by collapsing multiple types of variability into a single construct, e.g., the distribution of the individual reproductive number (4, 5, 6, 7). This latter set of models, which aggregate multiple types of variability in host behavior and infectivity into heterogeneous transmission, have been effective at illustrating how heterogeneity impacts transmission dynamics and risk. But this approach conceals specific mechanisms that can be points of high leverage for controlling norovirus outbreaks. In this paper, we address this gap by modeling specific social processes using simulated social networks whose structure is a function of 1) age, 2) variability in individual sociability, 3) neighborhood of residence and 4) clustering of

network relationships. To do this, we take advantage of data on: 1) contact networks, with particular attention to age structured mixing (8, 9); 2) heterogeneous durations of pathogen shedding resulting from variable host immunity(10); and 3) general patterns of norovirus (NoV) outbreaks, to examine how social and biological processes impact incidence patterns of NoV, with specific focus on the explosive nature of NoV outbreaks, and their duration.

Norovirus epidemiology

Norovirus outbreaks are characterized by small, explosive events that affect households (11, 12), workplaces (13, 14), schools (15) and healthcare facilities (16). While NoV outbreaks in communities are typically explosive and relatively short in duration, at the regional and national level, NoV infection is persistent and stable. In fact, norovirus is the most common cause of nonbacterial gastroenteritis in the U.S. and worldwide (17). One possible explanation for this pattern of local instability and global persistence may be found in wide variability individuals have in their ability to create infections. Such variability has been shown to make outbreaks sporadic, but more explosive when they occur (4). This variability may be a consequence of biological processes that make shedding duration variable across individuals (10), social processes that cause variability in the number and transmissibility of contacts, or both. For example, biological variation in individual infectiousness is a key feature of norovirus epidemiology (10): immunocompetent individuals exhibit variability in norovirus shedding lengths, and immune compromised individuals and very young children (aged < 1 years) have been observed to have very long asymptomatic shedding periods, with a mean of 87 days and maximum durations in excess of a year (10). Likewise, social

variability can be observed at the individual level. For example, a single food handler may expose from one to several thousand people at once. At the group level, individuals may have more ties within their neighborhoods than outside of them, resulting in isolated local transmission networks with relatively few links to and from adjacent neighborhoods.

NoV outbreaks are often started by a common-source event, such as the exposure of many people at one time to an asymptomatic but infectious food handler or health care worker (13, 18). Although much is known about the molecular links between norovirus outbreaks (e.g.,(19)), less is known about the way these types of outbreak events are linked to each other via chains of person-to-person and person-to-environment-to-person (20) transmission. For example, an originating point-source event may begin a series of undetected infections that are transmitted by contact in households and the community and terminate in another large, common-source outbreak that comes to the attention of public health officials (e.g., (11)). In this manuscript we focus on social processes but include biological heterogeneity in shedding duration. The outcomes of these social processes are modeled as community transmission networks. We conceptualize the transmission modes of norovirus in the community using a model that includes point-source events, such as contamination by food handlers, household transmission, and asymptomatic transmission via contacts in the community.

Community transmission networks

Much published work on norovirus epidemiology describes moderate to large-sized outbreaks in a variety of contexts including the home (11, 12), workplaces (21), and other venues, such as sporting events (22) and cruises (23). But there has been little emphasis on the way transmission events in these contexts are linked via casual contacts. Here, we outline a model of community contact networks that will be used to generate contact networks for our outbreak simulations. This model is based on two broad classes of behavior, which may a) increase, and b) decrease, heterogeneity in contact at the individual or group level. Factors in our model that amplify variability in contact include the following two items:

*1) Age-structured contacts***.** People tend to have relationships that cluster, i.e. match, within age groups or to span, i.e. mix across, age groups in a systematic fashion (e.g., students and teachers. This makes contacts more homogeneous at the level of age groups but increases variability across age groups in the population. Previous research (8, 9) has shown that age-dependent variability in contact patterns can be useful for explaining outbreak dynamics. These data typically reflect the expected number of hours individuals spend together on an average day as a function of their ages. When individuals are of the same or similar age, this is referred to as *assortativity* by age, or a tendency to have connections to similar individuals. By contrast, when individuals of differing ages have contact, this is referred to as *dissassortativity* by age. Here, we employ survey data from Zagheni et al. (8) on the amount of time individuals spend together, outside of the home, to construct realistic community networks.

2) Individual sociability. Variability in the propensity of individuals to have relationships, i.e. their *sociability*, drives variability in the number and patterning of

contacts at the individual level. This can result in long-tailed distributions of contacts, which have been shown to impact the probability, size and timing of outbreaks (5).

The behavioral mechanisms in our model that decrease heterogeneity in contact patterns operate in a more uniform way across individuals or groups than the mechanisms above:

*3) Neighborhood clustering***.** Individuals are more likely to have contact with network alters who belong to the same social groups as they do (24, 25). Such group level assortativity in contact, evidenced by patterns such as residential segregation, has been shown to be important to outbreak dynamics (26). Here, we divide the set of households into neighborhoods where members make contact preferentially with other individuals who live in that neighborhood. When the average number of contacts in the community is held constant, this group-level assortativity makes individuals more likely to make contact with infectious individuals in their neighborhood, while making them less accessible to those outside of it. Variability in the structure of networks as a function of group-level variability (age-based mixing) and homogenizing factors (neighborhood clustering) is illustrated in Figure 3.2.

4) Relationship clustering. Relationships are clustered when individuals share network partnerships, resulting in a closed triad. A closed triad is a group of nodes, *a*, *b*, *c*, each of which is connected to the others and are a common feature of socially cohesive networks (27). An increase in this kind of clustering has been shown to decrease the likelihood of outbreaks in networks with equivalent average degree (28). At the same time, a tightly clustered group of nodes is more likely to transmit infection internally than

is a loosely connected group, but this infection is less likely to spread into nodes outside of the clustered group.

Analysis of the role of social networks in the epidemiology of infectious diseases has typically focused on isolated aspects of contact networks, e.g. the clustering coefficient and parameters of the contact network degree distribution. Relying solely on such top-down measurements of network structure makes it difficult to understand whether it is these average, macro-level properties of the network or specific micro-level processes, such as relationship clustering and variation in sociability, and group-level processes such as age-group and neighborhood-based clustering, that drive transmission. Integrating these specific sources of variability in social behavior and pathogen shedding in a single model is an important step towards understanding the extent to which micro and group-level behavioral features drive the heterogeneity and unpredictability in outbreaks of NoV and other pathogens.

Methods

Our norovirus outbreak simulation model utilizes transmission and community contact mechanisms based on results from empirical studies. We use this model to examine the effect of social variability on outbreak size as it interacts with the biological heterogeneity outlined above. We are interested, specifically, in the impact that these processes play in the size of average size of community outbreaks aver point-source events and heterogeneity in the distribution of these outbreak sizes.

Community Transmission Model

The community transmission model links the social and biological mechanisms underlying norovirus (NoV) transmission in a single framework. Model runs occur in a simulated community of 10,000 people and begin immediately after a point-source

outbreak of norovirus. The community is partitioned into 4 equally sized neighborhoods with identical age distributions. When the model is initialized, households are sampled and populated with individuals using the procedure described in Algorithm S1 in the supplementary materials. Individuals have fixed contacts in the community network, i.e. contacts do not change over the course of an outbreak. Individuals who are food servers, i.e. those who can cause point-source outbreaks when asymptomatic, are identified at initialization but can only create point-source events when asymptomatically infectious.

Infection Process

We model the progression of norovirus infection using a stochastic SEIR model (3). We characterize heterogeneity in the duration of asymptomatic NoV infectiousness using a mixture distribution that describes the general population as well as sensitive population groups (see (10). In addition to the standard infectious state, we include an additional asymptomatic phase of infection to account for individuals with very long asymptomatic infectious periods. Individuals may be in any of the following states: susceptible (S), exposed/latency (E) symptomatic infectious (I_S), regular asymptomatic shedding (I_R), extended asymptomatic shedding (I_L), or recovered (R).

Below, we present a deterministic, differential equation based transmission model that illustrates how infected individuals progress through the disease states associated with norovirus infection. This basic model assumes homogeneous mixing, i.e., no community structure. We include this model as a compact illustration of the way individuals progress through infection states. After this, we will outline our community

model, which integrates this infection process with contact networks that reflect household structure and contact variability.

$$\frac{dS}{dt} = -\beta_{IS} S I_S + \beta_{IA}\left(I_R + I_L\right) \quad \text{(a)}$$

$$\frac{dE}{dt} = \beta_{IS} S I_S + \beta_{IA}\left(I_R + I_L\right) - \varepsilon E \quad \text{(b)}$$

$$\frac{dI_S}{dt} = \varepsilon E(1-\tau) - \gamma_S I_S \quad \text{(c)}$$

$$\frac{dI_R}{dt} = \gamma_S I_S + \varepsilon\tau E - I_R \gamma_R \quad \text{(d)}$$

$$\frac{dI_L}{dt} = \rho\gamma_R I_R - I_L \gamma_L \quad \text{(e)}$$

$$\frac{dR}{dt} = (1-\rho)\gamma_R I_R + \gamma_L I_L \quad \text{(f)}$$

Equation 3.1

For an analysis of this model, see Milbrath et al. (10). For a visual depiction of this infection process, see Figure 3.1.

Each simulation run begins with *N-q* individuals in the susceptible state (S), and *q* individuals in the incubating state (E), which represents the state of the system immediately after a point-source event in which *q* individuals were infected. Susceptible individuals are infected and enter the incubating/exposed state (E) at an average rate of $\beta_{IS} I_S + \beta_{IA}\left(I_R + I_L\right)$, where β_{IS} and β_{IA} are the transmission rates for symptomatic and

asymptomatic individuals, respectively. Infected individuals leave the incubating stage at an average daily rate of ε (Equation 1b).

A fixed fraction of the infected population, τ, is never symptomatic, and transitions directly from the incubating state to the regular asymptomatic shedding state (I_R). The remaining 1-τ individuals in *E* transition to the symptomatic infectious state (I_S) (Eq. 1c, 1d). In our stochastic implementation of this model, the duration of the latency period is assumed to follow a gamma rather than an exponential distribution, and, following Zelner et al. (12), has shape parameter, $\varepsilon_S = 4$.

Symptomatic infectious individuals transition from I_S into the regular-duration asymptomatic state, I_R, with average rate γ_S. Individuals in I_R leave with rate $1/\gamma_R$ (Eq. 1d). A fraction of these individuals, ρ, progress to the extended asymptomatic infectious phase, I_L (Eq. 1e) which they remain in for an average duration of $1/\gamma_L$, while the remaining 1-ρ immediately enter the recovered phase, R (Eq. 1f). The total number of asymptomatic infectious cases is denoted by $I_A = I_R + I_L$. Parameter definitions and estimated values used in simulations are listed in Table 3.1.

Household transmission

We model the transmission of norovirus within households using mechanistic parameters derived by Zelner et al. (12) from a series of household NoV outbreaks in Stockholm, Sweden (11). We assume that households are fully connected networks, i.e. all individuals in a household are connected to all other individuals in that household. We also assume that household transmission is density dependent, so that the per-contact rate of infection remains constant as the number of household members grows. To model

variability in household size and the distribution of ages within these households, we sample the age of adult householders from the U.S. Census distribution of ages (29) then sample the number of resident children and their ages using age-specific parity tables (30). Transmission parameter values used in simulations are presented in Table 1. For further details on household size and composition, see the supplementary materials.

Point source outbreaks

An important aspect of the heterogeneity in transmission in our model is the role played by asymptomatically infectious food servers who can infect many individuals at one time. At the beginning of each simulation run, a sample of individuals is selected from the population of adults in the simulated community, each with probability ζ, to be food servers. We represent food service by an asymptomatically infectious individual as a Bernoulli trial (i.e., flipping a biased coin) in which a success is defined as a time step in which no infections are generated, and a failure results in a point-source exposure. For example, a point-source exposure might occur when an asymptomatic individual who had previously prepared food while infectious but caused no infections, i.e. a series of successful trials, washes her hands in the same sink used for cleaning vegetables, resulting in a PSE, e.g. (31). Since asymptomatically infectious food handlers individuals are likely to be detected in the course of public health investigations, we assume that the individual is detected and removed from the workplace after the first exposure event. Such a process results in waiting times, represented by the random variable T_{PS}, from the onset of asymptomatic infection to a point source event that are distributed according to an exponential distribution, i.e.:

$$T_{PS} \sim Exponential(1/\lambda)$$

Where λ is the daily attack rate for point-source events generated by hub individuals and $1/\lambda$ is the expected waiting time from onset of asymptomatic infectiousness to a point-source event. If the waiting time for a PSE exceeds the food-server's asymptomatic infectious period, no PSE occurs. When a point-source exposure does occur, a group of individuals is sampled uniformly at random from the population and infected. The magnitude of the event is assumed to follow a geometric distribution with mean $\overline{Q}$.

For ease of reference throughout the next sections, we define the set of parameters governing household, community and point-source transmission to be a vector $\Phi = [\beta_{IS}, \beta_{IA}, \overline{Q}, \tau, \rho, \lambda\varepsilon, \varepsilon_s, \gamma_S, \gamma_R, \gamma_L]$.

Community transmission networks

We use exponential random graph (ERG) models (32, 33) to embed different types of social behavior in the contact networks used in outbreak simulations. ERG models specify a probability distribution from which one can generate networks where the probability of links forming between pairs of individuals in the network is a function of attributes of those individuals. Such models are ideal for understanding how multiple types of variability in social behavior impact epidemic outcomes. For instance, ERG models can account for age-specific behavior as well as assortativity by the neighborhoods that individuals live in. They may also include other determinants of network structure, such as a tendency for groups of individuals to have many shared relationships, and long-tail variation in individual-level propensities to have relationships.

When specifying an ERG model representing these behaviors, we define $M(x) = \left[M_A(x), M_S(x), M_N(x), M_R(x)\right]$ to be a vector of statistics measuring properties of an observed network, denoted x. $\theta = \left[\theta_A, \theta_S, \theta_N, \theta_R\right]$ is a vector of parameters controlling the relative strength of contributions of these network statistics to the resultant network. X is a random variable with a distribution from which the observed network, x, is drawn and $\bar{k}$ is the average degree of networks drawn from this distribution (34):

$$P(X = x | M(x), \theta, \bar{k}) = (1/\kappa)\exp\left\{\sum_{i \in \{A,S,N,R\}} \theta_i M_i(x)\right\}$$

Equation 3.2

We represent the network as a set, so $X = \{V, E\}$ where the set of nodes, or vertices in X, is a vector ordered by node ID and denoted V, and the number of elements in V is denoted as N. A is the adjacency matrix of x, so that $A_{ij} = 1$ if nodes i and j are connected in the community network and zero otherwise. Realized edges in A comprise a set, E, consisting of unordered pairs of nodes:

$$E = \left\{(i, j) \text{ s.t. } A_{ij} = 1\right\}$$

Because the normalizing constant κ in Equation 3.2 is difficult to compute, we use the Metropolis Hastings (MH) algorithm to generate random graphs from the distribution specified by Equation 3.2. Our MH algorithm is stepped forward by moving edges from uniformly selected pairs of connected nodes to uniformly selected pairs of nodes that are not connected. For a detailed explanation of the procedure used to sample graphs from this distribution, see (25, 32, 35, 36) and the supplementary materials.

Below, we provide detailed definitions of the network statistics in **M(x)** that are used to represent social behavior in the community contact model:

Age-structured contacts. The term representing age-dependent mixing, M_A, reflects the amount of time spent together by individuals in the network as a function of their ages, as measured in the time use data of Zagheni et al. (8).

The average number of hours that individuals of some pairing of age groups (or the same age group) spend together on any given day is presumed to be proportional to the odds of an edge forming between them. In our model, the time-use data are represented as a vector of parameters, Ω, where each entry corresponds to the expected number of hours per day spent together by individuals for each pair of ages in *x*, divided by the maximum number of hours in the time-use data. This makes the quantities in Ω analogous to odds ratios. The distribution of age-pairings is represented by a vector, **AP**. Entries in **AP** are counts of the number of elements in the set of edges, **E**, that connect individuals of the ages corresponding to each pair in Ω. We assume that edges that form as a function of age are independent of other realized edges, so we compute $M_A(x)$ as the sum of the natural logs of the parameters in Ω, which is equivalent to the natural log of the product of these values:

$$M_A(x) = \sum_{i=1}^{\Omega} \ln(\Omega_i) AP(x)_i$$

Equation 3.3

The formulation in Equation 3.3 implies the simplifying assumption that all community contacts are of uniform duration, so that increasing the amount of time spent

together by individuals in a pair of age groups is equivalent to increasing the number of network connections between individuals of these ages. Because the parameters in Ω control changes in the log-probability of the model, the model parameter θ_A is used only to include or exclude age-structured contacts and is therefore constrained to take the values 1 and 0.

Individual sociability. The term representing the sociability of individuals in the community, M_S, reflects the strength of individual-level tendency to create relationships with other individuals the community. Following results from existing social network studies, e.g. (5, 37), we conceptualize this propensity as an exponentially distributed quantity, the values of which are assigned uniformly at random to individuals in the community. We define a vector of length *N*, denoted *K*, where each entry is the sociability parameter of the corresponding node in *x*. As with age-structured mixing, the sociability of a node is analogous to the odds ratio, relative to an average individual, that the individual represented by the node will form a tie with any other individual. When the model is initialized, each node's *absolute* sociability is drawn from an exponential distribution with mean $\bar{S}$. This quantity is divided by $\bar{S}$ and entered in *K*. $\bar{S}$ is fixed at 10 to simplify analysis. As with age-structured contacts, the parameter for sociability, θ_S, is constrained to take only 1 and 0 as values. Again, as with age-structured mixing, we use the natural log of this value when computing M_S:

$$M_S(x) = \sum_{i=1}^{N} \ln(K_i) D_i$$

Equation 3.4

Neighborhood clustering. We define another network statistic, M_N, to be the number of edges in the network linking individuals who live in the same neighborhood. We define the vector $\boldsymbol{Z}$ to be the IDs of the neighborhoods to which individuals are assigned, where the elements of $\mathbf{Z}$ are indexed by node ID:

$$M_N(x) = \sum_{(i,j)\in E} I(i,j)$$
$$I(i,j) = 1, \text{ if } \mathrm{Z}_i = Z_j$$
$$I(i,j) = 0, \text{ otherwise}$$

Equation 3.5

The parameter θ_N expresses the change in the log-probability of *x* corresponding to the addition of a single edge that connects two individuals who live in the same neighborhood. Values of θ_N are constrained to be greater than zero. This indicates increasing assortativity by neighborhood.

Relationship clustering. Sharing of network partners, or local clustering, is accounted for by the globally weighted edgewise shared partner (GWESP) statistic of Hunter et al. (38), which we denote as $M_R(x)$. This is computed as:

$$M_R(x) = \exp(\eta)\sum_{i=1}^{N-2}\left\{1-(1-e^{-\eta})^i\right\}EP_i(x)$$

Equation 3.6

This statistic represents the extent to which nodes in the graph that are connected to each other will also share additional network partners. The distribution of edgewise

shared partners is denoted *EP* and indexed by *i*, so that $EP_i(x)$ is the number of edges in *x* that share exactly *i* partners. The parameter η controls the rate of geometric decay in the weight assigned to increasing values of this distribution, so that an edge with many shared partners is given only slightly more weight than an edge with a few shared partners. In all analyses, the value of η is fixed at 0.5. It has been shown that the output of the sampling algorithm is not very sensitive to the choice of a value for η (24). The parameter θ_R represents the change in log-probability of *x* associated with a 1-unit increase in this weighted measure of transitivity.

Finally, when sampling contact networks, we fix the number of edges so that the average degree of nodes in the network remains constant. This allows us to compare outbreak dynamics on community networks of differing average degree but with similar structure-generating mechanisms.

When transmitting over these networks, the parameter for infectivity in the community is scaled by the average degree, $\bar{k}$, so that the per-contact transmission rate is equal to $\beta_{IA}/\bar{k}$. This means that β_{IA} expresses the average number of infections the average asymptomatically infectious individual in the community will create on a given day.

Measurements of community networks

The individual and group-level behavior implied by the exponential random graph (ERG) model of community contact results in a variety of network structures. Given a set of transmission parameters, it is this structure of contacts that will determine the likelihood that outbreaks occur on the community network. To understand the impact that

the network parameters, $\theta = [\theta_A, \theta_S, \theta_N, \theta_R]$, have on variability in the resultant networks, we use three measures of network structure:

Coefficient of variation (CV) of degree distribution This is a measure of variability in the degree distribution of the community network, *x*. Previous network studies have shown that increasing variability in the degree distribution is associated with a greater likelihood of observing outbreaks (5). The CV of the degree distribution of x, which is denoted $D(x)$ is calculated as follows:

$$CV(D(x)) = SD(D(x))/E(D(x))$$

where SD(D(x)) and E(D(x)) are the standard deviation and expected value of D(x) , respectively. Increasing values of the CV indicate greater variability. A distribution with a CV of 1 is exponentially distributed, whereas if a network's degree distribution had a CV = 0, all nodes would have the same degree. This measure has the advantage of being dimensionless, so it can be used to directly compare networks of differing average degree.

Average shortest path length. This measures average number of network hops needed to connect any two nodes via the shortest path between them. It measures the overall accessibility of nodes to one another in the network. As this value becomes smaller, it is easier for infection to pass between two arbitrarily selected nodes. Longer path lengths between nodes make the slow propagation of infections across individuals more likely, while also increasing the likelihood of stochastic extinction and smaller outbreaks. We denote this statistic as *P(x)*.

Clustering coefficient. The clustering coefficient of X, denoted *C(x)*, measures the likelihood that open triangles of vertices, denoted open triples, in the community network will close and form a closed triplet. C(x) is measured as:

$$C(x) = \#\ of\ closed\ triplets\ /\ (\#\ of\ closed\ triplets + \#\ of\ open\ triples)$$

The clustering coefficient is best known for its role in measuring the small-world effect in social networks (39). Heuristically, it measures the redundancy of contacts amongst individuals. In social networks, greater clustering is often thought to imply greater social cohesion (27), whereas in epidemiological applications, clustering is thought of mainly for its role in inhibiting outbreaks (28, 40).

Simulating outbreaks

We simulate outbreaks from the community model using the epidemic percolation graph (EPG) framework of Miller (41). Using this approach, one can rapidly simulate the final size of outbreaks on networks with heterogeneous infectivities, which allows exploration of a large number of outbreak scenarios. All simulated outbreaks begin with a point source event, the size of which is drawn from a geometric distribution with mean $\overline{Q}$, in which a group of individuals is selected uniformly at random from the population and infected. The geometric distribution is chosen because it is a discrete probability distribution which a moderate amount of right-tail variability, analogous to an exponential distribution. This is in keeping with our heuristic understanding of point-source events as typically small-to-moderate sized with occassional large events. For each combination of model parameters listed in Table 3.1 (i.e., disease and social

network parameters) , we perform $5 \cdot 10^4$ EPG simulations. A detailed description of the implementation of this simulation framework is outlined in the supplementary materials.

Measures of simulated outbreaks

Average outbreak size. We measure the risk to the community from an initial point-source event by the number of cases caused in the course of the outbreak. In the outbreak simulations presented here, we present this in terms of the average number of secondary cases corresponding to a set of network and disease parameters. We measure the number of secondary cases as the number of *symptomatic* cases, as asymptomatic cases are unlikely to be observed in the course of an outbreak. This also facilitates comparison between simulations with varying levels of the proportion of fully asymptomatic infections, τ. We denote the average *symptomatic* outbreak size for a given set of network parameters, θ, and disease parameters, Φ, as $\mu_S(\Phi,\theta)$.

Coefficient of variation (CV) of outbreak size distribution. The CV of the distribution of outbreak sizes corresponding to a set of network and disease parameters reflects the degree of heterogeneity in outbreak sizes corresponding to that set of parameters. Because the CV is dimensionless, using it allows us to address the question of how model parameters impact heterogeneity in outbreak size across a wide range of average outbreak sizes. Similar to average size, above, we denote the CV of outbreak sizes as $CV_S(\Phi,\theta)$.

Parameterizing outbreak simulations

Wherever possible, disease parameter values used in simulation runs are obtained from published estimates of norovirus natural history and infectivity. For the progression of disease states after infection, we follow the parameterization of norovirus infection history used by Milbrath et al. (10), which is reproduced in Table 3.1 and represented graphically in Figure 3.1. We parameterized household transmission using the parameters obtained by Zelner et al. (12) in their analysis of household outbreaks of norovirus. Since no published estimates exist for the rate of transmission in the community, we perform simulations over a range of plausible values. Similarly, social network parameters for age-based mixing are derived from empirical time-use data. Other network parameters are picked to reflect realistic social behavior and to highlight the impact this behavior has on network structure and risk. Where reliable estimates of social network parameters are not available, we perform sensitivity anlysis to ensure the robustness of our results to variability in these parameters.

Household Transmission

Previous work has shown that estimates of the household transmission rate of norovirus are sensitive to the proportion of the population that is never symptomatic, τ (12). This value has been estimated to range from 15 to 50% (42). Zelner et al. (12) estimated that when $\tau = 0$, the daily, density-dependent household rate of transmission, $\beta_{IS} = 0.14$. For every 10% increase in τ, the value of the density-dependent transmission parameter is estimated to increase by an additional 25%, so that when $\tau = 0.5$, $\beta_{IS} =$

0.29. This increase in the proportion of individuals who are never symptomatic and the related increase in β_{IS} should not strongly impact within-household transmission dynamics at the beginning of an outbreak (12), but may decrease the visibility of connections between norovirus outbreak events. When simulating outbreaks, we use this relationship between β_{IS} and τ to vary the transmission rate and proportion of the population that is never symptomatic in an empirically-grounded fashion.

Point Source Events

There are few published reports that quantify the size and distribution of these kinds of point-source events in the context of food-borne and nosocomial outbreaks. Lynch et al. (13) provide some basic information on the number of norovirus outbreaks that were observed under the Foodnet program from 1998-2002 (657) and the total number of cases reported from these outbreaks (27,171), which yields an average of about 41 cases per reported event. Because there are likely many smaller events that are not recognized as originating from a common source, and up to 50% of individuals infected with norovirus, at a point-source or otherwise, are likely fully asymptomatic (43), there is reason to believe that this figure both overestimates the average size of the typical point-source event, and underestimates the risk stemming from them.

The exact number of individuals in the United States who work as food servers is unclear. However, the U.S. Bureau of Labor Statistics (BLS) does collect figures on the number of individuals working in food service and preparation management, and estimates that about 981,000 individuals were employed in this manner in 2008. We base our rough estimate on the assumption that each of these individuals supervises on average

2 people, resulting in a total of about 3 million individuals working directly with prepared food, or about 1 percent of the total U.S. population of 300 million. We use this as our estimate of ζ, the proportion of individuals in the population who are potential point-source hubs. The actual number is likely higher, but use of this figure ensures a conservative representation of the role of food-handlers in seeding point-source events.

Analysis of outbreak simulation data

Regression tree analysis. We analyze outbreak simulation results using several statistical tools. We perform non-parametric analysis of simulated networks and outbreaks using regression trees (RT), see e.g., (44). The RT methodology has the advantage of being non-parametric, particularly in that it does not make strong assumptions about the linearity of predictors and the distribution of errors. A regression tree is constructed by recursively partitioning the dataset, via binary splits, into progressively more homogeneous groups of cases. Each split in the tree corresponds to a logical condition. The set of simulated outbreaks that satisfy the logical condition are grouped under the left-hand side of the split, while the right-hand side of the split contains those cases for which the condition is false. The number of splits in the full regression tree is constrained by a parameter that imposes an increasing cost to the goodness-of-fit for a greater number of splits in the tree. In practice, this parameter controls the tradeoff between the granularity and parsimony in the resulting tree. In our analyses, we fix this parameter at 0.005. The RT approach also has the considerable advantage of implicitly including interactions between variables, which are represented

by the combination of logical conditions. Terminal nodes of the regression tree, i.e. cases grouped underneath the final split on a subtree, are labeled with the average value of the cases in this set.

Log-linear regression. We also use log-linear regression models to estimate the association between model parameters and the average size of outbreaks simulated from the epidemic percolation graph (EPG) community outbreak model (40). We interpret the exponentiated values of the coefficients from these log-linear models as the factor by which a one-unit increase in the model parameter increases or suppresses expected outbreak size, so that the exponentiated model intercept is multiplied by the exponentiated model coefficient to obtain the expected outbreak size.

Mediation Analysis. Because the behavioral parameters of the ERG model impact outbreak risk through their effect on structural features of the network, we also use log-linear models to perform mediation analysis. In these models, we assess the indirect effects of the behavioral parameters operating through structural characteristics of the community network. For example, we are interested in how age-structured mixing impacts outbreak risk via its effect on heterogeneity in the degree distribution, as measured coefficient of variation (CV) of the community network degree distribution. Following the logic of Baron and Kenny (45), we fit two log-linear regression models, one predicting the natural log of the expected outbreak size with the model parameter for, e.g. age structured mixing θ_A, as the independent variable:

$$\log(Y) = \alpha + \beta_A \theta_A + \varepsilon$$

Where $\log(Y)$ is the natural log of the average outbreak size. The second model predicts outbreak size as a function of both θ_A and $CV_D(X)$, so:

$$\log(Y) = \alpha + \beta_A^{'}\theta_A + \beta_V CV_D(X) + \varepsilon$$

We quantify the mediation effect, M, as the exponentiated difference in the values of β_A and $\beta_A^{'}$, so that:

$$M = \exp(\beta_A - \beta_A^{'})$$

This value can be interpreted as the ratio between the coefficients for the network model parameter when the structural mediator is and is not included in the model. We obtain 95% confidence intervals and *p*-values by bootstrapping the sampling distribution of *M*.

When assessing mediation by a variable, such as the average path length, which may be affected by multiple network model parameters, we adjust for this additional relationship in both models. For instance, the average length of the path separating any two arbitrarily selected nodes typically shortens with increasing average degree. So, when assessing how a model parameter, such as the one governing age-structured mixing (θ_A), impacts outbreak size through its effect on average path length, we include average degree of the community network as a covariate in both models, and then calculate *M* in the manner described above.

Results

Figure 3.3 shows the results of a regression tree analysis where average outbreak size is the dependent variable and the disease (in Φ) and social parameters (in θ) used in simulations, are predictors. The figure shows that for values of $\beta_{IA} < 0.07$, average

outbreak size is small, at 23.4 symptomatic cases. At a slightly higher level, $\beta_{IA} = 0.08$, average outbreak size grows to 137.3 cases, and is not strongly impacted by other social network or disease parameters. When $0.09 \leq \beta_{IA} \leq 0.1$, i.e., β_{IA} is large and asymptomatic cases are fairly infectious, outbreak size is also modulated by age-structured mixing and sociability. When $\theta_S = 1$ and $\theta_A = 0$, i.e. there is variation in individual sociability but no age-structured contacts, average outbreak size increases to 594.3 cases. When $\theta_S = 0$ and $\theta_A = 1$, average outbreak size grows again to 662.6. Finally, with both of these factors, the average number of symptomatic infections grows to 987 out of 10,000 susceptible individuals.

Table 3.2 shows results from several log-linear regression models where average outbreak size is the dependent variable and ERG model and disease parameters are predictors. The first column of the table shows a model that includes all values of the asymptomatic transmission parameter, β_{IA}. Outbreak size is strongly dependent on this parameter, and a 0.01 increase results in a 50% increase in outbreak size. Although the effect of a 0.01 increase in the daily rate of point source events (PSE), λ, is small (1.02), when this value increases over the full simulated range, from 0.01 to 0.1 events / day, the amplifying effect of point source events on outbreak size is considerable (1.21). Including either age-structured mixing or variation in individual sociability also results in larger outbreaks (θ_A: 1.65, θ_S:1.56). When both are included, the combined effect on outbreak size is somewhat smaller than that of the product of their coefficients ($\theta_A * \theta_S$: 0.83). Strong within-neighborhood clustering also suppresses outbreak sizes (θ_N :0.93). This effect becomes stronger when the community network is impacted by age-structured mixing and/or variation in individual sociability ($\theta_A * \theta_N$: 0.92; $\theta_S * \theta_N$:0.91).

The remaining columns of Table 3.2 show results from log-linear models predicting average outbreak size, stratified by levels of β_{IA}. The intercept for these models is the expected outbreak size in a fully random network for the selected value of β_{IA}, when all infections are symptomatic (i.e., $\tau = 0$) and there are no point-source events (i.e., $\lambda = 0$). When $\beta_{IA} = 0.01$ (Column 2), so that the average asymptomatic individual has a 1% chance of making a new case in the community on every day she is asymptomatically infectious, a 10% increase in the proportion of infections that are fully asymptomatic dampens outbreak size (τ: 0.92). Otherwise, when $\beta_{IA} = 0.01$, the remaining coefficients are near 1, indicating no effect of the parameter on outbreak size. This is to be expected because outbreaks of considerable size are uncommon at this low value of β_{IA}. Column 3 of table 3.2 shows average outbreak size when $\beta_{IA} = 0.05$, with results quantitatively similar to those in Column 1 for all values of β_{IA} (Column 1). Column 4 shows outbreak size for the largest simulated value of β_{IA}=0.1. Here, the baseline expected outbreak size is larger (170.64 cases) than in the previous two conditions and increasing the proportion of fully asymptomatic infections amplifies rather than suppresses outbreak size (τ: 1.05). As before, the independent amplifying effects of age-structured mixing and variability in individual infectiousness are large (θ_A:3.45, θ_S :3.45), but their joint effect is smaller than the product of their individual effects, so that when $\theta_A = 1$ and $\theta_S = 1$, the expected outbreak size is 960.7, or 5.6 times larger than the baseline outbreak size. When $\beta_{IA} = 0.1$, a one-unit increase in average degree above 10 is associated with a statistically significant increase in overall risk ($\bar{k}$: 1.05). The relative strength of the outbreak-size reducing effect of neighborhood clustering is also greater than in the first three models (θ_N :0.83). Figure 3.4 illustrates the joint effects of the

asymptomatic transmission parameter, β_{IA}, on variation in individual sociability (θ_S) and age-structured mixing (θ_A) on average outbreak size.

Table 3.3 presents results from a mediation analysis assessing the indirect effects on average outbreak size of the behavioral parameters of the exponential random graph (ERG) model as they operate through structural characteristics of the community network. The first column shows the effect of the ERG parameters operating through the coefficient of variation of the community network degree distribution. For all simulated values of β_{IA}, the effect of age-structured mixing on outbreak size operates in part through its effect on the CV of the degree distribution (M = 1.14, 95% CI = [1.12, 1.15]). When β_{IA} = 0.05, this effect is slightly weaker (M = 1.07, 95% CI = [1.068, 1.079]). It becomes stronger when β_{IA} = 0.1 (M = 1.35, 95% CI = [1.33, 1.38]). Variation in sociability also operates on risk for all values of β_{IA} by inducing variability in the degree distribution (M = 1.35, 95% CI = [1.33, 1.38]). This effect is particularly pronounced when β_{IA} = 0.1 (M = 4.44, 95% CI = [4.40, 4.73]). By contrast, the outbreak-dampening effect of increasing neighborhood clustering, θ_N, is suppressed by increasing variability in the degree distribution (M = 0.92, 95% CI = [0.91, 0.93]) and decreasing average path length (M = 0.88, 95% CI = [0.87, 0.89]). The effect of relationship clustering, θ_R, on outbreak size is weakly mediated by the CV of the degree distribution (M = 1.01, 95% CI = [1.005, 1.017]), instead operating through the clustering coefficient of the graph (M = 2.19, 95% CI = [2.01, 2.39]).

Figure 3.5 illustrates the relationship between disease and social parameters and variability in outbreak size. The figure shows a regression tree where the dependent

variable is the coefficient of variation (CV) of the distribution of outbreak sizes corresponding to a set of model parameters, and the model parameters are predictors. The figure shows that when the asymptomatic transmission parameter, β_{IA}, is small, from 0.01 to 0.03 cases/infected per day, the distribution of outbreak sizes is approximately exponential, indicating a moderate amount of variability. When $0.04 \le \beta_{IA} \le 0.06$, variability in outbreak sizes increases and is grows further with age-structured mixing and variation in individual sociability, so that when $\theta_S = 1$ and $\theta_A = 1$, CV = 1.5. By contrast, when $0.06 < \beta_{IA} < 0.09$, increasing the proportion of asymptomatic cases ($\tau \ge 0.2$), and with this the rate of household transmission, decreases variability in outbreak sizes. At the top end of the range of the asymptomatic transmission parameter values used in simulations, $0.09 \le \beta_{IA} \le 0.10$, outbreak size variability is lowest when proportion of asymptomatic infections is large ($\tau \ge 0.2$) and age-structured mixing and variation in individual sociability are present (e.g., $\theta_S = 1, \theta_A = 1$; CV = 0.96).

Discussion

Our results show that community networks with age-structured contact patterns are more likely to have large outbreaks than random networks. The strength of this effect grows with increasing asymptomatic infectivity (β_{IA}). Our results also demonstrate that exponentially distributed variation in individual sociability has an effect on outbreak size that is similar in magnitude to that of age-structured mixing. This effect also grows with an increasing strength of asymptomatic transmission. As β_{IA} increases, the joint effect of age-structured mixing and variation in sociability on outbreak size becomes progressively smaller than the product of their individual effects. Since average outbreak sizes, even for

the largest simulated level of asymptomatic infectivity, $\beta_{IA} = 0.1$, do not exhaust the susceptible population (~1000 cases), this indicates that the mechanisms by which age-structured mixing and sociability impact epidemic outcomes are overlapping.

Our results also show the extent to which point-source events (PSE) drive outbreak size in our community outbreak model. We find that increasing the rate of point-source events consistently amplifies outbreak size, although not as dramatically as age-structured mixing or exponential variation in sociability. The most notable effect of point-source spreading is at the low end of the range of simulated asymptomatic infectivity ($\beta_{IA} = 0.01$), where a high point-source event rate, $\lambda = 0.1$, increases average outbreak size by 10%, whereas other factors have either very small or non-significant effects on outbreak size.

Our mediation analysis exposes how the behavioral mechanisms in the community model impact outbreak size through their effects on elements of network structure. We find that some of the outbreak amplifying effect of age-structured mixing and much of the effect of individual sociability can be explained by their effects on increasing variability in the degree distribution of the community network. Age-structured mixing and sociability also amplify outbreak size by shortening the average path length between arbitrarily selected nodes in the community network, which makes outbreaks more likely. Neither of these factors is strongly mediated by the clustering coefficient of the community network.

In contrast to the behavioral features that increase variability in contact and average outbreak size, we find that increasing neighborhood clustering results in smaller

average outbreak size. Our mediation analysis shows that this effect is attenuated in situations in which other factors increase variation in degree, indicating that this factor suppresses outbreaks partially by homogenizing the distribution of degree and lengthening the paths connecting arbitrarily selected pairs of nodes. The fact that the strength of this suppressive effect grows with the inclusion of age-structured mixing and variation in individual sociability exemplifies why it is important to account for the effect of potentially countervailing social behaviors.

Our model is also able to address the impact of social and disease parameters on heterogeneity in outbreak sizes. Our regression tree analysis of the drivers of outbreak size variability shows a non-linear relationship between the rate of asymptomatic transmission and the coefficient of variation (CV) of the distribution of outbreak sizes. We see that the CV of the distribution of outbreak sizes is smallest when asymptomatic infectiousness is low. This is because, when β_{IA} is small, most of the originating point-source outbreaks result in few or no community cases and most secondary cases come from household contacts and point-source events. When asymptomatic infectivity is increased into a more moderate range, $0.06 \leq \beta_{IA} \leq 0.08$ cases/infected per day, is when outbreak size variation is greatest. Age-structured mixing and exponentially distributed variation in sociability modulate this heterogeneity in outbreak size. When both of these factors are present and the proportion of fully asymptomatic infections is low ($\tau \leq 0.2$), we see the greatest overall variability in outbreak sizes (CV = 1.7). At the highest levels of simulated asymptomatic infectivity, $\beta_{IA} \geq 0.09$, this effect is reversed, so that outbreak simulations including age-structured mixing, variation in sociability and low levels of fully asymptomatic infection have less variability in final size (CV = 1.2), although the

expected magnitude of these outbreaks is larger. Results for both average outbreak size and variability in size indicate that the most realistic parameter ranges for norovirus asymptomatic infectivity may be in the moderate range from 0.06 to 0.08, where outbreak events are moderately sized (~137 cases on average). However, there is wide variability in the distribution of these events, resulting in a many small outbreaks and a few large ones.

One limitation of our results regards the relationship clustering mechanism in the community network model. From the results presented here, the effect of relationship clustering mechanism in the community contact network model on outbreak size appears to be minimal, which runs counter to existing findings showing that such clustering tends to suppress outbreaks (46). Analysis of the structure of sampled networks indicates that the clustering coefficient of sampled networks, *C(x)*, even for the largest sampled level of the relationship clustering parameter, θ_R= 10.0, has a minimal impact on the overall clustering coefficient of the graph, as the greatest sampled value of C(x) = 0.04. Test runs using networks with fewer (1000) nodes yielded significantly larger values of the clustering coefficient (~0.3) when θ_R= 10.0. For networks with more nodes, such as the ones used here, greater values of θ_R may be necessary to obtain significant clustering. The lack of clustering in our simulated networks may also be the result of the choice of the value for the rate of decay in the shared partner weighting, η. As network size grows, the raw number of shared partnerships grows as well, and the choice of η = 0.5 may impose too steep of a penalty on the weight an increasing number of shared relationships contributes to the calculation of the network statistic $M_R(x)$. Consequently, additional

simulation and sensitivity analysis are necessary to understand the true role that relationship clustering can play in the outbreak dynamics observed in this model.

Our results on the heterogeneity of outbreak sizes indicates that age-structured mixing and exponential variation in individual sociability make outbreak size less predictable for moderate levels of β_{IA}. This suggests that outbreaks with these parameters are also more likely to have the explosive and unpredictable characteristics associated with norovirus epidemiology. Additional discrete-time simulations that assess the evolution of outbreaks over time are needed to fully understand the extent to which the mechanisms discussed here impact explosiveness as well as heterogeneity in norovirus outbreaks.

Conclusions

By including variation in social behavior and heterogeneity in duration of infection in a single, parametric framework, we are able to shed light on the specific ways social behavior and disease processes independently, and interactively, influence epidemic outcomes. Our results show how the exponential random graph (ERG) framework can incorporate multiple types of social behavior into outbreak simulation models in a reasonably transparent way. In the simulations presented here, we selected model parameters to represent largely non-overlapping types of social behavior. The task of specifying an ERG model to be fit to observed social networks is more difficult than using one in simulation (33), but there are large potential benefits from the integration of empirical data sources on social behavior and disease processes that this framework allows.

An alternative approach to inference with this type of model is to utilize the plug-and-play approach, e.g., (47, 48, 49). In this approach, it is only required that it is possible to simulate sample paths, in this case time series of symptomatic cases, from the theoretical model. Using these techniques, it is possible to estimate social parameters, such as the strength of within-neighborhood clustering, from observed outbreak time series, which is an important new direction for both social scientists and epidemiologists.

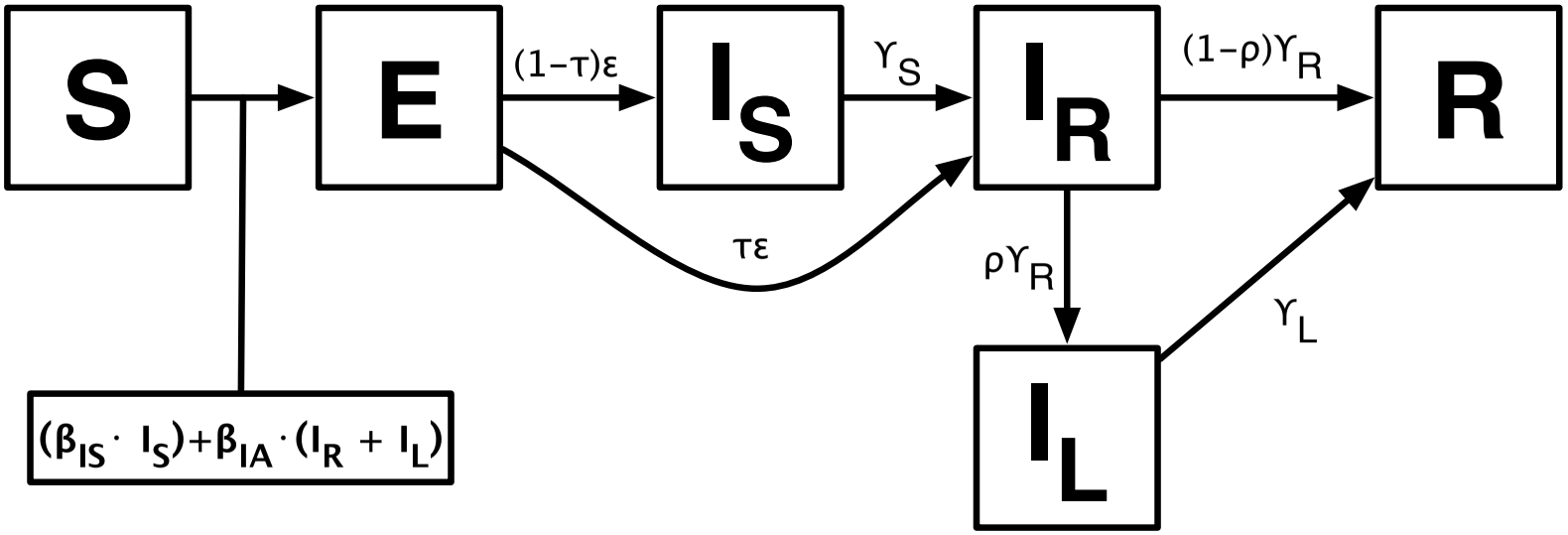

Figure 3.1. Stages of Norovirus (NoV) infection. States of infection are indicated by boxes. Arrows indicate transitions between states and are labelled with parameters for average rates of state transition. Individuals are infected at a rate proportional to the number of symptomatic and asymptomatically infectious individuals they are exposed to in their home and community networks. β_{IS} and β_{IA} are the rates at which symptomatic and asymptomatic individuals infect susceptible ones, respectively. $\beta_{IA} \ll \beta_{IS}$. Immediately after infection, susceptible (**S**) individuals enter the exposed/latent state (**E**). Individuals spend, on average, $1/\varepsilon$ days in the latent state before progressing to either the symptomatic infectious (I_S) or regular-duration asymptomatic (I_R) states with probabilities $(1-\tau)$ and τ, respectively. Individuals remain in I_S for an average of $1/\Upsilon_S$ days before entering the regular asymptomatic phase I_R. After approx. $1/\Upsilon_R$ days, the majority of individuals, represented by $(1-\rho)$, become recovered/immune (**R**) and are no longer infectious. A small fraction of asymptomatic individuals, represented by ρ, progress from I_R to an extended asymptomatic shedding phase, denoted I_L.

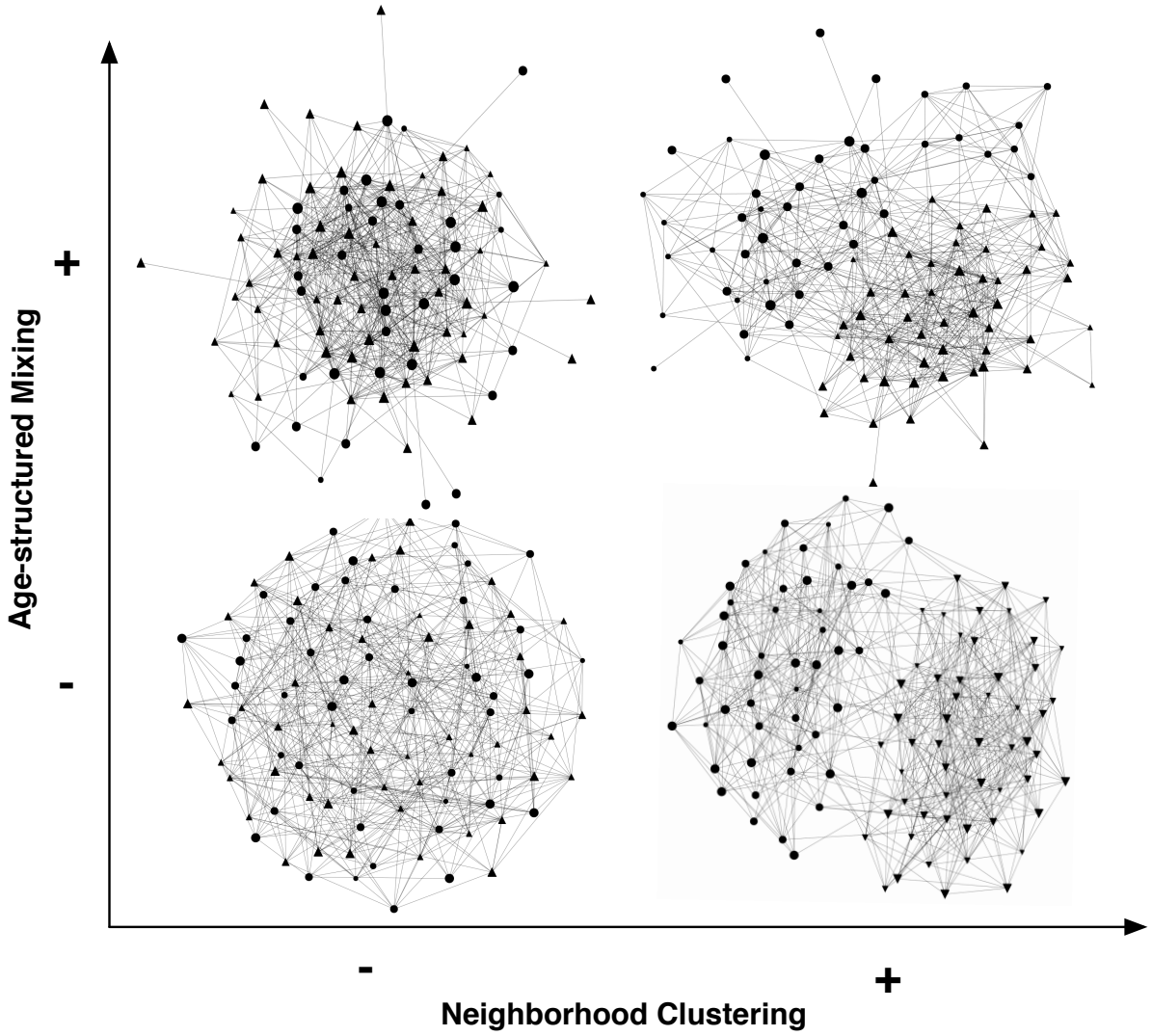

Figure 3.2. Network structures resulting from different types of social mixing. The figure illustrates a community consisting of 100 individuals with average degree of 12. The community is partitioned into two equally sized neighborhoods, represented by triangular and circular nodes. Larger node size indicates increasing age. The network in the bottom-left quadrant (-,-) is and Erdos-Reny random graph, in which connections in the community are distributed without respect to node attributes. In the bottom right quadrant (+,-) individuals preferentially form network connections with individuals from the same neighborhood. In the network in the top-left quadrant, (-,+) individuals make relationships preferentially based on age, according to the age-dependent time-use data of Zagheni et al (2006). In the top-right quadrant, network connections form as a function of both age and neighborhood (+,+).

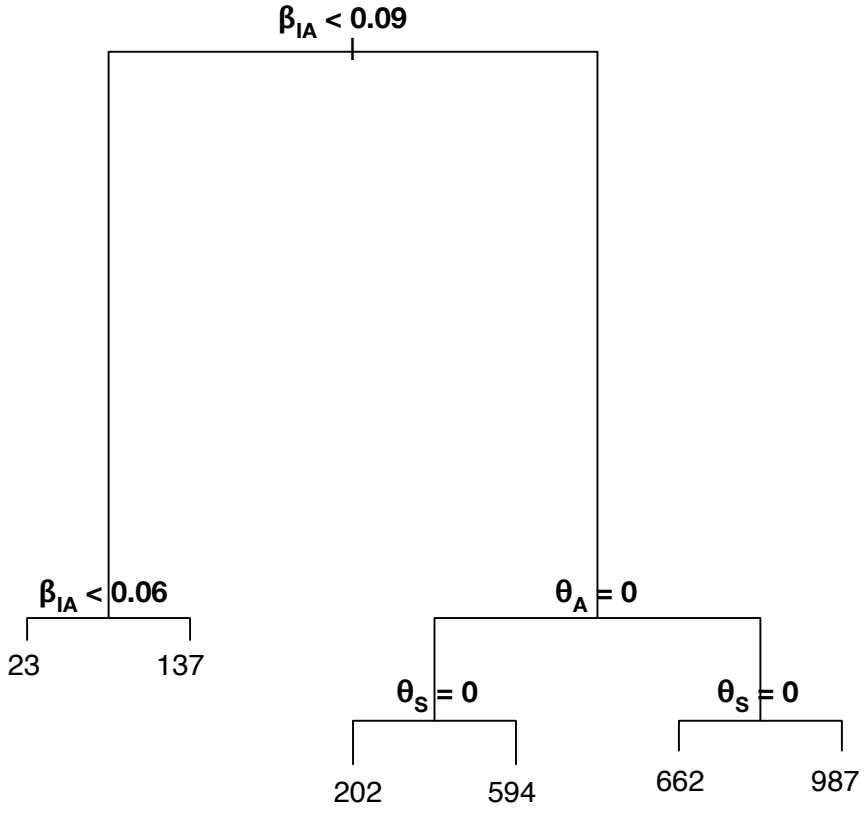

Figure 3.3. Regression tree showing average outbreak size as a function of model parameters. Splits in the tree are labeled by a logical condition. Simulation runs satisfying this condition are contained in the subtree or node underneath the left-hand side of the split. Runs that do not satisfy the logical condition are underneath the right-hand side of the split. Subtree arm lengths represent the amount of variance explained by the split. Terminal nodes are labeled by the average outbreak size of simulation runs satisfying all of the conditions leading to that node. Parameter values and definitions are in Table 3.1.

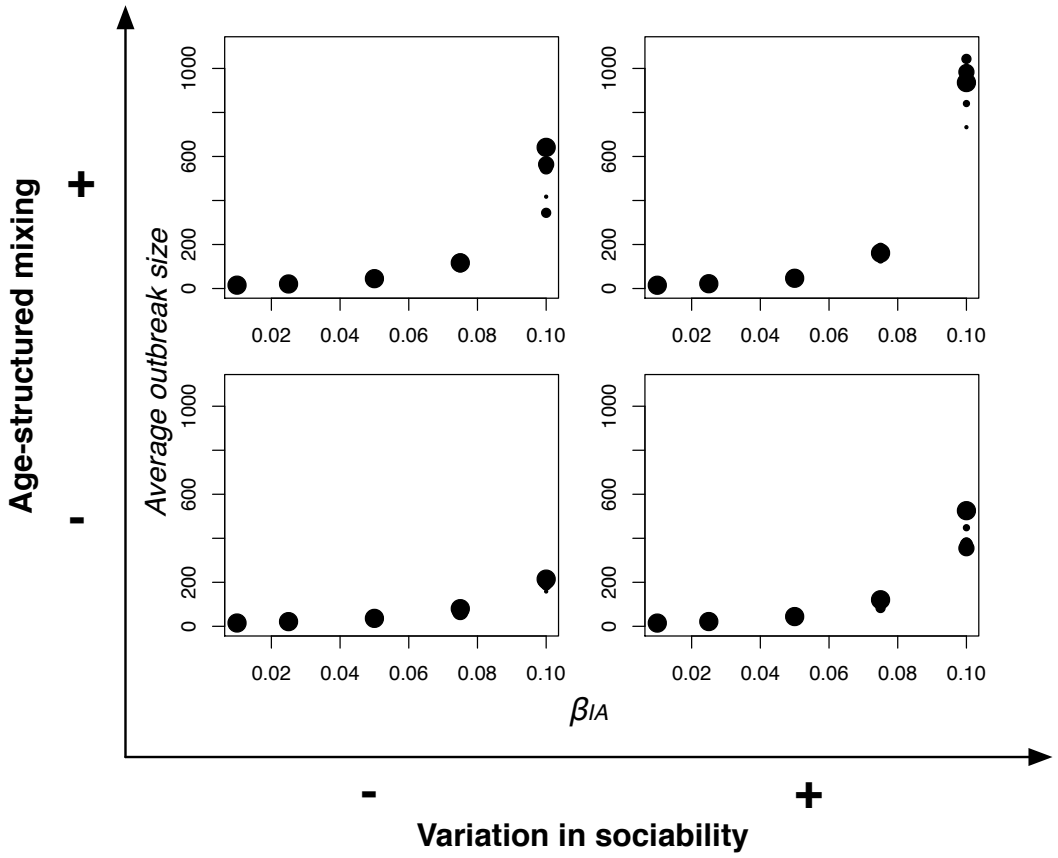

Figure 3.4. Relationship between transmission and social network parameters and average outbreak size. The asymptomatic transmission parameter, β_{IA}, is on the x-axis of the inner panels and the y-axis is the average number of symptomatic secondary cases resulting from an outbreak after a point-source spreading event. Increasing dot size in the plot indicates greater values of the point-source spreader attack rate, λ. The bottom-left panel (-,-) represents outbreaks in a random network. The top-left panel (-,+) shows outbreaks in a network with strong age-structured mixing ($\theta_A = 20.0$), and the bottom-right panel (+,-) shows outbreaks in networks with heterogeneous individual sociability. In the top-right panel (+,+) outbreaks occur on a network conditioned by both factors. In all plots, the value of τ, the proportion of individuals who are never symptomatic is fixed at 10%, and the value of the relationship and neighborhood clustering coefficients, θ_R and θ_N, respectively, are fixed at 1 to have no effect on network structure.

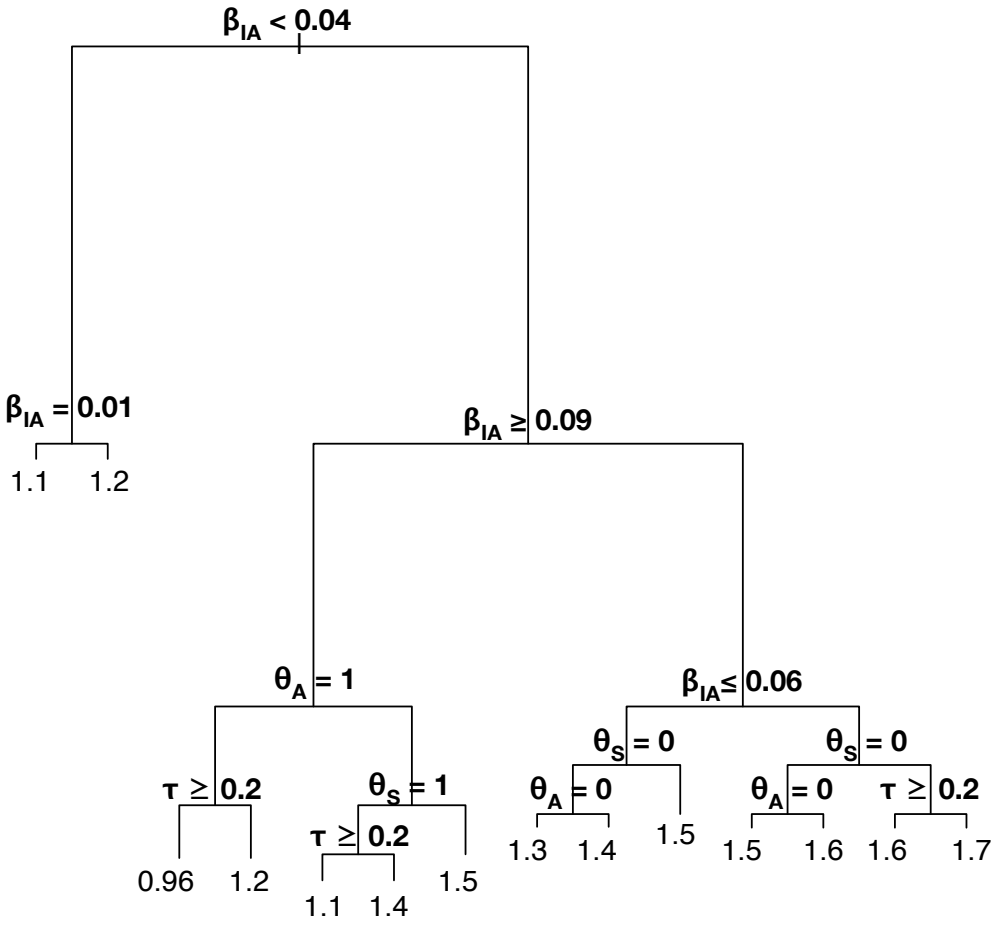

Figure 3.5. Regression tree of coefficient of variation (CV) of outbreak sizes as a function of community model parameters. The figure shows a regression tree where the terminal nodes are labeled with the average coefficient of variation for outbreak simulations corresponding to the logical conditions above the node. Definitions and notation for parameter values are in Table 3.1.

Parameter	Definition	Values	Units	Source
Transmission parameters				
β_{IS}	Household transmission rate for symptomatic infectious individuals	0.07	(infections / contact) / day	Zelner et al. (2010)
β_{IA}	Household and community transmission rate for asymptomatic infectious individuals	[0.01, ..., 0.1]	Avg. infections / day	See text
Disease natural history				
τ	Proportion of infections that are fully asymptomatic	[0.0, ..., 0.5]	-	Milbrath et al. (2010)
ρ	Proportion long asymptomatic shedders	0.05	-	Milbrath et al. (2010)
$1/\varepsilon$	Average duration of latency period	1.5	days	Zelner et al. (2010)
ε_S	Variability in gamma distribution of latency duration (shape parameter)	4	-	Zelner et al. (2010)
$1/\gamma_S$	Average duration of symptomatic infectious period	2	days	Zelner et al. (2010)
$1/\gamma_R$	Average duration of regular asymptomatic infectious phase	10	days	Milbrath et al. (2010)
$1/\gamma_L$	Average duration of long asymptomatic infectious phase	100	days	Milbrath et al. (2010)
Community transmission network				
$\overline{k}$	Average number of community contacts	[8, 10, 12]	contacts	Bansal et al. (2007)
θ_A	Presence or absence of age-based mixing	[0, 1]	-	See text
Ω	Intensity of mixing by age	See text	odds ratio	Zagheni et al. (2008)
θ_S	Presence or absence of variability in individual sociability	[0, 1]	-	See text
$\overline{S}$	Average sociability of individuals in network	10	-	See text
θ_M	Relative strength of within-neighborhood assortativity	[1.0, ..., 20.0]	odds ratio	See text
h	Number of neighborhoods	4	#	See text
θ_P	Intensity of network partner sharing	[1.0, ..., 10.0]	odds ratio	See text
η	Rate of geometric decay in partner sharing distribution	0.5	-	Kitts et al. (2009)
Point source events				
ζ	Proportion of individuals who are food servers	0.01	-	See text
λ	PSE attack rate for asymptomatic food servers	[0.01, 0.1]	1/day	See text
$\overline{Q}$	Average size of point-source event	10	individuals	See text

Table 3.1. Parameter values for outbreak simulations. The table presents and defines parameters for the disease process and social networks used in outbreak simulations. Parameter values presented in brackets are used in sensitivity analysis. Parameter values are obtained from published sources unless otherwise noted.

	All β_{IA}	β_{IA} = 0.01	β_{IA} = 0.05	β_{IA} = 0.1
Intercept	**4.69**	**14.70**	**32.16**	**170.64**
Disease Parameters				
Asymptomatic infectivity, β_{IA}	**1.50**	-	-	-
Proportion fully asymptomatic, τ	**0.97**	**0.92**	**0.95**	**1.05**
Point-source event rate, λ	**1.02**	**1.01**	**1.01**	**1.02**
Network Parameters				
Age, θ_A	**1.65**	**1.02**	**1.35**	**3.45**
Sociability, θ_S	**1.56**	**1.01**	**1.27**	**3.20**
Neighborhood clustering, θ_N	**0.93**	**0.99**	**0.97**	**0.83**
Relationship clustering, θ_R	1.00	0.99	1.00	**1.02**
Average degree	**1.01**	**0.99**	**1.00**	**1.05**
Network Parameter Interactions				
$\theta_A * \theta_S$	**0.83**	**0.99**	**0.95**	**0.51**
$\theta_A * \theta_R$	**0.98**	1.00	**0.98**	**0.97**
$\theta_A * \theta_N$	**0.92**	0.99	**0.92**	**0.91**
$\theta_S * \theta_R$	0.98	1.00	**0.98**	**0.96**
$\theta_S * \theta_N$	**0.91**	1.00	**0.98**	**0.85**
$\theta_R * \theta_N$	**1.02**	1.01	**1.02**	**1.03**
Goodness of fit (R^2)	0.93	0.92	0.82	0.92

Table 3.2. Amplification of average outbreak size. The table shows results from a log-linear regression model where the natural logarithm of average outbreak size is the dependent variable and community model parameters corresponding to these outcomes are predictors. The table presents exponentiated values of the regression coefficients, which are interpreted as the factor by which the expected outbreak size grows with a 1-unit increase in the predictor. Model parameters have been re-scaled to facilitate interpretation: a 1-unit change in regression parameters for β_{IA} and λ represent a .01 change in the corresponding simulation model parameters. A 1-unit change in the regression coefficient for τ represents a 10% increase in the proportion of individuals who are never symptomatic. The coefficient for average degree reflects deviations from an average degree of 10. Coefficients for θ_N and θ_R reflect the difference between maximum and minimum values of the parameter used in simulations. Column 1 shows results for all values of β_{IA}. The remaining columns show results when the value of β_{IA} is fixed at 0.01, 0.05,and 0.1, respectively.

	Degree CV (95% CI)	Avg. Path Length (95%CI)	Clustering Coef (95% CI)
All β_{IA}			
Age, θ_A	**1.14 (1.12, 1.15)**	**1.12 (1.11, 1.14)**	**1.002 (1.001, 1.003)**
Sociability, θ_S	**1.07 (0.95, 1.18)**	**1.63 (1.56, 1.70)**	1.005 (0.998, 1.013)
Neighborhood clustering, θ_N	**0.92 (0.91, 0.93)**	**0.88 (0.87, 0.89)**	**0.991 (0.987, 0.994)**
Relationship clustering, θ_R	**1.01 (1.005, 1.017)**	0.0 (0.99, 1.004)	**2.19 (2.01, 2.39)**
β_{IA}=0.05			
Age, θ_A	**1.07 (1.068, 1.079)**	**1.07 (1.066, 1.077)**	1.001 (1.0, 1.002)
Sociability, θ_S	**1.50 (1.48, 1.51)**	**1.36 (1.35, 1.37)**	1.003 (1.0, 1.005)
Neighborhood clustering, θ_N	**0.96 (0.94, 0.96)**	**0.93 (0.92, 0.94)**	**0.995 (0.993, 0.997)**
Relationship clustering, θ_R	1.006 (0.999, 1.014)	0.0 (0.99, 1.01)	**1.60 (1.55, 1.64)**
β_{IA}=0.1			
Age, θ_A	**1.35 (1.33, 1.38)**	**1.34 (1.32, 1.37)**	1.005 (0.999, 1.01)
Sociability, θ_S	**4.44 (4.40, 4.73)**	**3.25 (3.18, 3.33)**	**1.02 (1.01, 1.02)**
Neighborhood clustering, θ_N	**0.82 (0.79, 0.85)**	**0.71 (0.70, 0.74)**	**0.98 (0.97, 0.982)**
Relationship clustering, θ_R	1.03 (0.99, 1.05)	0.0 (0.97, 1.03)	**7.08 (6.61, 7.61)**

Table 3.3. Mediation of exponential random graph (ERG) network model parameters by structural features of the community network. Parameters of the ERG model, which represent social behavior, impact outbreak size through their effect on structural characteristics of the community network. The table shows results of a mediation analysis, where these behavioral parameters are conceptualized as distal factors which impact outbreak size through their effect on three characteristics of the community contact network: 1) the coefficient of variation (CV) of its degree distribution, 2) the average length of the path connecting arbitrarily selected nodes, and 3) the global clustering coefficient. The disease outcome is the natural log of the expected outbreak size. We present results for all levels of the asymptomatic transmission parameter, $\boldsymbol{\beta_{IA}}$ as well as results stratified by $\boldsymbol{\beta_{IA}}$ = 0.05 & 0.1. Mediation effects, *M*, are calculated as the ratio of the log-linear regression coefficients associated with the ERG parameter in two log-linear models: 1) with only the distal variable as a predictor, where the coefficient for the distal variable is denoted *b;* and 2) one adjusted for the distal variable and the mediator, where the coefficient of the distal variable is denoted *b'*. So, M = exp(b - b'). 95% confidence intervals are obtained by inspecting the quantiles of the bootstrapped sampling distribution of M. Bolded items are statistically significant at the $p < 0.05$ level.

References

1. CDC. Preliminary FoodNet Data on the incidence of infection with pathogens transmitted commonly through food--10 States, 2008. MMWR Morb Mortal Wkly Rep 2009;58:333-7.

2. Longini I, Nizam A, Xu S, et al. Containing Pandemic Influenza at the Source. Science 2005;309:1083-7.

3. Anderson R, May R. Infectious Diseases of Humans: Dynamics and Control. 1992.

4. Lloyd-Smith J, Schreiber SJ, Kopp PE, et al. Superspreading and the effect of individual variation on disease emergence. Nature 2005;438:355-9.

5. Bansal S, Grenfell BT, Meyers LA. When individual behavior matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface 2008;4:879-91.

6. Riley S, Fraser C, Donnelly C, et al. Transmission dynamics of the etiological agent of SARS in Hong Kong: Impact of Public Health Interventions. Science 2003;300:1961-6.

7. Salathé M, Jones J. Dynamics and Control of Diseases in Networks with Community Structure. PLoS Comput Biol 2010;6:e1000736.

8. Zagheni E, Billari F, Manfredi P, et al. Using Time-Use Data to Parameterize Models for the Spread of Close-Contact Infectious Diseases. American Journal of Epidemiology 2008.

9. Mossong J, Hens N, Jit M, et al. Social contacts and mixing patterns relevant to the spread of infectious diseases. Plos Med 2008.

10. Milbrath MO, Zelner JL, Moe CL, et al. Heterogeneity in Duration of Infectiousness: Impact on Norovirus Outbreak Dynamics. 2010.

11. Gotz H, Ekdahl K, Lindback J, et al. Clinical Spectrum and Transmission Characteristics of Infection with Norwalk-Like Virus: Findings from a Large Community Outbreak in Sweden. Clinical Infectious Diseases 2001;33.

12. Zelner J, King A, Moe C, et al. How Infections Propagate After Point-Source Outbreaks. Epidemiology 2010;21:000-.

13. Lynch M, Painter J, Woodruff R, et al. Surveillance for foodborne-disease outbreaks—United States, 1998-2002. MMWR Surveill Summ 2006;55:1-42.

14. Zomer TP, De Jong B, Kühlmann-Berenzon S, et al. A foodborne norovirus outbreak at a manufacturing company. Epidemiology and Infection 2010;138:501-6.

15. Hoebe C, Vennema H, de Roda., A. Norovirus outbreak among primary schoolchildren who had played in a public fountain. The Journal of Epidemiology and Community Health 2004.

16. Fretz R, Svoboda P, Lüthi T. Outbreaks of gastroenteritis due to infections with Norovirus in Switzerland, 2001–2003. Epidemiology and Infection 2005.

17. Atmar R, Estes M. The epidemiologic and clinical importance of norovirus infection. Gastroenterology Clinics of North America 2006;35:275-90.

18. Baert L, Uyttendaele M, Stals A, et al. Reported foodborne outbreaks due to noroviruses in Belgium: the link between food and patient investigations in an international context. Epidemiology and Infection 2009;137:316-25.

19. Lopman B, Gallimore C, Gray J, et al. Linking healthcare associated norovirus outbreaks: a molecular epidemiologic method for investigating transmission. BMC Infectious Diseases 2006.

20. Li S, Eisenberg JNS, Spicknall IH, et al. Dynamics and control of infections transmitted from person to person through the environment. American Journal of Epidemiology 2009;170:257-65.

21. Cheesbrough JS, Green J, Gallimor CI, et al. Widespread environmental contamination with Norwalk-like viruses (NLV) detected in a prolonged hotel outbreak of gastroenteritis. Epidemiology and Infection 2000;125:93-8.

22. Becker KM, Moe CL, Southwick KL, et al. Transmission of Norwalk Virus During a Football Game. New England Journal of Medicine 2000;343:1223-7.

23. Fankhauser R, Monroe S, Noel J, et al. Epidemiologic and molecular trends of 'Norwalk-like viruses' associated with outbreaks of gastroenteritis in the United States. The Journal of Infectious Diseases 2002;186:1-7.

24. Goodreau S, Kitts J, Morris M. Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. Demography 2008;46:103-25.

25. Morris M, Kurth AE, Hamilton DT, et al. Concurrent Partnerships and HIV Prevalence Disparities by Race: Linking Science and Public Health Practice. American Journal of Public Health 2009;99:1023-31.

26. Acevedo-Garcia D. Residential segregation and the epidemiology of infectious diseases. Social Science & Medicine 2000;51:1143-61.

27. Moody J, White D. Social cohesion and embeddedness: A hierarchical conception of social groups. American Sociological Review 2003.

28. Aparicio J, Pascual M. Building epidemiological models from Ro: an implicit treatment of transmission in networks. Proceedings of the Royal Society: B 2007;274:505-12.

29. Bureau USC. United States Census, 2010. (http://www.census.gov/).

30. Swanson DA, Siegel JS. The Methods and Materials of Demography. New York: Academic Press, 2004.

31. CDC. Norovirus Outbreak Among Evacuees from Hurrican Katrina -- Houston, Texas, September 2005. MMWR 2005;54:1016-8.

32. Robins G, Pattison P, Kalish Y, et al. An introduction to exponential random graph (p*) models for social networks. Social Networks 2007.

33. Robins G, Pattison P, Woolcock J. Small and Other Worlds: Global Network Structures from Local Processes 1. American Journal of Sociology 2005;110:894-936.

34. Robins G, Snijders T, Wang P, et al. Recent developments in exponential random graph (p*) models for social networks. Social Networks 2007;29:192-215.

35. Goodreau S. Advances in exponential random graph (p*) models applied to a large social network. Social Networks 2007.

36. Snijders T. Markov chain Monte Carlo estimation of exponential random graph models. Journal of Social Structure 2002.

37. Bearman P, Moody J, Stovel K. Chains of Affection: The Structure of Adolescent Romantic and Sexual Networks. American Journal of Sociology 2004;110:44-91.

38. Morris M, Handcock M, Hunter D. Specification of Exponential-Family Random Graph Models: Terms and Computational Aspects. Journal of statistical software 2008;24:1548-7660.

39. Watts DJ. A simple model of global cascades on random networks. Proceedings of the National Academy of Sciences 2002;99:5766-71.

40. Miller JC. Spread of infectious disease through clustered populations. Journal of the Royal Society Interface 2009;6:1-14.

41. Miller J. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. Physical Review E 2007;76.

42. Phillips G, Tam CC, Rodrigues LC, et al. Prevalence and characteristics of asymptomatic norovirus infection in the community in England. Epidemiology and Infection 2010:1-5.

43. Atmar RL, Opekun AR, Gilger MA, et al. Norwalk virus shedding after experimental human infection. Emerging Infectious Diseases 2008;14:1553-7.

44. Prasad AM, Iverson LR, Liaw A. Newer classification and regression tree techniques: Bagging and random forests for ecological prediction. Ecosystems 2006;9:181-99.

45. Baron RM, Kenny DA. The moderator-mediator variable distinction in social psychological research: conceptual, stategic, and statistical considerations. Journal of Personality and Social Psychology 1986;51:1173-82.

46. Aparicio J, Capurro A, Castillo-Chavez C. Transmission and Dynamics of Tuberculosis on Generalized Households. Journal of Theoretical Biology 2000;206:327-41.

47. He D, Ionides EL, King AA. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. Journal of the Royal Society Interface 2010;7:271-83.

48. Ionides E, Breto C, King AA. Inference for nonlinear dynamical systems. Proceedings of the National Academy of Sciences 2006;103:18438-43.

49. Marjoram P, Molitor J, Plagnol V, et al. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences 2003;100:15324-8.

APPENDIX B

Chapter 3 Supplementary Materials

S1. Simulating household sizes and age distribution. We make some simplifying assumptions to generate households in a transparent way. For instance, we assume that individuals younger than 20 years are not parents and that households with householders older than 50 years do not have children. We also assume that each household has a maximum of two adults, i.e. individuals aged over 18 years, and that these adults are of the same age. Although these simplifications guarantee deviations from the exact distribution of household sizes, our procedure covers the great majority of households, and therefore serves the heuristic purpose for which it is intended. When generating a collection of households, we follow the following procedure:

```
population size = 0
IF population size < N:
    Sample Householder 1 age from age distribution of adults 20+ years of age.

    With p = .5, add Householder 2:
        Householder 2 age = Householder 1 age

    IF Householder 1 age < 50, sample resident children:
        Sample number of children, given Householder 1 age

        IF number of children > 0:
            Sample mother's age at first birth from Normal(23,2)

            years since first birth = 0
            For child in number of children:
                child age = mother's age at first birth + years since first birth
                years since first birth = years since first birth + min(1,Exponential(2))
        ELSE:
            STOP
    ELSE:
        STOP

    population size = population size + household size

ELSE:
    STOP
```

Algorithm 3.1

The distribution of householder ages is obtained from the U.S. Current Population Survey. The distribution of number of children by mother's age is obtained using age-specific parity tables from Swanson and Siegel (1).

S2. Sampling from an exponential random graph (ERG) model using the Metropolis-Hastings algorithm

When sampling community contact networks for outbreak simulations, we begin with an Erdos-Renyi random graph, *x*. This network has *N* nodes representing the individuals in the community. The average degree of nodes in the network is $\bar{k}$. In order to maintain a constant value of $\bar{k}$, we utilize a sampling algorithm that, on each step, moves an edge from a connected pair of nodes to a randomly selected pair.

We define the adjacency matrix of *x* to be A, so that the entry corresponding to an edge between nodes *i* and *j*, A_{ij}, equals 1. On each step of the algorithm, a random edge, A_{ij}, is selected from the set of existing edges in the set of edges in *x*, denoted *E*. A candidate edge A_{kl} is selected from the set of all possible combinations of unconnected vertices not in *E*. We then evaluate the probability of a new graph, x', in which $A_{ij} = 0$ and $A_{kl} = 1$. The probability of accepting x', *a,* is a function of the ratio of the likelihoods of x' and *x,* given the network statistics in θ and the average degree, $\bar{k}$. This ratio, denoted as *r*, is computed as follows:

$$r = \left(\frac{P(x'|\theta,\bar{k})}{P(x|\theta,\bar{k})} \right) = \exp\left\{ \sum_{i\in\{1\ldots n\}} \theta_i \left(S_i(x') - S_i(x) \right) \right\}$$

Equation 3.7

In the equation above, θ and S are vectors of network parameters corresponding network statistics, respectively, both of length *n*. Using the value of *r* obtained from Equation 3.7, we can calculate the probability of accepting x', denoted p_{accept}, and move the chain to x' with probability proportional to p_{accept}:

$$x^{t+1} = \begin{cases} x' \text{ with probability } p_{accept} \\ x \text{ with probability } (1 - p_{accept}) \end{cases}$$

Equation 3.8

When sampling networks, we sample a set of ages form the household model and then run the algorithm above for a burn-in period of 500,000 steps, which ensures that sampled graphs will not contain traces of the initial state of the graph. Thereafter, we sample random contact networks at intervals of 10^3 steps and use these to evaluate the contact heterogeneity and outbreak dynamics associated with the social processes described in the text.

S4. Epidemic Percolation Graph Model of Community Transmission

We utilize the epidemic percolation graph (EPG) method of Miller (3) to simulate the final size of NoV outbreaks. In this model, a weighted network $G = \{V, E\}$ is constructed, where the vertices, *V*, consist of the individuals in the community and the edges in *E* are directed links weighted by the probability that node i will infect node *j*, given that *i* is infectious and *j* is susceptible. Outbreak realizations are generated by randomly selected edges to remain in the graph with probability proportional to their transmissibility. The size of outbreak associated with infection of a single index node is equivalent to the size of the out-component of the node in the resultant graph. For a point-source outbreak involving multiple index nodes, the size of an outbreak is equivalent to the set of nodes included in all out-components of the index nodes.

In our EPG model, household and community contacts are condensed into a single network where nodes are connected by weighted edges, and the weights correspond to transmissibility. Transmissibility is a function of phase of infection, i.e. symptomatic vs. asymptomatic, and context, i.e. household vs. community.

When the model is initialized, all individuals are assigned the duration of symptomatic, $d_{S,i}$, regular asymptomatic infection, $d_{A,i}$ and long asymptomatic infection, $d_{L,i}$, they would have if infected. The average duration of symptomatic infectiousness is denoted $1/\gamma_S$. The average duration of regular and long asymptomatic shedding are denoted, $1/\gamma_R$ and $1/\gamma_L$, respectively. The durations of these infectious phases that are used in simulations are realizations of the random variables D_S, D_R, D_L, respectively, with the distributions defined below:

$$D_s \sim \text{Exponential}(1/\gamma_S)$$
$$D_R \sim \text{Exponential}(1/\gamma_R)$$
$$D_L \sim \text{Exponential}(1/\gamma_L)$$

We also define X to be a random variable with a uniform distribution, where *x* is a random realization of X:

$$X \sim \text{Uniform}(0,1]$$

Using all of these values, we can assign durations of symptomatic and asymptomatic infectiousness to all nodes in the community:

$$\text{for all nodes } v \in V:$$
$$d_{S,v} = \begin{cases} 0 \text{ if } x < \tau \\ d_s \text{ otherwise} \end{cases}$$
$$d_{A,v} = \begin{cases} d_R + d_L \text{ if } x < \rho \\ d_r \text{ otherwise} \end{cases}$$

Algorithm 3.1

Using these duration values, we can weight the directed edges connecting nodes in the community and household networks. $Q_{H,i}$ is the transmissibility of a household network edge emanating from node *i*, while $Q_{C,i}$ is the transmissibility of an edge in the community emanating from *i*:

$$Q_{H,i} = 1 - \exp(-\beta_{IS} d_{S,i} + \beta_{IA} d_{A,i})$$
$$Q_{C,i} = 1 - \exp(-\beta_{IA} d_{A,i})$$

Equation 3.9

Because individuals do not transmit in the community during the symptomatic phase of infection, we omit the strength and duration of symptomatic infectiousness from the calculation of the transmissibility of community links, $Q_{C,i}$.

Finally, we outline the mechanism by which point-source spreaders transmit in the EPG model. At initialization, a set of nodes is chosen, each with probability ζ, to be food servers. In order to preserve the burst-like quality of point-source events (PSE), when a PSE occurs, the food handler infects all of their uniformly selected point-source contacts at one time. The size of these point-source exposures are drawn at initialization and directed edges are placed between the food-server and exposed individuals. We denote T_{PSE} to be a random variable representing the waiting time until a PSE event occurs. $t_{PSE,i}$ is a realization of T_{PSE} associated with individual *i*. *Z* is the set of individuals exposed to individual *i*.

$T_{\text{PSE,i}} \sim \text{Exponential}(\lambda)$
for all nodes $v \in Z$:
  $Q_{\text{PSE,v}} = 1$, if $t_{\text{PSE,i}} < d_{A,i}$
  $Q_{\text{PSE,v}} = 0$, otherwise

Algorithm 3.2

References

1. Swanson DA, Siegel JS. The Methods and Materials of Demography. New York: Academic Press, 2004.

2. Zagheni E, Billari F, Manfredi P, et al. Using Time-Use Data to Parameterize Models for the Spread of Close-Contact Infectious Diseases. American Journal of Epidemiology 2008.

3. Miller J. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. Physical Review E 2007;76.

Chapter 4

Social connectedness can inhibit disease transmission: Social organization, cohesion, village context and infection risk in rural Ecuador

Abstract

Epidemiologists typically approach social networks as conduits that transmit infectious disease. But dense networks of relationships may also manifest social organization that suppresses exposure to pathogens. We analyze the influence of social connectedness and organization on infectious disease transmission in 19 villages in rural coastal Ecuador that varied by their remoteness (measured by the difficulty and cost of travel to the nearest major town). Social connectedness is assessed using data from community-wide surveys of social networks, sanitation, and illness outcomes. We estimate risk and protective factors associated with two village social networks that represent relationships characterized by 1) discussing important matters (IM) or 2) passing time together (PT). The most remote village in our sample has a 50% lower risk of disease (OR = 0.49, 95%CI = [0.29,0.84]) than the least remote one, and this can be explained by increased social connectedness as well as improved water treatment, sanitation and hygiene. Controlling for exposure to illness in ego's community and household, for average households, a 1-unit increase in a village's average network

degree is associated with a reduction in risk of illness (IM: OR = 0.83, 95%CI=[0.72-0.95]; PT: OR = 0.89, 95%CI=[0.81-0.98]). Other network exposure factors within households and contacts outside villages were shown to be risk factors. Social networks are not only conduits for the transmission of infectious disease but also adaptive resources that facilitate the development of social mechanisms for preventing the spread of disease.

Introduction

Studies of the transmission of infectious diseases (1, 2) often use social networks as maps of direct contacts that facilitate person-to-person transmission of pathogens. But dense social networks may represent social cohesion that can foster education and social organization. These, in turn, may lead to better water sanitation and hygiene practices, which can prevent population-level exposure. For pathogens whose community-level transmission is strongly impacted by the efficacy of water sanitation infrastructure and household-level hygiene practices. In some instances, the preventive effects of networks might even outweigh their transmission potential so that individuals experience lower risk of infection in communities with denser and more cohesive social networks, as illustrated by Figure 4.1. This contrasts with the prevailing view of networks as transmission systems in which an increasing number of contacts is associated with greater individual-level risk (3). We test the idea that increased social network connectedness predicts diminished risk of infection using a sample of 19 villages in rural, northern coastal Ecuador. A road was recently built that connects these villages to the nearest large town in the area, which itself has about 5000 inhabitants, was recently built.

Consequently, these villages now vary in their remoteness, measured by their distance and difficulty of travel to this trading center. Our previous analysis suggests a causal relationship between remoteness and the more proximal social network variables (4) as well as connectivity within and outside of the study region (5) Less remote villages have more transient inhabitants and are more socially fragmented and less able to build and maintain water sanitation infrastructure and promote hygiene practices than more remote ones. To understand these effects, we analyze infection risk in these villages in terms of three specific mechanism that are downstream from remoteness and may affect diarrheal disease risk: 1) the rate of introduction of pathogens into villages as measured by contact with individuals from outside; 2) social cohesion that may provide protection through improved infrastructure; and 3) contact patterns that may spread disease from either person-to-person or person-to-environment-to-person (4, 5).

We define a network comprised of relationships that are likely to facilitate transmission of pathogens as a *contact* network; i.e., a structure of connections through which an individual, denoted as *ego*, may be infected by his or her network neighbors, denoted as *alters*. This network contains all of the pathways an infection may follow through the community via direct human contact. In contrast to these contact networks, we define links in *sociality* networks as connections between people that represent specific types of social engagement. Connections in sociality networks can correspond to casual acquaintance, close friendship and trust, or economic exchange. These relationships have a meaningful impact on infection risk because they often determine whether communities have effective sanitary infrastructure and health services (6). In this

way, more network connections (e.g., friends), may be protective and indicative of social support, instead of transmission-enhancing, as in a contact-only network (7).

Although the social cohesion and organization represented by these sociality networks is critical to the functioning of communities (8, 9, 10), the influence of these factors on infectious disease risk at the community level is typically neglected. Social relationships have long been employed as contact networks in transmission models, e.g., (1, 11, 12, 13), and as factors that are protective against chronic disease (*14, 15*). But outside of the literature on sexually transmitted diseases, e.g., (*16, 17*), there are comparatively few examples of the protective and organizing role these social relationships may play in the epidemiology of infectious diseases. Cohesive communities are more likely to perceive their economic and social interests as common to the group and may be more motivated and better able to pursue collective goals such as building and maintaining effective water and sanitary infrastructure. This means that understanding the drivers of infectious disease risk at the community level requires understanding the elements of social structure which condition the infrastructure and behavior that can prevent population-level exposure. In this paper, we operationalize specific risk and protective effects of social relationships via survey and social network analysis methods and show that social networks can reduce as well as enhance risk in the epidemiology of infectious diseases.

Community social structure and risk

Understanding how sociality networks influence infection risk in these villages requires consideration of how social organization and action can inhibit or enhance

pathogen transmission via the environment. Figure 4.1 illustrates our posited mechanism of the effect of social cohesion on illness risk. Poor quality sanitary infrastructure is a leading population-level risk factor for infection by enteric pathogens such as cholera (18, 19, 20), and such infrastructure is usually a public good that requires ongoing funding and management by the community. Transmission of many such pathogens is often conceptualized as person-to-environment-to-person, with water acting as the environmental reservoir (*21*). Greater community cohesion may predict better overall water quality through education in sanitary practices and social organization that produces infrastructural improvements, such as water filtration and sewage treatment. If this is true, we would expect to find an that the average number of social network connections in a village and risk of infection by enteric pathogens are inversely related.

For example, if ego has many relationships in her village sociality network and belongs to a community organization focused on improving local water quality, she may help reduce the entire village's exposure to pathogens. Although her social relationships can also function as transmission connections, it is possible that the salutary effects of her social engagement will render transmission via those connections less relevant by reducing village-wide exposure to enteric pathogens in the first instance. This idea is supported by the work of Chang et al. (*22*), who found that higher levels of poverty and violent crime, which indicate diminished social control and organization, predict increased shigellosis incidence at the level of U.S. counties. On a similar note, Troesken (*23*), shows that a switch from private to public provision of water services, and a corresponding shift in incentives, in early 20th century New Orleans predicted a sharp decline in Typhoid incidence among that city's black residents.

Measuring effects of sociality and contact networks on risk

When assessing the role of social networks in infection risk, the distinction between the sociality and contact aspects of relationships is not always clear. We address this by analyzing our illness data with respect to two different sociality networks and comparing the results of each analysis. We define the person-to-person contact network in all analyses as the network comprised of individuals, excluding ego's household members, with whom ego reports having spent time with in the previous week. This is called the *passing time* or PT network. We use this inclusive definition of contact because a wide range of casual and close contacts can transmit gastrointestinal pathogens (*24*).

In our first analysis, the PT network represents both village sociality and contact networks. This definition of a sociality relationship is inclusive and likely to highlight many connections between people in the community but is a general network that may only weakly reveal underlying social structure. Therefore, if a minimal level of attachment by individuals to the community is necessary to impact infection risk, we should expect that relationships in the PT network will predict diminished risk. In the second analysis, membership in the sociality network is constrained to relationships corresponding to the question: "Outside of members of your household, with whom do you talk about important matters?" We define this network as the *important matters*, or IM network. This network typically contains fewer individuals than the PT network, but may reveal more about the underlying structure of the community. If a level of attachment to the community stronger than that implied by the PT network is important for reducing illness risk, relationships in the IM network should be better predictors of

risk than those in the PT network. By comparing results obtained using each of these sociality networks, we can better understand how the definition of relationships in the sociality network impacts our ability to assess sources of risk and understand the effects social relationships have on risk.

Our analysis of contact networks focuses on ego's risk of infection from contact with ill individuals in her contact network. By contrast, the analysis of sociality networks focuses on risk associated with the sociality network's aggregate, village-level features and ego's position in this village-wide network. This allows us to examine separate effects of these different aspects of each type of network on disease outcomes.

Methods

Setting

Our data consist of 2938 observations among individuals aged 13 and older, from a 2007 survey of sociality networks, contact networks, household wealth, village remoteness, and disease outcomes (diarrhea and fever) in 19 villages in the northern coastal Ecuadorian province of Esmeraldas. These villages are situated along 3 rivers, the Cayapas, Santiago, and Ónzole, all of which drain towards Borbón, which is the major population center of the region. In 1996, a new paved road was built westward from Borbón to the coast, and in 2001 a road connecting Borbón to the Andes was completed. A network of smaller roads linking villages to the main road is under continual construction. These villages vary by remoteness, a function of their time and cost of travel to Borbón (see (5)). Remoteness influences social relationships and network

structure, migration in and out of the region, and other factors that impact both social network characteristics and exposure to infectious diseases.

Previous work with this collection of villages has shown that a village's remoteness is inversely associated with the level of lab-confirmed enteric disease in that community (5). This finding was observed in parallel with the fact that the network of social contacts was significantly denser in the villages far from roads (4). Existing theory would predict that dense villages would be more susceptible to outbreaks of enteric pathogens, because the average infected individual would have more available susceptible contacts in the denser network than in the more diffuse one.

Additional Network Data

In addition to the important matters (IM) and passing time (PT) networks described above, we represent contacts between households by links in a food-sharing network. Relationships in this network were obtained using the survey question: "Outside of your household, with whom did you share food with in the past week." In this network, two households are connected if food was shared between them during the week prior to the survey. To gauge the effect of these relationships at the individual level, the household network is represented as a bipartite network, where the two sets of nodes correspond to individuals and to households, and an individual is connected to all of the other households in his or her household's food-sharing network.

Outcome Measure

Recent infectious illness. Our outcome measure is ego's self-reported diarrheal disease or fever in the week prior to the survey. Diarrheal illness is defined as ego having

3 or more liquid stools in one day (*25*). We combine these two categories of illness into a single binary response variable that indicates whether the individual has recently experienced illness likely of infectious origin.

Measuring community cohesion and household attachment

We take several approaches to measuring social cohesion and organization, utilizing data on the structure of community social networks, participation in community organizations, and perceptions of trust within the community.

Average degree of individuals in sociality network. This measures the average number of relationships in the sociality network for individuals aged 13 and above. As the number of connections per person grows, the cohesion of the community is expected to grow as well, e.g., (4, 26). Unless otherwise noted, this quantity is measured in 1-unit increments.

Household sociality degree. We expect the salutary effects of social connectedness to operate at the household level. Because of this, we measure the effect of sociality network degree on risk using the sociality degree of the most connected individual in ego's household, which we define as *household degree*. In order to measure the effect of a relative increase in degree at the village level, we standardize each village's distribution of household degree, D, to have mean zero and unit variance. In all analyses, we present household degree in standard deviation (SD) units from village mean household degree.

Clustering coefficient of sociality network. Multivariate models with network covariates are adjusted for the global clustering coefficient of the village sociality network. For a detailed description of clustering, see the supplementary materials.

Participation in community organizations. We also collected data in the survey on participation in community organizations, which reflects the engagement of individuals in the civic life of their community. We measure this using the mean number of organizations that individuals aged 13 and above in the village belong to. We measure the participation of households as the maximum number of memberships by anyone living in ego's household, which we define as *household membership*. As with household degree, this value is standardized to have unit variance and equal zero at its village mean.

Other Covariates

Village remoteness: Remoteness is measured using a scale that is a composite of distance and cost of travel to Borbón, the commercial center in the region. We normalize the value of remoteness so that the closest village has a remoteness value of zero and the most remote village has remoteness equal to 1.

Individual and village demographics: All analyses are adjusted for ego's age. We also adjust the number of inhabitants of ego's village included in the survey, and ego's household size.

Contact with individuals outside the village: We estimate risk of illness associated with two variables indicating a) the proportion of households in which at least one person has left the village or b) received visitors from outside the village, both in the previous week.

Contact network exposure: We estimate the change in ego's illness risk associated with each additional alter in her contact (PT) network who reported symptoms in the previous week.

Household exposure: We estimate the change in ego's risk of disease associated with each additional infected household member. We subtract the sample mean household size from each household so that our risk estimates reflect the effect of deviation from the average household.

Food-sharing exposure: We include the count of the number of individuals in ego's food-sharing network who reported illness in the previous week.

Wealth: We estimate the impact of household-level wealth, measured by the ownership of material goods, on risk. We assess wealth with a scale from 0 to 10, that is a weighted sum of the number of items belonging to a household. These include televisions, motorcycles, refrigerators, chainsaws, etc. For details on the measurement of household wealth, see the supporting materials.

Education: More years of formal education may lead to greater awareness of methods for preventing infection by diarrheal pathogens through, for example, sanitation and water treatment. We estimate the change in risk of infection associated with a 1-year increase in the average number of years of education for all individuals aged 13 and older in each village.

Water sanitation and hygiene: We measure sanitation and water quality using several community-level indices: *Observed hygiene* is the average of 23 indicators of household cleanliness across all households in the village. *Improved sanitation* is the

proportion of households in the village with improved sanitation, defined as a septic tank or a latrine. *Improved water source* measures the proportion of households using a water source such as piped water or rainwater collection. *Water treatment* measures the proportion of households in the community reporting that they used some kind of water treatment. For values of these measures by village, see Table 4.6 in the supplementary materials.

Modeling risk for individuals nested in communities

Since we are conceptualizing individual-level outcomes as the consequence of village-level factors, we are making multiple observations of the same village and expect that responses will be correlated. We deal with this correlation in all regression models by using mixed-effects logistic regression models with village-level random intercepts and estimates of individual-level fixed effects for all covariates over all villages (*27, 28*). Because such models are fit via maximum-likelihood, likelihood-based goodness-of-fit measures such as the Akaike information criterion (AIC) (*29*) may be used for model comparison and selection, assuming correct model specification. The AIC balances parsimony and goodness-of-fit by imposing a penalty on the likelihood of the model for each additional parameter.

Indirect effects of village-level characteristics on individual risk

We assume that remoteness does not directly impact disease, but acts through (or is mediated by) more immediate factors (e.g. sanitation) as illustrated by Figure 4.1. Since quantifying the mediation effect through the difference of regression coefficients, e.g. (*30*), is not readily extended to binary response variables, we use an alternate

approach. We estimate the mediation effect as the difference between a) the crude association of remoteness with illness and b) the association of remoteness and illness, adjusted for the mediator variable. These are quantified by the ratios of the expected probability of illness for individuals in Far versus Near villages, with and without the mediator in the model. Statistical significance of this effect is assessed using a non-parametric bootstrap. This analysis is repeated to estimate the mediation effect of average village level degree. For a detailed discussion of this mediation analysis, see the supplementary materials.

Results

Descriptive Results

Village-level descriptive statistics of remoteness, illness, water sanitation, water quality and household hygiene are listed in Table 4.1. Villages are listed in order of increasing remoteness. Descriptive characteristics of interest for the important matters (IM) and passing time (PT) networks for each village include average degree and the global clustering coefficient (Table 4.2). Additional village-level descriptive statistics, on organization membership, education and wealth, are available in Table 4.5 in the supplementary materials.

Individual and village-level predictors of infectious illness risk

Logistic regression models were analyzed for the effects of household and village-level social network characteristics, village-wide socioeconomic status (wealth, education), social capital (membership in community organizations), and contact with individuals outside of ego's village on risk of illness. Model 1 (AIC = 2110) shows

sources of exposure risk, adjusted for age and village size. This model shows that: 1) a 10% increase in the proportion of households with visitors from outside the community in the week prior to the survey predicts an increased risk of illness (OR = 1.12, 95%CI = [1.00-1.25]); 2) a 1-person increase in the number of ill individuals in ego's household predicts increased risk of illness (OR = 1.59, 95%CI = [1.41-1.79]); and 3) a 1-person increase in the size of ego's household is associated with diminished risk (OR = 0.86, 95%CI=[0.82,0.91]).

Living in the villages with the largest values of average IM and PT degree compared to the lowest is associated with a significant unadjusted reduction in risk (IM: OR = 0.38, 95%CI = [0.22-0.68]; PT: 0.41, 95%CI = [0.20-0.84]). Models 2 and 3 are adjusted for household and village-level network characteristics. Both networks show that a 1 unit increase in average village-level degree is associated with diminished risk when household degree is fixed at its village mean (IM: OR = 0.83, 95%CI = [0.72-0.95]; PT: OR = 0.89, 95%CI = [0.81-0.98]). The statistically significant interaction between village average and household IM degree (OR = 1.17, 95%CI = [1.04-1.32]) suggests that the protective effect of village level average degree applies to households with degree less than 0.6 standard deviations above the village mean. Above this level the associations become non-significant and therefore our data cannot resolve the association. For further discussion of this relationship, see the supplementary materials.

Indirect effects of distal (remoteness) and intermediate (social network) factors on risk

Residence in the most remote versus the least remote village in our sample is associated with a large decrease in ego's unadjusted risk of infectious illness (OR = 0.49, 95%CI = [0.29-0.84]). This effect operates through five statistically significant village-level mediators ($p \leq .10$): the percentage of households in the village with an outside visitor in the previous week ($\hat{m} = 0.061$, $p = 0.012$), village-wide average degree in the PT network ($m = 0.047$ $p = 0.054$), improved sanitation ($\hat{m} = 0.042$, $p = 0.012$), improved water treatment ($\hat{m} = 0.072$, $p = 0.04$) and ego's household size ($\hat{m} = 0.016$, $p = 0.004$). We also find that the mean number of organization memberships by adults in ego's community ($\hat{m} = -0.046$ $p = 0.081$) suppresses the effect of remoteness. In order to assess whether the four mediators can fully explain the association between remoteness and illness, we fit a logistic regression model predicting ego's illness with remoteness, household size, village average PT degree and improved sanitation and water treatment as covariates. In this model, the relationship between remoteness and illness is no longer significant and the point estimate is closer to the null (OR: 0.71, 95%CI: 0.46, 1.10), suggesting that these five variables are important mediators linking remoteness to illness.

The mediation analysis thus far suggests that a distal factor such as remoteness is mediated through social network factors as well as more proximal water and sanitation factors. Village-level social network factors are also mediated by some of the more proximal factors. For example, observed hygiene is a statistically significant mediator of the effect of village-wide average important matters (IM) and passing time (PT) degree (IM: $\hat{m} = 0.093$, $p = 0.010$; PT: $m = 0.134$, $p = 0.001$). Improved sanitation mediates IM degree ($\hat{m} = 0.039$, $p = 0.001$) while suppressing PT degree ($\hat{m} = -0.042$, $p = 0.004$). This finding, that improved sanitation suppresses the village level protective effect of the PT

network, can be explained by a single village with high-quality water sanitation but low average degree. Dropping this village from the analysis eliminates the suppressive effect of improved sanitation for PT ($\hat{m}$= 0.006, p = 0.455).

Discussion

Contrary to the standard representation of highly connected social networks as efficient transmission systems (e.g., (3)), we show that greater connectivity at the village level can inhibit the prevalence of self-reported diarrheal disease and fever. While controlling for sources of exposure to illness, our analysis shows that increasing village-wide average degree is associated with decreasing risk for households of average degree in the important matters (IM) network and all households in the passing time (PT) network.

Because our multivariate models assess risk factors for the same group of individuals and villages arranged into different village-wide networks, we can compare the relative ability of these networks to explain risk using the AIC. We find that model 3 in table 4.3, which includes both individual and village-level features of the IM network, provides the best overall fit to the data. This suggests that relationships in the IM network are more useful for revealing the social backbone of the community than those in the PT network. This may be because IM network relationships imply greater intensity and trust than those in the PT network. Their more stringent definition may also result in less variability in the meaning and intensity of relationships in the network across individuals and communities that makes these networks more useful for understanding how social

support in a sample of communities affects risk. In addition to the measures of social connectedness and support presented here we also explored the effect of other factors related to social cohesion, such as levels of mutual trust in the village on risk, but found these associations to be non-significant.

Our analysis also connects these more proximal social network, water sanitation, and hygiene factors to the more distal social and environmental context in which the village is situated, i.e. its remoteness. Specifically, we employ a mediation analysis to examine how remoteness affects social networks, which in turn affect the quality of water and sanitation infrastructure. Previous studies postulate that remoteness impacts risk via its effects on contact networks and village cohesion (5). Results from our mediation analysis agree with this theory, showing that more remote villages experience decreased risk because the average individual in them has more relationships in the village passing time (PT) network, and lives in a larger household than a comparable person in a less remote village. Further mediation analysis suggests that villages with high average degree experience a decreased risk of illness through improved water quality and sanitation. In such villages, these protective effects render the increased risk associated with living in a low-degree household non-significant.

Our finding that household size predicts decreased risk is explained by the finding that increasing household size mediates of the protective effect of remoteness. Although living in a larger household means that ego has more contacts, larger households may also indicate tighter integration of individuals into their communities. This would be consistent with the overarching finding that the protective effect of remoteness operates through its effects on increased social cohesion.

Our causal model (Figure 4.1), hypothesizes a relationship between village remoteness and risk that operates through village social organization and cohesion. This social organization is postulated to lead to activities, such as building water sanitation infrastructure, hygiene practices, etc., that decrease disease prevalence in the village. Because we conceptualize sanitation and hygiene as village-level constructs, the relatively small number of villages in our sample makes it difficult to directly test the hypothesis that water sanitation and hygiene are outcomes of village-level social cohesion. Nonetheless, results illustrating the mediation of village remoteness and average degree by these factors strongly suggest that this is the case.

In addition to the protective effects of social behavior outlined here, we also find that migration between villages, measured by whether ego had contact with a visitor from outside the village in the previous week, predicts increased risk of infection. This confirms previous findings from these villages (5). The result that this exposure suppresses the protective effects of village-level average IM and PT degree underscores the effectiveness of between-village migration as a means of disseminating pathogens.

Networks of social relationships can reduce the risk of illness from infectious diseases by mitigating population-level sources of risk in ways that preempt the role of social networks in person-to-person transmission. These results expand on theory that social connectedness and support are important predictors of chronic illness (*31, 32*) as well as the theory that social connectedness can mitigate the spread of pathogens such as tuberculosis and HIV (e.g., (*33*)). Infectious disease epidemiologists and social scientists should incorporate insights presented here into mechanistic models that can explain outbreak and epidemic time series in terms of both the contact and sociality functions that

networks have. Such models can provide a more nuanced analysis of the relative contributions of social organization and contact to the risk of infectious diseases.

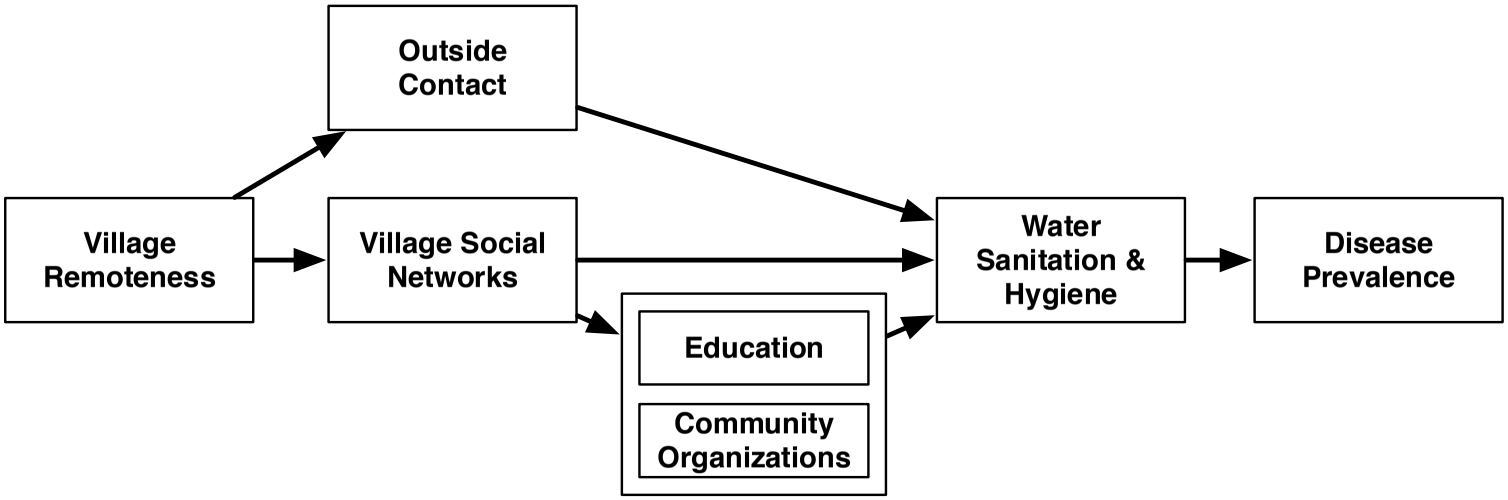

Figure 4.1. Postulated causal diagram. Solid arrows illustrate the hypothesized causal pathway by which remoteness impacts risk of infection. Increasing remoteness is postulated to increase social cohesion and organization, which are evidenced by an increased number of relationships in village social networks and greater participation in community organizations. More socially cohesive villages are believed to be more likely to 1) build and maintain effective water sanitation infrastructure and 2) promote hygiene through education, thereby decreasing disease prevalence in the village.

	Remoteness		Sample Size	Fever or diarrheal disease	Households w/ water treatment	Households w/ improved sanitation	Households w/improved water source	Observed household hygiene index
Village	Continuous	Category	N	Cases /100	%	%	%	Mean
1	0.06	Close	234	15	25	43	43	0.64
2	0.07	Close	785	15	74	33	49	0.70
3	0.11	Close	35	11	33	50	0	0.62
4	0.13	Close	465	13	18	55	59	0.69
5	0.20	Medium	114	12	14	61	7	0.69
6	0.20	Medium	52	16	0	64	15	0.63
7	0.20	Medium	47	19	93	11	2	0.53
8	0.25	Medium	56	7	33	100	0	0.79
9	0.25	Medium	47	28	72	55	100	0.51
10	0.31	Medium	106	12	0	15	0	0.45
11	0.40	Medium	69	14	0	26	100	0.68
12	0.57	Medium	91	18	23	50	77	0.71
13	0.62	Medium	121	20	19	7	19	0.31
14	0.71	Far	89	7	13	52	48	0.38
15	0.78	Far	224	10	33	55	55	0.71
16	0.80	Far	81	0	15	86	99	0.74
17	0.83	Far	311	8	0	41	82	0.73
18	0.96	Far	346	7	13	56	64	0.73
19	1.00	Far	140	14	5	50	28	0.68

Table 4.1. Descriptive characteristics of villages. Village characteristics include remoteness, the number of individuals surveyed (sample size), prevalence of having either non specific fever or diarrheal disease, the percentage of households in the village using some form of water treatment, improved sanitation (i.e., septic tank or latrine), or an improved water source (i.e., piped water or rainwater collection), and an index, from 0 to 1, of observed hygiene practices.

		Important Matters			Passing Time		
Village	Remoteness	Degree		Global Clustering	Degree		Global Clustering
		Mean	Max		Mean	Max	
1	0.06	1.9	8	0.13	5.1	16	0.18
2	0.07	2.8	15	0.17	5.2	27	0.19
3	0.11	1.2	2	0.00	2.9	9	0.16
4	0.13	3.1	13	0.15	5.1	21	0.14
5	0.20	3.6	10	0.19	7.7	20	0.27
6	0.20	2.6	5	0.33	5.4	13	0.39
7	0.20	2.6	6	0.37	4.4	9	0.27
8	0.25	3.8	14	0.28	7.6	16	0.46
9	0.25	1.8	4	0.25	3.3	7	0.38
10	0.31	2.0	9	0.12	4.0	11	0.25
11	0.40	2.6	7	0.33	3.3	8	0.26
12	0.57	3.9	14	0.25	6.2	15	0.33
13	0.62	2.7	9	0.18	5.5	14	0.29
14	0.71	2.7	11	0.26	5.2	15	0.25
15	0.78	3.9	15	0.22	4.5	12	0.23
16	0.80	4.3	11	0.33	7.3	21	0.39
17	0.83	4.3	15	0.19	5.8	17	0.21
18	0.96	4.5	18	0.18	6.5	25	0.17
19	1.00	4.6	15	0.31	6.9	23	0.28

Table 4.2. Characteristics of passing time (PT) and important matters (IM) networks by village. Characteristics of village networks include mean and maximum degree, and the global clustering coefficient for both the IM & PT networks. For a full description of these network variables, see the text.

	Model 1	Model 2	Model 3
Sociality Network Type	None	Passing Time	Important Matters
Demographics			
Age (decades)	**0.90 (0.84, 0.96)**	**0.90 (0.84, 0.96)**	**0.90 (0.85, 0.96)**
Village Size	**1.11 (1.03, 1.19)**	**1.05 (0.99, 1.10)**	**1.04 (0.99, 1.10)**
Wealth			
Ownership of material goods by household	0.86 (0.35, 2.12)	0.90 (0.37, 2.20)	0.86 (0.36, 2.09)
Pathogen Exposure			
Outside Contact			
% households w/outside visitor	**1.12 (1.00, 1.25)**	1.10 (0.99, 1.22)	1.08 (0.97, 1.21)
% households w/outside trip	1.03 (0.91, 1.16)	1.03 (0.92, 1.15)	0.96 (0.86, 1.08)
Food-Sharing Exposure	0.84 (0.45, 1.56)	0.84 (0.45, 1.55)	0.89 (0.48, 1.66)
Within-Household Exposure			
# Infected in household	**1.59 (1.41, 1.79)**	**1.55 (1.37, 1.74)**	**1.54 (1.36, 1.73)**
Mean-centered household size	**0.86 (0.81, 0.90)**	**0.86 (0.82, 0.91)**	**0.87 (0.82, 0.92)**
Contact Network Exposure			
# Infected Alters in PT network	0.91 (0.74, 1.11)	0.97 (0.80, 1.19)	0.95 (0.78, 1.16)
Sociality Network			
Household degree		0.64 (0.37, 1.10)	**0.59 (0.40, 0.85)**
Average degree		**0.89 (0.81, 0.98)**	**0.83 (0.72, 0.95)**
Avg. degree x household degree		1.06 (0.96, 1.17)	**1.17 (1.04, 1.32)**
Graph clustering		1.18 (0.94, 1.48)	1.12 (0.89, 1.42)
Goodness of fit			
Log-likelihood	-1045	-1038	-1037
AIC	2110	2107	2103

Table 4.3. Multivariate models for risk of infection in previous week. Results are presented as OR with 95% CI in parentheses. Statistically significant results ($p \leq .05$) are in bold. Column 1 presents sources of exposure risk, adjusted for household wealth as well as individual and village demographics. Columns 2 and 3 show protective effects of household and village-level characteristics of the passing time (PT) and important matters (IM) networks, respectively. Descriptions of these variables can be found in the text.

	Remoteness	Avg. PT Degree	Avg. IM Degree
Pathogen Exposure			
Outside Contact			
% households w/outside visitor	**0.061 (0.012)**	**-0.106 (0.001)**	**-0.047 (0.001)**
% households w/outside trip	0.009 (0.368)	**0.059 (0.008)**	**0.075 (0.043)**
Within-Household Exposure			
Mean-centered household Size	**0.016 (0.004)**	-	-
Wealth			
Ownership of material goods by household	0.006 (0.225)	-0.007 (0.179)	0.005 (0.143)
Social Capital			
Sociality Network			
Average degree (IM)	0.067 (0.156)	-	-
Average degree (PT)	**0.047 (0.054)**	-	-
Education			
Mean village years of education	-0.035 (0.123)	**0.065 (0.004)**	**0.054 (0.043)**
Participation in Community Organizations			
Mean # of org. memberships in village	**-0.046 (0.081)**	**0.059 (0.020)**	0.038 (0.211)
Max # of org. memberships in HH	0.001 (0.209)	0.007 (0.167)	**0.006 (0.084)**
Water Quality & Sanitation			
Observed hygiene index	0.005 (0.181)	**0.134 (0.001)**	**0.093 (0.010)**
Community improved sanitation	**0.042 (0.012)**	**-0.042 (0.004)**	**0.039 (0.001)**
Community water treatment	**0.072 (0.04)**	0.002 (0.205)	-0.018 (0.333)
Community water source	-0.032 (0.146)	-0.024 (0.222)	-0.019 (0.281)

Table 4.4. Indirect effects of remoteness and village-level average degree on risk of illness. Table entries indicate the indirect effect, with bootstrap based *p*-values in parentheses. Positive values indicate mediation, while negative values indicate suppression. Statistically significant results ($p \leq .1$) are in bold.

References

1. Klovdahl AS, Graviss EA, Yaganehdoost A, et al. Networks and tuberculosis: an undetected community outbreak involving public places. Social Science and Medicine 2001;52:681-94.
2. Jolly A, Muth S, Wylie J, et al. Sexual networks and sexually transmitted infections: a tale of two cities. Journal of Urban Health 2001;78:433-45.
3. Newman M. Spread of epidemic disease on networks. Physical Review E 2002;66.
4. Trostle JA, Hubbard A, Scott J, et al. Raising the Level of Analysis of Food-Borne Outbreaks. Epidemiology 2008;19:384-90.
5. Eisenberg J, Cevallos W, Ponce K, et al. Environemntal change and infectious disease: How new roads affect the transmission of diarrheal pathogens in rural Ecuador. Proceedings of the National Academy of Sciences 2006;103:19460-5.
6. Ostrom E. Collective action and the evolution of social norms. The Journal of Economic Perspectives 2000;14:137-58.
7. Christley R, Pinchbeck G, Bowers R. Infection in social networks: using network analysis to identify high-risk individuals. American Journal of Epidemiology 2005.
8. Pahl-Wostl C, Craps M, Dewulf A, et al. Social learning and water resources management. Ecology and Society 2007;12:5.
9. Hunt RC, Hunt E. Canal irrigation and local social organization. Current Anthropology 1976;17:389-411.

10. Wallace R. A Synergism of Plagues: "Planned Shrinkage," Contagious Housing Destruction, and AIDS in the Bronx. Environmental Research 1988;47:1-33.

11. Meyers L, Newman M, Martin M, et al. Applying network theory to epidemics: Control measures for Mycopasma pneumonia Outbreaks. Emerging Infectious Diseases 2003;9.

12. Aparicio J, Pascual M. Building epidemiological models from Ro: an implicit treatment of transmission in networks. Proceedings of the Royal Society: B 2007;274:505-12.

13. Bansal S, Grenfell BT, Meyers LA. When individual behavior matters: homogeneous and network models in epidemiology. Journal of the Royal Society Interface 2008;4:879-91.

14. House J, Landis K, Umberson D. Social Relationships and Health. Science 1988;241:540-5.

15. Berkman LF, Glass T. Social integration, social networks, social support, and health. In: Berkman LF, Kawarchi I, eds. Social Epidemiolgy. New York: Oxford University Press, 2000.

16. Holtgrave D, Crosby R. Social capital, poverty, and income inequality as predictors of gonorrhoea, syphilis, chlamydia and AIDS case rates in the United States. Sex Transm Infect 2003;79:62.

17. Cohen JM, Wilson ML, Aiello A. Analysis of social epidemiology research on infectious diseases: historical patterns and future opportunities. Journal of Epidemiology and Community Health 2007;61:1021-7.

18. Checkley W, Gilman RH, R.E. B, et al. Effect of water and sanitation on childhoold health in a poor Peruvian peri-urban community. Lancet 2004;363:112-8.

19. Rego RF, Moraes LR, Dourado I. Diarrhoea and garbage disposal in Salvador, Brazil. Trans R Soc Trop Medi Hyg 2005;99:48-54.

20. Tumwine JK, Thompson J, Katua-Katua M, et al. Diarrhoea and effects of different water sources, sanitation and hygiene behaviour in East Africa. Trop Med Int Health 2002;7:750-6.

21. Koelle K, Pascual M. Disentangling extrinsic from intrinsic factors in disease dynamics: A nonlinear time series approach with an application to Cholera. American Naturalist 2004;163:901-13.

22. Chang M, Groseclose SL, Zaidi AA, et al. An ecological analysis of sociodemographic factors associated with the incidence of salmonellosis, shigellosis, and E. coli O157:H7 infections in US counties. Epidemiology and Infection 2009;137:810-20.

23. Troesken W. Race, disease, and the provision of water in American cities, 1889–1921. The Journal of Economic History 2002;61:750-76.

24. Anderson RM, May RM. Infectious Diseases of Humans: Dynamics and Control. Oxford: Oxford Science Publications, 1992.

25. WHO. Persistent diarrhoea in children in developing countries: memorandum from a WHO meeting. . Bulletin of the World Health Organization 1988;66:709-17.

26. Bates S, Trostle J, Cevallos W, et al. Relating Diarrheal Disease to Social Networks and the Geographic Configuration of Communities in Rural Ecuador. American Journal of Epidemiology 2007;166:1088-95.

27. Laird N, Ware JH. Random-effects models for longitudinal data. Biometrics 1982;38:963-74.

28. Stiratelli R, Laird N, Ware JH. Random-effects models for serial observations with binrary response. Biometrics 1984;40:961-71.

29. Burnham K, Anderson D. Model selection and inference: a practical information-theoretic approach. New York: Springer-Verlag, 1998.

30. Baron R, Kenny D. The moderator-mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. Journal of Personality and Social Psychology 1986;51:1173-82.

31. House J. Understanding Social Factors and Inequalities in Health: 20th Century Progress and 21st Century Prospects. Journal of Health and Social Behavior 2002;43:125-42.

32. Klinenberg E. Heat Wave: A Social Autopsy of Disaster in Chicago. Chicago: University of Chicago Press, 2002.

33. Wallace D, Wallace R. A Plague on Your Houses: How New York Was Burned Down and National Public Health Crumbled. New York: Verso, 1998.

APPENDIX C

Chapter 4 Supplementary Materials

S1. Household Ownership Scale.

Household wealth is assessed using a scale that assesses ownership of material goods. Items are weighed according to their relative value, so that, e.g., ownership of the family's home makes a greater contribution to overall wealth than ownership of a motorcycle or television. The weights are listed in Table 4.4.

Ownership of	**Weight**
House they live in	4
Farm	3
Animals (cows)	3
Motorcycle	3
Solar	1
Motorcycle	3
Canoe	2
Chainsaw	2
Bicycle	2
Cellular Telephone	2
Television	2
Gas Stove	2
Refrigerator	2

Blender	1
Stereo	1
Sewing machine	1
DVD Player	1
Computer	2

Table 4.5

The scale is computed for each household by summing the weights corresponding to the items in table S1 owned by ego's household. This value is then divided by the sum of all the weights to obtain a normalized value, from 0 to 1, for ego's household.

S2. Clustering coefficient of village sociality network.

The clustering coefficient of an entire network is the probability that an open triple in the graph will be closed. A closed triplet is a group of 3 nodes, *a,b,c* each of which is connected to the other two. An open triple is then defined as 3 nodes, *a,b,c* where *a* is connected to *b* and *c*. The clustering coefficient is the probability that *b* and *c* are connected, given that they are both connected to *a*, and is computed as:

C = # of closed triplets / (# of closed triplets + # of open triples)

Equation 4.1

Higher levels of overall clustering in a sociality network may indicate cohesion at the village level, indicating an increased propensity for individuals to form relationships with

those connected to their other alters (*1*). In addition, increasing average degree is often associated with an increase in the clustering coefficient of the graph. In transmission networks, holding average degree constant, increased clustering is generally associated with decreased population-level outbreak risk (*2*). Consequently, it is important to adjust for clustering when assessing the effect of average degree.

S3. Additional measures of village social cohesion and socioeconomic status.

Village	Organization Memberships	Years of Education	Household Wealth
	Mean	Mean	Median
1	0.50	4.4	3.51
2	1.10	5.8	3.51
3	0.94	5.4	3.24
4	0.86	5.5	4.05
5	1.22	5.2	4.59
6	1.08	5.0	3.38
7	0.92	3.5	3.78
8	1.20	3.2	5.27
9	1.04	4.6	3.51
10	0.29	4.4	4.19
11	0.59	3.6	4.05
12	0.93	3.8	3.78
13	1.13	3.3	3.78
14	1.13	4.4	2.97
15	1.25	3.9	3.51
16	1.31	5.5	3.65
17	1.07	4.4	3.24
18	0.88	5.1	4.05
19	1.70	5.2	3.24

Table 4.6

Table S2 presents village-level measures of social cohesion, and socioeconomic status (SES), ordered by village remoteness. The mean number of memberships in community organizations by individuals older than 13 years is an additional measure of community cohesion. The mean number of years of education by individuals older than 13 years, and the average score of households in the village on the wealth scale presented in Table S1 are measures of SES.

S4. Interactive effect of household and village-level average degree on risk.

In our data, there is a joint effect of village average important matters (IM) degree and household IM degree on risk of illness. In villages with high average degree, individuals are always protected regardless of the degree of their household. But in villages where average degree is lower, household degree becomes protective. This relationship is analogous to herd immunity obtained through high vaccine coverage. It is illustrated in Figure 4.2, below:

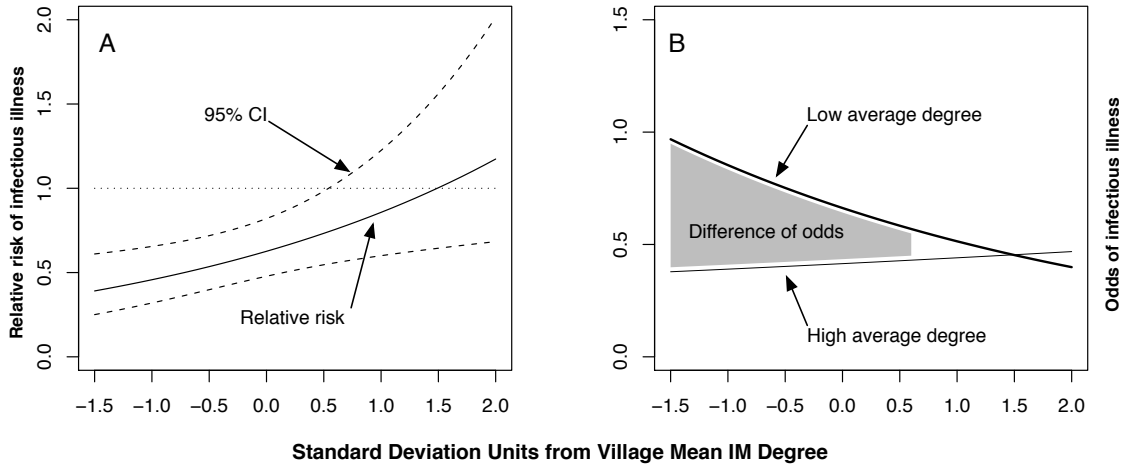

Figure 4.2

In each panel of the figure above, values of household degree, on the x-axis, are normalized so that 0 represents the mean household degree in ego's village and the units are standard deviations from the mean. The 75th and 25th percentile of average IM degree (3.8, 1.7, respectively) are used as the thresholds for 'high' and 'low' in each plot. Panel A shows relative risk of illness, comparing individuals in high vs. low average

degree villages. The solid line is the relative risk and the dashed lines are pointwise 95% confidence intervals. Panel B shows the odds of infectious illness stratified by 'high' and 'low' average degree. The shaded region of the plot displays the difference of odds where this difference is statistically significant.

This relationship highlights the fact that connectedness at the household level can also be protective against risk. Because households are embedded in a hierarchical system in which household-level risk is strongly impacted by the quality of village-level infrastructure, this relationship is only operative in those villages in which levels of social cohesion do not foster consistent, high-quality water sanitation and hygiene practices.

S4. Calculation of Mediation Effects. The adjusted effect of remoteness is quantified by:

$$R = E(Y \mid D, Far) / E(Y \mid D, Near),$$

Equation 4.2

Where D denotes the mediator. The corresponding 'unadjusted' effect of remoteness is defined analogously as:

$$R' = E(Y \mid Far) / E(Y \mid Near).$$

Equation 4.3

Each of these quantities is estimated by the average of the sample predictions of probability of disease produced by non-linear mixed effects (NLME) models that have been fit to our illness data using the lme4 package in R 2.10. The indirect effect of remoteness acting through a mediator, D, is then quantified by $\hat{m}$ = R-R'. Negative values suggest that the factor is a suppressor, i.e., something that buffers or inhibits the effect of remoteness. This analysis was conducted separately for each potential mediator so that the pathways through which remoteness is operating can be assessed. Although a joint mediation analysis would be equally simple to perform, it would be more difficult to discern the pathways by which remoteness operates on risk.

In this analysis, positive values of *m* indicate mediation, and negative values indicate suppression. This is because the effect of living in a far village is protective

against illness, so the ratio of the average probability of infection for individuals living in far villages vs. those in near villages, R', is less than 1. If adjusting for the mediator eliminates the ability of residence in a in a far vs. near village to discriminate between illness outcomes, i.e. full mediation, the adjusted effect of remoteness, R, would then equal 1 and R - R' > 0. So, where $m > 0$, we are seeing mediation. The choice of the ratio in equations 4.2 & 4.3 is arbitrary and should not impact the qualitative results of the mediation analysis. If we were to assess mediation in terms of the ratio of near vs. far, however, positive values of *m* would then indicate suppression while negative values would indicate mediation.

The sampling distribution of our estimator of *m* is not analytically tractable, so we obtain standard errors using a bootstrap resampling procedure (e.g. (*3*)). Estimates are obtained via bootstrap resampling and refitting of the NLME models to these sampled datasets, which approximates the sampling distribution of *m*. For each covariate listed in Table 4, the full dataset was resampled with replacement 10^3 times, and the value of $\hat{m}$ for each sampled dataset was assessed using a random-intercept mixed effects model grouped by village, as described above. Positive values of $\hat{m}$ indicate mediation of remoteness by the selected factor, whereas negative values indicate suppression or buffering. *P*-values for significance testing are produced by inspecting the quantiles of the bootstrap estimates.

References

1. Goodreau S, Kitts J, Morris M. Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. Demography 2008;46:103-25.

2. Kenah E, Robins JM. Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing. Journal of Theoretical Biology 2007;249:706-22.

3. Efron B, Tibshirani R. Bootstrap Methods for Standard Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science 1986;1:54-75.

Chapter 5

Conclusion

This dissertation addresses a number of challenges involved in fostering and sustaining an integrated sociological and social epidemiological approach to infectious diseases. In the introduction, we focused on the theoretical issues that have prevented growth in this area of inquiry. Here, I will discuss some equally important methodological challenges and future directions through the lens of the three projects that comprise this thesis.

Interfacing dynamic models and data

In the first paper, we obtained transmission parameter estimates for household transmission of norovirus following a point-source outbreak event in which a food handler infected many people at once. This project underscores the challenges of interfacing dynamic models with data in a meaningful way. Its success is owed as much to the relative cleanliness of the data, and the point-source event that brought them about, as it does to the modeling choices made when analyzing the data. Explaining transmission systems that involve more moving parts, particularly the types of social behavior presented in our model of community-level norovirus outbreaks, or in other relevant social models, such as those of residential segregation, e.g. (1), involves an escalation of this technical challenge. The increased dimensionality of dynamic models

that involve both social behavior and disease natural history makes defining a likelihood function measuring the correspondence between model and data difficult.

In general, there are two alternative approaches to inference in models that have many moving parts, both of which are simulation based. These alternative approaches have been described variously as 'equation-free' and 'plug-and-play' techniques, e.g., (2). The equation-free approach is still within the frame of likelihood-based inference and is based on the work of Ionides, Breto & King (3). Their approach involves simulating from the theoretical model one wishes to use to explain the data. Model parameters are estimated using the likelihood of the observed data in these simulated observations. Ionides et al. developed a novel filtering technique that allows one to find the maximum likelihood (ML) model parameters without resorting to brute-force simulation of the universe of parameter combinations. But generating enough simulations to obtain a stable likelihood is nonetheless a computationally intensive affair that only becomes more difficult with models of increasing scale and scope, in terms of population size and the range of behaviors they can address.

The second set of simulation-based approaches dispenses with ML estimation and focuses on optimizing the match between simulated values from the model and measures of the target dataset, e.g. (4) using some measure of the distance between the model and the data. This approach is somewhat analogous to the method of moments technique in which the fit of a statistical distribution to a dataset is assessed by the correspondence between the moments – mean, standard deviation, etc – of each, rather than the likelihood of the observed data, given the parameters of the distribution. These likelihood-free approaches to fitting dynamic models to data can be informative and often necessitate

fewer stochastic samples than likelihood-based methods like that of Ionides et al. However, like the method of moments, likelihood-free approaches lack the efficiency of maximum-likelihood estimation and are susceptible to bias.

These simulation-based, equation-free approaches to inference are compelling because they blur the line between the approach presented in paper one, which is focused on fitting a theoretical model to data, and the approach taken in paper two, where we analyze data generated by a more involved community outbreak model.

From the plug-and-play perspective on inference, there is no theoretical reason preventing us from using the simulation model in paper two for both exploratory analysis and maximum-likelihood estimation of parameters for an observed outbreak. However, this does not mean that we can dispense with care in the process of model building and continually throw ideas against the wall until one sticks. Dynamic models are as susceptible to overfitting and multiple testing bias as any other statistical model and should be approached with the care and forethought that this implies. Instead, by loosening some of the constraints on fitting the model to data, we put the burden precisely on the process of design and theoretical development as opposed to working around inferential constraints. For example, such approaches open up the possibility of developing dynamic models of infectious disease transmission that include the kinds of layered social behavior implied by the results of the third paper. The greater flexibility of this computational approach does of course come at a price, because it involves much greater demand on computational resources than more analytical approaches.

Massively parallel epidemiology

One of the more exciting computing advances of recent years is the increasing accessibility and affordability of massively parallel computing. These tools allow us to take many samples at the same time, sometimes numbering in the hundreds or thousands, from computationally intensive stochastic models, rather than drawing these samples in series and waiting a long time for the result. Accessible parallel computing has the potential to improve the modeling enterprise at two both the theoretical and public health policy levels. At the level of theoretical development, models that run faster can speed development and testing; at the level of public health policy and practice, faster runtimes mean that these models can be used to understand the progression of evolving outbreaks and devise and test potential interventions in real time, e.g. (5).

Model development

In keeping with their computational underpinnings, the process of developing the kinds of mechanistic models discussed here is typically more of a loop than a linear process. Model building is usually iterative: models are tested, refined, re-built and re-tested. The program implementing the model should facilitate rather than hinder this process of creative destruction. Consequently, implementing the ideas expressed by the theoretical model in as transparent and flexible way as possible is challenging, as is measuring and optimizing the correspondence between model and data. Despite this, the process of implementing the model is often regarded as secondary to addressing the theoretical issues the model is aimed at. This often results in poorly written code that allows the program to be mistaken for the model. Thoughtful, modular code allows theoretical and empirical considerations determine modeling decisions.

For builders of dynamic models, this process of theoretical refinement is typically enmeshed with, and limited by, the process of writing, maintaining and re-writing computer code. For example, it is not unusual for models to be coded in a procedural way in which the events that occur in the model happen one after another, rather than putting model elements into modular components that can be modified independently without breaking the entire program. This procedural approach makes it more likely that theoretical assumptions will become baked into the model code and become progressively less likely to be subject to further refinement or revision over time.

This kind of procedural programming often comes out of a single-minded focus on the optimization of model code for speed. When code is inefficient and simulations take a very long time to run, modelers may be tempted to cut their losses by looking at only a narrow band of the model's behavior, or take too few samples at each point in the model's parameter space to draw meaningful conclusions. This undermines the exploratory and empirical potential of the model, because the implications of the ideas encoded by the model cannot thoroughly explored. Consequently, we end up taking too few trips through the theory-simulation-refinement loop and end up with final products that are not as strong as they otherwise might be. Completing this loop and developing effective and meaningful models that integrate social and biological dynamics requires a willingness to embrace an approach to computation that is as sophisticated as our theoretical ideas and statistical techniques.

Future directions

The third paper of this dissertation poses but does not directly address the issue of integrating the effects that networks of social relationships have on exposure of entire populations to pathogens, inside of a unified model. Approaching this challenge involves drawing together the lessons of each of the projects that comprise this dissertation. These projects build up to progressively greater levels of scale and aggregation, from households to the regional level.

The next step involves working backwards through the concepts motivating each of these projects and augmenting and refining them to address these ambitious goals. In the third paper, we highlight the nuance involved in understanding the role of different types of social relationships in transmission risk. Understanding these issues is important for designing useful dynamics models including the role of social action in community and region-level infectious disease transmission. In the second paper, we developed and implemented a framework that allows for multiple types of social behavior in the context of an outbreak simulation model and that should be expanded to deal with the kind of social behavior that conditions population-level susceptibility.

Finally, in the first paper, we explained household transmission subsequent to a point-source outbreak using a likelihood function derived from a dynamic outbreak model. In order to apply the insights of the second and third chapters to real-world data, we will need to go beyond this approach and use techniques, such as the plug-and-play methods discussed above, to develop our understanding about the correspondence between behavior, biology and infectious disease outcomes.

References

1. Bruch E, Mare R. Neighborhood Choice and Neighborhood Change. American Journal of Sociology 2006;112 667-709.
2. He D, Ionides EL, King AA. Plug-and-play inference for disease dynamics: measles in large and small populations as a case study. Journal of the Royal Society Interface 2010;7:271-83.
3. Ionides E. Inference and filtering for partially observed diffusion processes via sequential Monte Carlo. Submitted to Journal of Computational and Graphical … 2004.
4. Marjoram P, Molitor J, Plagnol V, et al. Markov chain Monte Carlo without likelihoods. Proceedings of the National Academy of Sciences 2003;100:15324-8.
5. Staes CJ, Xu W, LeFevre SD, et al. A case for using grid architecture for state public health informatics: the Utah perspective. BMC Medical Informatics and Decision Making 2009;9:32.