SANDIA REPORT

SAND2008-6044
Unlimited Release
Printed October 2008

Distributed micro-releases of bioterror pathogens: threat characterizations and epidemiology from uncertain patient observables

J. Ray, B. M. Adams, K. D. Devine, Y. M. Marzouk, M. M. Wolf, and H. N. Najm

Prepared by
Sandia National Laboratories
Albuquerque, New Mexico 87185 and Livermore, California 94550

Sandia is a multiprogram laboratory operated by Sandia Corporation,
a Lockheed Martin Company, for the United States Department of Energy's
National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

Issued by Sandia National Laboratories, operated for the United States Department of Energy by Sandia Corporation.

Available to DOE and DOE contractors from
U.S. Department of Energy
Office of Scientific and Technical Information
P.O. Box 62
Oak Ridge, TN 37831

Telephone: (865) 576-8401
Facsimile: (865) 576-5728
E-Mail: reports@adonis.osti.gov
Online ordering: http://www.doe.gov/bridge

Available to the public from
U.S. Department of Commerce
National Technical Information Service
5285 Port Royal Rd
Springfield, VA 22161

Telephone: (800) 553-6847
Facsimile: (703) 605-6900
E-Mail: orders@ntis.fedworld.gov
Online ordering: http://www.ntis.gov/ordering.htm

SAND2008-6044
Unlimited Release
Printed October 2008

Distributed micro-releases of bioterror pathogens: threat characterizations and epidemiology from uncertain patient observables

J. Ray, B. M. Adams, K. D. Devine, Y. M. Marzouk and H. N. Najm
Sandia National Laboratories, P. O. Box 969, Livermore CA 94551
and
Michael M. Wolf
4332 Siebel Center, MC-258, 201 N. Goodwin,
University of Illinois, Urbana-Champaign
Urbana, IL 61801
jairay,briadam,kddevin,ymarzou,hnnajm@sandia.gov
mmwolf@uiuc.edu

Abstract

Terrorist attacks using an aerosolized pathogen preparation have gained credibility as a national security concern since the anthrax attacks of 2001. The ability to characterize the parameters of such attacks, i.e., to estimate the number of people infected, the time of infection, the average dose received, and the rate of disease spread in contemporary American society (for contagious diseases), is important when planning a medical response. For non-contagious diseases, we address the characterization problem by formulating a Bayesian inverse problem predicated on a short time-series of diagnosed patients exhibiting symptoms. To keep the approach relevant for response planning, we limit ourselves to 3–5 days of data. In computational tests performed for anthrax, we usually find these observation windows sufficient, especially if the outbreak model employed in the inverse problem is accurate. For contagious diseases, we formulated a Bayesian inversion technique to infer both pathogenic transmissibility and the social network from outbreak observations, ensuring that the two determinants of spreading are identified separately. We tested this technique on data collected from a 1967 smallpox epidemic in Abakaliki, Nigeria. We inferred, probabilistically, different transmissibilities in the structured Abakaliki population, the social network, and the chain of transmission. Finally, we developed an individual-based epidemic model to realistically simulate the spread of a rare (or eradicated) disease in a modern society. This model incorporates the mixing patterns observed

in an (American) urban setting and accepts, as model input, pathogenic transmissibilities estimated from historical outbreaks that may have occurred in socio-economic environments with little resemblance to contemporary society. Techniques were also developed to simulate disease spread on static and sampled network reductions of the dynamic social networks originally in the individual-based model, yielding faster, though approximate, network-based epidemic models. These reduced-order models are useful in scenario analysis for medical response planning, as well as in computationally intensive inverse problems.

Acknowledgment

This work was supported Sandia National Laboratories' LDRD (Laboratory Directed Research and Development) funds, sponsored by the Computational and Information Sciences Investment Area.

Contents

Appendix

Figures

Tables

1 Introduction

The anthrax attacks of 2001 [3] are generally cited as the impetus for raising the specter of rare pathogens as a terrorism tool. Yet pathogens have seen extensive use in war. There are recorded accounts of malicious smallpox-infected blanket distribution during the French and Indian Wars (1756-1763) and attempts to spread veterinary diseases among pack mules and horses used in front-line positions during World War I [4]. Also, the Japanese Army, during World War II, included a unit devoted to biowarfare [5] and the Soviet Union allegedly weaponized pathogens on an industrial scale [6]. The Sverdlovsk anthrax accident, where an aerosolized anthrax preparation was inadvertently released from a military institution [7] provided one example of the potential effect of an outdoor aerosolized pathogen release; the "Amerithrax" attacks [3] provided a bioterrorism counterpart. Thus people's ability to weaponize pathogens and intent to use them aggressively is not in doubt.

Pathogenic preparations have both tactical and strategic uses. Large attacks with a non-contagious disease can severely degrade human population viability in a tightly circumscribed theater of war, while a contagious disease may spread uncontrollably, seriously (and unpredictably) disrupting a nation's operation. It has been estimated that a smallpox attack infecting a small, but significant, fraction of a country's population could completely undermine its war effort – this decimation would primarily be effected by the public health measures (social distancing and quarantine) required to combat spread and the required care of infected people, rather than by any widespread morbidity due to the disease [4]. The threat posed by weaponized pathogens should not be taken lightly.

While prevention of a bioattack against a civilian population remains the obvious preferred option, the question of how to best mount a medical response is never far behind. This was investigated first in the "Dark Winter" exercise [8] and thereafter in many TOPOFF ("Top Official") exercises conducted by the Department of Defense. "Dark Winter," which investigated the mechanics of mounting a medical response to a smallpox attack on an American city, revealed that detecting a bioattack and identifying its causative agent were insufficient. Logistics planning (for medical personnel and infrastructure) required knowledge of the number of infected individuals and the rate of disease spread. Further, this information would be required quite soon after detection to mount a timely response. Since such information is not readily available, any estimates provided would be "rough."

The difficulty in deciding the response parameters has two origins. Many pathogens considered for bioattack use, e.g., *Bacillus anthracis*, rarely cause diseases in humans and never in epidemic numbers; hence there is no "prior art" regarding countermeasures. Others, e.g., *Yersinia pestis* (plague) and *Variola major* (smallpox) are old human scourges, but any epidemic data were collected decades ago in regions with socio-economic characteristics drastically different from those in contemporary American society, so it is not clear that similar countermeasures would be applicable or effective today. Substantial evidence indicates that socio-economic factors, or more specif-

ically, social mixing patterns, play an enormous role in determining how quickly and broadly a disease spreads. For example, the basic reproductive ratio, R_0, (the average number of susceptibles infected by a single infectious person) for the 1972 Yugoslav smallpox outbreak was around 5.4 [9] (before countermeasures were introduced) and 17 within the confines of a West German hospital during winter [10], yet an average value of 3.0 is generally used, given the preponderance of data collected in rural areas of the Indian subcontinent during the 1960s and 1970s [9, 11]. Therefore, while historical outbreaks may provide guidance regarding medical and public health responses, whole-hearted singular reliance on them would be foolhardy.

Despite the sparsity and lack of direct applicability of recorded data, the technical issues behind questions raised by the "Dark Winter" exercise are clear. There is a need to create models of rare pathogen behavior in individuals that capture the diversity, i.e., the stochastic variability, of humans. Then, if some individuals were infected, stochasticity would guarantee that a small sample of those infected would develop symptoms early and be diagnosed. The problem then reduces to inferring/estimating the characteristics of the (unseen) infected population from the small (diagnosed) sample drawn from it. This inference process is not, theoretically, far-fetched. For contagious diseases, it relies on constructing models that characterize pathogenic transmissibility separately from social mixing. Such models, calibrated to historical outbreaks, permit the estimation of transmissibility (and the social mixing model, though that can be discarded). This calculated transmissibility, in conjunction with a social mixing model for contemporary society, could be used to predict epidemic evolution with a reasonable degree of confidence. Finally, there is the obvious requirement to construct a model of the social mixing observed in contemporary society, for this purpose.

These are precisely the aims of our research effort. Given the paucity of data, any estimates drawn will contain significant uncertainties; prudence dictates these be quantified. This clearly demands a statistical approach; we adopt a Bayesian approach and develop parameter estimates as probability distribution functions. For contagious diseases, we adopt a Poisson process-based model of disease transmission and social network models of social mixing; these allow us to directly gauge the applicability of mixing parameters that we infer from historical epidemic records, while cleanly separating the pathogenic transmissibility from social factors of disease spread. Finally, we approach the problem of constructing an epidemic model, commensurate with contemporary American society, with an individual-based technique; this simplifies the comparison to real life for validation purposes. While this may appear tedious and intractable, current literature offers much help.

Methodologically, the creation and "calibration" of the models present some stiff algorithmic issues. Individual-based urban population models can be large and unwieldy so we appeal to parallel computing. Scalable algorithms for individual-based models is an emerging field; we explore and develop new techniques in this work. Markov Chain Monte Carlo (MCMC) techniques are an efficient way to solve the statistical inverse problems arising in this context and will be adopted for the purpose; however, the high-dimensionality of the contagious-disease problem, especially for inference of social networks, presents novel challenges. Our research effort therefore has both modeling and algorithmic contributions in equal measure.

In the remainder of this report, we describe the inferential and modeling capabilities developed and

their performance with both simulated and recorded data. In Section 3 we describe our individual-based epidemic model and its approximations, which were developed for computational celerity. In Section 4, we formulate and solve an inverse problem to estimate the characteristics of an infected population, given a small sample. In Section 5, we develop a technique that estimates both pathogenic transmissibility and a social network from observations of a smallpox epidemic. We conclude in Section 6 and assess the extent to which we achieved our research goals. Throughout this report, anthrax serves as a prototypical non-contagious disease; smallpox as its contagious counterpart.

2 Literature review

In this section, we survey available literature pertinent to the problems considered here. This will be done separately for the three sub-problems under investigation: the individual-based epidemic simulation, the characterization of non-communicable disease outbreaks and finally, that of communicable diseases. We finish with a short discussion of existing literature on Markov Chain Monte Carlo methods, which are used to solve the inverse problems arising in our work.

2.1 Individual-based models of epidemics

Epidemiologically-based mathematical models and associated computer simulations are widely used to understand historical disease outbreaks. Given sufficient characterizations of pathogen dynamics and transmissibility, they can be used to predict the severity of future epidemics and the impact of potential interventions. Perhaps the most prolific epidemic models are compartmental differential equation models which divide people into bulk categories such as susceptible, infectious, and recovered (immune or dead), SIR models. Transition between the compartments is then based on contact rates between susceptible and infectious people and average recovery time. Simplifying a diverse population in this manner yields a system of differential equations amenable to analysis and useful for assessing key outbreak features as a function of time. A good overview of such models, including mathematical analysis techniques and relevance for policy making can be found in [12].

Compartmental models predict the time evolution of population fractions in various disease stages, under mass-action (well-mixed population) assumptions. They can address questions about the rate of spread of a disease, whether it will become endemic, and how one might control seasonal or other cyclic epidemic waves. Potential enhancements to basic SIR models include: adding disease progression stages (e.g., to create S-Exposed-I-R models), tracking multiple cohorts of people (e.g., to add age structure or account for immunocompromised individuals), and explicitly modeling vaccinated (people with partial or full immunity) or quarantined groups.

A central goal of this project is to use extremely limited observations to invert epidemic models, characterizing outbreak sources. Most aggregate population models (stochastic ODE models possibly excepted) are insufficient in this regime, where only a small fraction of the population is infected and potential transmission paths need to be analyzed to reveal likely index and secondary cases. Agent-based simulation models, with social contact networks and in-host pathogenesis models at their core, offer one means to predict epidemic scenarios while tracking individuals and detailed transmission paths.

A number of intermediate model types accounting for social contact structure, but short of fully agent-based simulations, are possible. One approach involves constructing a large differential equation system, with one or more equations explicitly modeling each individual's disease state and coupling via a static transition matrix weighting links between pairs of individuals [13]. This could be viewed as analogous to electrical circuit network simulations or a continuum model similar to the static network model presented in Section 3.1. It includes sufficient detail to study the effect of

social distance between people, as represented in pairwise transmission force.

Lloyd and co-authors offer a gentle introduction to epidemic models on network topologies, and analysis comparing the potential effect of local versus global mixing. [14, 15] These papers bridge the gap from mass action to network-based models and illustrate challenges in analyzing the effect of social structures. Many have addressed disease spread on variously connected networks, including the authors of [16, 17, 18] and [19], some of whom use census data to inform social network construction. The importance of considering contact heterogeneity is emphasized in [20], where given a particular disease's basic reproductive ratio R_0 (a dimensionless measure often used to quantify spread in a population), a variety of epidemiological outcomes may be realized. Like compartmental models, contact epidemiology-based models can be used to assess containment strategies (quarantine or prophylaxis) or vaccination impact. See *The structure and function of complex networks* [21] for a thorough discussion of network analysis, [22] for complex network metrics capturing the most salient topological features, and [23] for algorithmic generation and analysis of the social networks used in our present work.

Many of the papers cited above consider analysis of static network structures, and the correlation between complex network metrics and summary epidemic outcomes. We now turn to truly individual- or agent-based time-stepped simulations. They offer an alternate means to assess disease spread in different societal structures, though have associated validation challenges (as discussed in Section 6). These models typically have social contact network structure and detailed in-host pathogenesis models, but could also include agent cognition models representing human behavior during an epidemic. In [24], SEIR type models are explicitly compared to agent-based models of the same phenomena, and effect of assumptions on network structure explored. The EpiSims model for smallpox transmission and its associated contact network [25] is an early example on which our simulation is based, and among the first simulations of its scale.

Several stochastic individual-based models are presented with a strong emphasis on control strategies. An age-dependent probability of transmission model is used in [26] to model pandemic influenza spreading among 281 million U.S. citizens. Burke, et al. [27] and Longini, et al. [28] also consider social structures, but use estimates of smallpox characteristics from a recent expert panel and, in part, calibrate smallpox models to historical data, before assessing vaccination strategies. Agent-based simulations can leverage substantial computing power to model with extreme detail, capturing the subtle effect of social connectivity and performing precise scenario analysis.

2.2 Characterization of anthrax outbreaks

Bacillus anthracis is an aerobic Gram-positive, spore-forming nonmotile *Bacillus* species. The non-flagellated vegetative cell is about 1-8 μm in length and 1-1.5 μm in width. Spores are approximately 1 μm in size and grow readily on laboratory media [29]. Anthrax spores germinate when situated in a media rich in amino acids, nucleotides and glucose (e.g. blood and tissues of humans). When vegetative cells run out of nutrients, they form spores. Vegetative cells are not robust; they disappear almost completely within 24 hours of being injected into water [30]. Spores, on the other hand, are hardy and can survive for decades [31].

Inhalational anthrax follows deposition of spore-bearing particles (1-5 μm in size) in the alveolar spaces. Macrophages ingest the spores which are mostly destroyed. The survivors are transported via lymphatics to the mediastinal lymph nodes where germination can occur up to 60 days later [32, 33]. In Sverdlovsk, cases occurred from 2 to 43 days after exposure [7].

Few studies have used statistical methods to characterize the genesis of a partially observed epidemic. Walden & Kaplan [34] introduced a Bayesian formulation for estimating the size and time of a bioterror (BT) attack and tested it on a low-dose (less than ID_{25}, the dose at which a person has a 25% probability of incurring the disease) anthrax release corresponding, approximately, to the Sverdlovsk outbreak [7] of 1979. Their formulation incorporated an incubation period model developed by Brookmeyer *et al.* [35] and demonstrated the use of prior distributions on N to reduce uncertainty in the inferred characteristics. Brookmeyer & Blades [36] used a maximum likelihood approach, along with the anthrax incubation model in [35], to infer the size of the 2001 anthrax attacks [3] before estimating the reduction in casualties due to the timely administration of antibiotics. Both [34] and [36] developed similar expressions for the likelihood function, i.e., the probability of observing a patient time series given an attack at time τ with N infected people. The incubation period model in [35] was not dose-dependent, and hence no doses were inferred in these two studies.

Significantly more effort has been spent in characterizing the incubation period of inhalational anthrax. Most work has been experimental, with non-human primates subjected to anthrax challenges [1, 32, 37, 38, 39, 40]. Brookmeyer *et al.* [35], on the other hand, used data from the Sverdlovsk outbreak to fit a log-normal distribution of incubation periods valid at low doses; their more recent work, based on a competing risks formulation, includes dose-dependence [41]. Wilkening [42] compares four dose-dependent models for the incubation period distribution, one of which (termed Model D) is structurally identical to Brookmeyer's [41], with updated parameters. Compared to Model D, Wilkening's Model A2 provides slightly better agreement with the spatial and temporal distribution of anthrax cases observed in Sverdlovsk; the median incubation period predicted by Model D is consistently larger than that predicted by Model A2 (see Figure 7). Yet experimental results by Ivins *et al.* [40] and Brachman *et al.* [1] show significant departures from the results of both models, especially in the 10^3–10^4 spore dose range (see Figure 7). Thus both A2 and D must be considered approximate, though useful, predictive tools. In this study, we explore the impact of model error by using Model D to simulate bioterror attacks while using Model A2 for inference. A more detailed discussion of the anthrax incubation period models is provided in Section 4.2.

The issue of dose-response functions, which indicate whether a person exposed to a number of spores will actually contract the disease, will not be addressed in this study. We concentrate on inferring the number of people who are actually infected, not merely exposed to the pathogen. The problem of estimating the probability of infection from D spores was addressed by Brookmeyer *et al.* [41] as well as by Glassman [43] and Druett *et al.* [44]. Haas [45] has established that exposure to low doses can still pose a statistically significant risk to large populations.

The BARD effort [46] also seeks to characterize a BT attack from presentation of symptoms. It attempts to estimate the location, height, and time of an airborne anthrax release, as well as the number of spores. The observables consist of respiratory visits to emergency departments, as might be

obtainable from syndromic surveillance systems such as RODS [47]. The model that relates these observables to outbreak characteristics includes a Gaussian dispersion plume [48], Glassman's infection relation [43], and a log-normal distribution of incubation periods, with dose-dependent mean and standard deviation. However, BARD's use in an urban context is only approximate since Gaussian plumes are suited mainly for open spaces [48].

In this study, we develop a Bayesian formulation for inferring BT attack characteristics in the form of probability distributions for N, τ, and D, using data from the first 3–5 days of an outbreak. We restrict ourselves to temporal analysis; that is, we do not take the location of diagnosed patients into consideration. All tests are performed with anthrax as the pathogen. In this study, a hypothetical infected population receives a broad range of doses, commensurate with atmospheric dispersion over a 10 km square domain. We explore how the accuracy and uncertainty of inference are affected by the size of the outbreak, the dose received, and the frequency with which patient data is collected. In the interest of realism, we also consider cases in which the anthrax model used to generate the observed data (via simulated outbreaks) is different from the model used in inference. We conclude with an application of this method to the Sverdlovsk outbreak of 1979 [7].

This study adds a new degree of realism to outbreak data and its analysis compared to those conducted in [34, 46]. Unlike [34], we consider dose-dependent incubation periods and populations infected by a range of doses, as might be obtained by atmospheric dispersion, and infer a representative dose for the population. Since aerosol releases in confined spaces can lead to high doses (comparable to or greater than ID_{50}), the inferred dose serves as a useful indicator of the indoor versus outdoor nature of the release. Unlike [46], model uncertainty—when the disease model used in the inference procedure is only a partially accurate representation of the disease's behavior—is considered here to explore how large an inference error one might encounter under realistic conditions. Further, since our analysis is strictly temporal, we do not take into account the geographical location of patients; in a mobile population, this can be a significant source of (observation) error, especially if the time of infection is not known and a detailed movement schedule of the infected patients is unavailable. We also consider correlations between the inferred parameters of the attack, demonstrating realistic cases in which scarce data might support multiple characterizations. These were not explored in [34, 46].

2.3 Characterization of smallpox outbreaks

Smallpox is a highly contagious and frequently fatal disease. Its causative agent is an Orthopox virus *Variola major*. The best sources of epidemiological information on smallpox are [11, 49]. The disease has 12 manifestations, ranging from the uniformly fatal hemorrhagic and flat manifestation to the relatively mild "modified" manifestation [49]; overall the mortality rate is approximately 30%. The disease follows a typical incubation-prodromal-contagious-removed sequence; removal by recovery bestows immunity. The distributions for the incubation, prodromal and contagious periods can be found in [2], where they were modeled as Γ distributions. The R_0 for smallpox, a measure of the spreading rate in a virgin population, has been estimated to vary between 3 and 17 [9]; the upper limit of spread was observed in a hospital in Meschede, W. Germany, in 1970 [10], where the contagion "leaked out" from the isolation ward with warm air into an in-

sulated hospital (it was winter).

The threat from a release of smallpox would be of a strategic nature; apart from its high mortality rate, it spreads rather quickly. Smallpox has been used in warfare in the past (infected blankets were distributed to the Indians during the French and Indian Wars, 1754-1767 by the British forces in North America [4]) and it has been alleged that the Soviet Union weaponized it [6]. However, its ability to spread in a contemporary society is unknown, though attempts have been made to model it [25, 50]. These models attempt to separate the effects of social mixing from pathogen characteristics, but are frequently forced to use parameters whose values are largely guessed. Thus being able to extract the pathogenic transmissibility from observations of historical epidemics, separate from the effect of social mixing on disease spread, can be of help in informing such models. Also, given this degree of uncertainty in crucial pathogenic and epidemiological parameters, a real-time approach to measure the instantaneous spreading rate of such an outbreak would be helpful, if only to measure the efficacy of epidemiological countermeasures. Thus, methods to characterize outbreaks of contagious diseases, from full and partial observations, can be useful.

Such efforts have already started, mostly for emerging infectious diseases. Recent studies [51, 52, 53, 54] have concentrated on estimating the spread rates (R_0) of various emerging strains of influenza, from sparse observations; however, they have generally used conventional, ordinary differential equation-based SEIR models. Further, these studies did not include any effect of a structured population. Of late, there has been some interest in addressing problems of statistical inference, predicated on incomplete data, which involve stochastic epidemics in a structured population. Typically, the structure involves clustering, most commonly, a family or a household. The in-household rate of spread is assumed to be larger than the rate at which households themselves get infected. Cauchemez *et al.* [55] performed such a study on the spread of influenza, as observed over a period of 15 days, in Epigrippe, France. Their recent work, however, structures the population into adults and children and infers the importance of children as vectors for the diseases, conditioned on Sentinel data [56]. Eichner and Dietz [2] divided the population in the Abakaliki into 3 groups and estimated inter- and intra-group spread rates of smallpox. In both these studies, homogeneous mixing was assumed inside each group i.e. there was no notion of a social graph.

Introduction of an unobserved social graph into an inference problem renders it high-dimensional (since the social graph itself becomes a model "parameter" to be inferred) and has generally been addressed using Markov Chain Monte Carlo (MCMC). MCMC has been used to infer epidemiological models, even when social graphs were not involved [57, 58, 59]. Britton and O'Neill [60] investigated gastroenteritis and shigellosis outbreaks where they explicitly introduced a social graph into a stochastic epidemic model. They assumed an SIR model and formulated a Bayesian inverse problem for the dates of infection and the average contagious period of the disease (assumed exponentially distributed). A closed population was assumed, and a binomial graph, with an uncertain connection probability, used to model interpersonal relations. Disease transmission over a social link was modeled as a Poisson process, whose rate was inferred as a part of the solution. The authors formulated a Bayesian inverse problem, predicated on the removal dates of the epidemic victims, and solved it using an MCMC procedure. A mixture of Gibbs and Metropolis-Hastings updates were used to sample the high-dimensional parameter field, which include the social graph and the infection pathway. The size of the problem was generally small (10-40 patients in a population of roughly 100-200). Demiris and O'Neill [61] extended Britton and O'Neill's approach

to address two-level mixing, i.e., where the social graphs for inter-household and intra-household connections assumed different contact probabilities. However, they retained the SIR model, assumed that the contagious period was known, and modeled the social graphs as binomial graphs.

Modeling social connections with a binomial graph is rather restrictive; studies have shown that human contacts rarely follow such a distribution [62]. Britton's recent work has addressed the generation of random graphs that follow a given degree distribution [63], but they have not yet been incorporated into an epidemic inference problem.

2.4 Markov Chain Monte Carlo methods

We conclude our discussion of prior work with a quick review of Markov Chain Monte Carlo (MCMC) methods which we use to solve our Bayesian inverse problem. Inverse problems are most profitably formulated in a Bayesian framework if (1) the data are diverse and (2) the data are sparse. Diverse data, which may not be linked together via a model in an inverse problem, can be accommodated directly via *prior beliefs* in a Bayesian inverse problem. Bayesian methods allow estimation of inverse problem unknowns as probability density functions; this is critical when data are sparse and point estimates could be insufficient/misleading. [64] is an excellent reference on the formulation and mathematical aspects of such problems; [65] adopts a more practical approach to formulating Bayesian inverse problems. Such problems result in an expression for the joint posterior probability distribution for the problem unknowns (and thus can be high-dimensional). The joint probability distribution is evaluated by sampling from it, which is most efficiently done using MCMC methods. Metropolis-Hastings samplers, which can address arbitrary posterior distributions, will be used in this work; [66, 67, 68] provide an excellent and detailed treatment of the matters, including many practical issues (e.g. "convergence" of the MCMC chain to a stationary state in a finite number of steps, for which no theoretical metric exists).

MCMC methods are not without problems; they often have difficulty sampling from multimodal distributions. However, "mode-hopping" MCMC methods, which directly address this problem have been studied [69, 70, 71]. MCMC also have difficulty with narrow and skewed posterior distributions; these may be resolved by either transforming the unknowns yielding a better-behaved distribution (i.e., more circular) [66] or adapting the proposal distribution [72], particularly when dealing with high-dimensional problems [73]. An intuitive way to deal with higher dimensional problems (where a single chain may have difficulty visiting the entire space in a reasonable number of iterations) or problems where evaluating the posterior distribution is computationally expensive (where a single chain may not be able to take many steps in a reasonable amount of time) is to have multiple chains distributed among multiple CPUs in a parallel supercomputer; Population Monte Carlo methods to do so have been investigated [74, 75].

3 Individual-based models of epidemics, including approximations.

A detailed model for inter-person disease transmission is essential for solving inverse problems to characterize outbreak source, strength, and number of people infected. It also enables crucial scenario analysis including potential transmission paths, effect of interventions such as vaccination and quarantine, and advance disaster response placement decisions. Detailed individual person-based models can also inform construction of aggregate (mass-action) epidemic models, such as SEIR models, potentially with structured populations.

The inversion work in this project concerns source identification given extremely limited observations. Aggregate population models are insufficient in this regime, where the total number infected is low and potential transmission paths need to be analyzed to reveal likely index and secondary cases. To meet these needs, we constructed a stochastic, individual-based model for disease transmission, including social contact network structure and detailed in-host pathogenesis models. Key features of the initial model prototype include:

- object-oriented C++ implementation;
- text-based input file specifying simulation characteristics;
- accepts bipartite person/location or unipartite person/person contact graph;
- accepts static or dynamic (time-varying) contact graph;
- capable of simulating multiple diseases simultaneously;
- scales well via MPI parallelism for rapid individual scenario analysis, or can be run massively serial for ensemble Monte Carlo analysis;
- aggregates multiple events in time to coarser time scale; and
- can perform reduced-order simulations using user-provided person and/or location samples.

In this section we describe our individual-based disease transmission model, including social networks and associated transmission models, available social network data, potential network reduction approaches, and in-host pathogenesis models. We present sample simulation results which demonstrate bulk epidemic properties and spatial spread, including for reduced-order simulations.

3.1 Social contact networks and transmission models

While disease model stochasticity could derive in part from stochasticity of the social contact network or uncertainty of its specification, the network specification to our disease model is explicit deterministic input. The specification of a social network for disease spread may be static or

dynamic, and may be bipartite (e.g., with nodes representing people and locations) or unipartite, modeling only person-to-person interactions. While any of these combinations are feasible, those implemented in the present disease model are discussed here.

The available transportation-based data (described in Section 3.2) include dynamic, bipartite graph characterizations similar to those used by Eubank, et al. [25]. Graph nodes consist of people and locations, and dynamic edges connect people to locations, indicating their transient behavior throughout the day. In this specification, when people are collocated (same geographic coordinates and same sublocation, e.g., room within a building), there is potential for disease transmission.

The static person-to-person networks considered consist of nodes (people) and weighted edges indicating time of collocation. While weights could be specified as percent time collocated (relative to other people pairs), the data specify collocation in hours for a 24 hour period. The network structure is currently supplied to the disease model in augmented compressed sparse row (CSR) format, with a specified period (24 hours), and edge weights in hours of collocation. The static network-based disease simulation is therefore time stepped using the 24 hour period specified for the connectivity data. Static equivalents of the dynamic contact graphs are included in the available data sets.

We consider two models for inter-host disease transmission. The first is a physics-based model, where people (and locations when considering a bipartite representation) have a disease "load" associated with each pathogen they may acquire. [25] Contagious individuals shed pathogen, either to the location or directly to other people, at a potentially disease state- or covariate-dependent rate, influencing their load. Susceptible individuals absorb pathogen from the environment (at a potentially demographics-dependent rate), and update their load and in-host pathogenesis model accordingly, as described in Section 3.4. Load is not intended to strictly quantify the amount of disease present in an entity, but rather provides a means to model transmission and progress the relative state of disease in an individual. In this report, the load-based model is closely associated with the dynamic bipartite case, so individuals shed to and absorb from locations (as in the Eubank, et al. "environment-mediated transmission" model), but could be generalized to the direct transmission case.

In the load transmission model, initial conditions (pathogen levels) may be specified by contaminating locations, or directly infecting people. In contaminated locations and people who have not yet reached their infectious dose (ID), pathogen decays or grows exponentially at a fixed rate. This allows modeling of diseases with varying vectors, including contamination, and differing requirements for sustained growth or decay. A person becomes infected once they reach their assigned infectious dose.

The second transmission model is simpler, based on the probability of transmission between two collocated individuals. Based on the overall probability of transmission in a 24 hour period, we assume that the probability increases with time collocated, according to

$$p_j = 1 - \exp\left\{ -\sum_{i=1}^{N_N} p_i(t) w_{ij} \right\} \tag{1}$$

where p_j denotes the probability of susceptible person j being infected, given collocation with N_N neighbors over a 24-hour time period. The time-dependent $p_i(t)$ is the scaled probability of neighbor i infecting a susceptible person and is dependent on their disease state and therefore simulation time t. The weight w_{ij} indicates the time person i and j are collocated in a 24-hour period. While this could be implemented for both the static and dynamic graph cases, we presently only consider such transmission models for the former.

Initial disease conditions (outbreaks or attacks) may affect people or locations, and may be specified by person or location. For example, the simulation input may specify that location 101 is contaminated, that all people in location 101 are directly infected, that specific people individuals are infected, or the location (or its occupants) corresponding to a particular individual is attacked (as in the malicious case).

3.2 Available network data

Results presented in this report use social contact network data available from the Network Dynamics and Simulation Science Laboratory (NDSSL) of the Virginia Bioinformatics Institute at Virginia Tech. These synthetic data for the population of Portland, Oregon, are generated from detailed simulation models implemented in Simfrastructure, which aims to simulate functioning virtual cities at the individual level. [76, 77] Simfrastructure includes TRANSIMS for agent-based large-scale transit simulations [78] and EpiSims for disease outbreak modeling [25], and is designed to create a prototypical urban population and infrastructure and together with individuals' movements and interaction with the infrastructure. The most relevant features of these synthetic data include:

- approximately 1.6 million people, with demographic attributes, and assignment to households;
- 243,423 geographically distinct locations that people may visit, with (x, y) coordinate pairs in meters (roughly two locations per roadway link/city block);
- activity data, indicating movement of people among locations, and their purpose in occupying a location, for a typical day (these characterize a dynamic, bipartite social contact graph); and
- a static projection of the social contact network, based on the dynamic activities, that results in a person-to-person social contact graph with edges weighted according to time collocated.

The populations were generated to be statistically equivalent to Portland census data at the block level. Collocation of two individuals at a geographic location is insufficient for contact, as a location may include multiple sublocations. For example, there may be several households and/or work sublocations assigned to a single location.

The NDSSL released two versions of synthetic data for Portland. Initial model development used Data Set Version 1.0 [76], released in January 2006, but several limitations required moving to

Data Set Version 2.0 [77] when it was released in 2007. Potential limitations with the Version 1.0 data include:

- a total time span of approximately 30 hours, and no clear means to select a 24 hour slice for simulation purposes;
- many individuals would "disappear" for long periods of time (5–6 hours), i.e., were not assigned to locations for these time spans, would occupy a location for potentially unrealistic durations (order seconds), or only have movement data for a small fraction of the 30 hour simulation period;
- some overlapping people events, i.e., a person may be assigned to two distinct locations at the same time;
- lack of explicit sublocation data (as generated by the underlying EpiSims sublocation model);

For model prototyping and testing, most of these are not fatal limitations. However the lack of sublocation-resolution movement data prohibited comparison between the dynamic, bipartite and static, unipartite graph models. The latter static network instances were generated from social contact based on the sublocation model, which requires people to not only occupy the same geographic location, but also the same sublocation. The Version 2.0 data set explicitly included sublocation data for both person movement and household locations, enabling comparison to the static graph case.

Other minor differences between the Version 1.0 and 2.0 data sets include addition of boolean worker information, relationship of people to household head, and number of household vehicles. Also, there are minor differences in numbering/indexing of entities in the data. The disease simulation can accommodated pre-processed data from either original data format.

A potential benefit of the Version 2.0 data is inclusion of several realizations of disease spread simulated with the Epi-Fast model, including various initial conditions and simulated public health interventions. These could be used to to compare disease transmission assumptions, with a common underlying social structure. For further details on the construction and characteristics of the synthetic data used, the interested reader should consult the technical reports from NDSSL references above and references therein.

3.3 Description of population sampling approaches

The static unipartite network is one example of a surrogate for (approximation of) the full dynamic, bipartite graph. Strategic population and/or location sampling offers another type of surrogate or reduced-order model for the full fidelity simulation. One model reduction technique involves clustering or partitioning people, based on a similarity metric, and then selecting one or more (perhaps a percentage of) people from each cluster. The sampled representatives can then be used in a disease simulation as a surrogate for the full population.

For illustration purposes, consider a simple matrix representation, with one row per location, one person per column, and non-zero entries C_{ij} that indicate person j visited location i. Instead of indicators the entries could be weighted by time spent in that location over a 24-hour period. The inner product of two columns C_{*j} and C_{*k} gives a measure of similarity between persons j and k in terms of locations they visited that could then be used for clustering.

We consider two specific clustering variants; the first based on person-to-location clustering and the second on person-to-person clustering.

1. **Person-to-location clustering with Zoltan:** This approach clusters people who visit the same locations, but does not represent collocation of people with respect to time.

   In this variant, hypergraph coarsening [79] from the Zoltan toolkit was used to cluster the people in the simulation. Vertices of the hypergraph represented people; hyperedges represented locations. A hyperedge, then, consisted of all people who visited a given location during a 24-hour period. This hypergraph representation is more compact and retains more information than a person-to-person graph constructed by connecting (with graph edges) all individuals who visited a location. In our coarsening algorithm, we computed the number of locations shared by pairs of people. People who shared the most locations were considered to be most similar and, thus, were grouped together in a greedy manner. The coarsening process was terminated when a coarse hypergraph of user-specified size was obtained.

   This process was used to generate people samples of sizes 107359 and 6874 using Data Set Version 1.0. Adaptations of this process accounting for percentage of time collocated through summary weights on hyperedges are possible, but not considered here.

2. **Person-to-person clustering with PVXORD:** This approach clusters people who are often collocated at the same locations, or collocated for similar amounts of time.

   In contrast to the person-to-location approaches, where the data size is M locations by N people, this variant operates on larger $N \times N$ data, and requires examination of time-dependent collocation data to generate the connectivity matrix. Dynamic movement data for hour 6 only (software was limited by data size), were used to construct a person-to-person contact graph with edges weighted by percentage of simulation time individuals are truly collocated (normalized to 1.0). We used the parallel VXORD software PVXORD (Wylie, Martin, Brown) to assign people to clusters and then sample one representative person (or again a percentage of people) from each cluster. By this means a population sample of 238370 individuals was generated.

A challenge in performing this kind of network model reduction is properly scaling contact rates and disease properties so the reduced-order graphs are faithful to the full-fidelity simulation; see comparative results in Section 3.6.

Table 1. Representative smallpox disease periods, with probability distributions for durations.

period	distribution	log (load threshold)	notes
exposure	n/a		
incubation	normal(12.0, 2.0)	0.5	
prodromal	constant(3.0)	2.0	symptomatic
contagious burst	constant(0.5)	6.0	symptomatic, infectious
contagious decline	normal(6.5,1.0)	7.0	symptomatic, infectious
recovery	uniform(6.0,12.0)	6.0	

3.4 In-host pathogenesis (micro-models)

In-host disease progression models are based on stages and at a minimum, include those corresponding to an SIR model: susceptible, infectious, and recovered. The residence times in each disease period are sampled from probability distributions, introducing a stochastic element. Associated with each phase are indicators for whether a person is susceptible, contagious, and/or symptomatic for that phase. This could readily be extended to include severity of susceptibility or contagiousness dependent on a person's disease state or even covariate (e.g., demographic) information.

For example, for smallpox, we consider the periods with distributions specified in Table 1, similar to the original EpiSims smallpox model. Here the normal distribution is parametrized by mean and standard deviation and the uniform by its lower and upper bounds. The exposure period is of variable length, as a person remains in it until reaching their assigned infectious dose. At the end of the recovery period, a person could have acquired immunity, otherwise is dead. There is wide speculation about the pathogenesis of smallpox in modern society, and slightly different periods together with discrete probability distributions are reported by Longini, et al. [28], with reference to consensus of a recent Smallpox Modeling Working Group working group commissioned by the United States Department of Health and Human Services.

For load-based disease models, as in [80], we assume that the human ID_{50} (infectious dose, 50% probability) is 5 PFUs (plaque forming units), and that 500 PFUs provide an almost 100% probability of infection. Infectious doses for individuals in the simulation are therefore sampled from a log-normal probability distribution with mean 18 and standard deviation 68.

While disease stage lengths vary across individuals, the associated disease loads (with the exception of that required to enter incubation) are the same for all people. During a stage, the disease load grows exponentially with (potentially negative) rate required to attain the specified starting and ending load over the sampled period length. The in-host load model can therefore be represented as an ordinary differential equation

$$\frac{dL(t)}{dt} = a(t)L(t) \tag{2}$$

with piecewise constant growth rate $a(t)$. While the generality of a differential equation is not necessary for the present simulation, which uses piecewise constant growth rates over each disease stage and therefore has an explicit analytical solution, it permits the replacement of the in-host models with more complex pathogenesis models if desired.

For the results here, we consider fixed shedding and absorption rates $1.0e-4$ and $5.0e-3$, respectively. The model also permits specifying these in a demographics- or disease state-dependent fashion. For example, data might support age-dependent transmission rates, or specification that infected individuals are more contagious in specified time windows or given certain variants of a disease.

3.5 Parallel implementation

An initial implementation of load-based time stepping with the dynamic network proved computationally intensive (requiring over 26 CPU hours on a single 3.60 GHz Intel Xeon processor to simulate three weeks of disease spread). The first parallel implementation replicated locations on all processors and partitioned people among processors. An MPI all-to-all communication pattern updated loads, where contributions to loads (load updates) were computed on all processors and then all-reduced. This did not scale well with the number of processors.

The current model incorporates a point-to-point communication model, which leverages services from Sandia's Zoltan framework. [81] Along with the new parallel communication model, we partitioned both locations and people among processors, reducing memory requirements per processor. With point-to-point communication, each processor registers needed load updates (either local people needing off-processor location data or local locations needing off-processor people data), and the Zoltan communicator orchestrates the parallel data movement when requested. For static networks a single communication initialization suffices, whereas for dynamic networks the communication pattern must typically be re-initialized at each time step.

Even with this additional overhead, the new parallel model scales nearly linearly up to 64 processors, as shown in Figure 1. Formal profiling studies have not yet been conducted, but we conjecture that for this (relatively small) social network, communication begins to dominate local processing at around 64 processors. If scalability tapered for larger data sets, load balancing via Zoltan's hypergraph partitioning would likely help.

3.6 Sample simulation results

The section includes sample disease model results for smallpox. A preliminary plague model, implemented, but not yet exercised.

In Figure 2 we present an example of the spatial spread of disease in Portland, Oregon, over 21 days. The initial condition (I1) exposed 1599 collocated people to 100 PFUs of smallpox each, 312 collocated people to 30 PFUs each, and contaminated another location with 1000 PFUs. Based

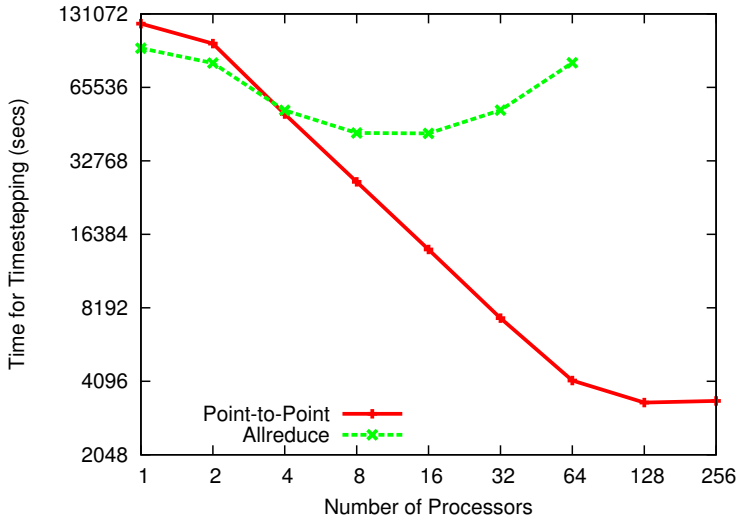

Figure 1. Parallelism: improved scalability with point-to-point communication model.

on the dynamic load-based model, the intensity represents the disease load at the location. The geographic spread is not simply diffuse, but indicates social connections are responsbile for first bringing disease to a new region. Ongoing work is exercising the model with carefully selected initial locations to assess spread across key city features and natural partitions.

Comparison between the dynamic and static network approaches is shown in Figures 3 and 4, with comparative SEIR data for one and six months, respectively. These results correspond to an initial condition affecting 7 locations, with varying amounts of pathogen in the load-based case and a 95% probability of initial transmission in the direct transmission case. While there are some temporally locally deviations, the simulations agree qualitatively. The plots indicate more secondary infections in the load-based case. The parameters for the dynamic, load-based simulation are adapted from [25] while those used in the static, simple contact model are adapted from [28]. Given the independence of the parameters and differences in simulation formulations, the agreement is surprisingly good. If calibrated, this discrepancy could likely be reduced.

We present examples of model reduction for both the person-to-location (1) and person-to-person (2) sampling cases, using Data Set Version 1.0, initial condition I1. Figure 5 depicts overall epidemic spread with the full population and with the 107,359 person sample. The initial outbreak wave is captured well by the reduced-order model, but the wave of secondary infections is considerably stronger in the full population. The geographic spread of the disease in the reduced-order case is shown in Figure 6. Qualitatively, the epidemic characteristics are similar, but clearly reduced-order population models must incorporate scaled rates in order to properly represent the full population. Our network sampling techniques are heuristic and could be improved by adding analytic rigor in the sampling process to preserve particular graph and/or simulation characteristics in the sample.

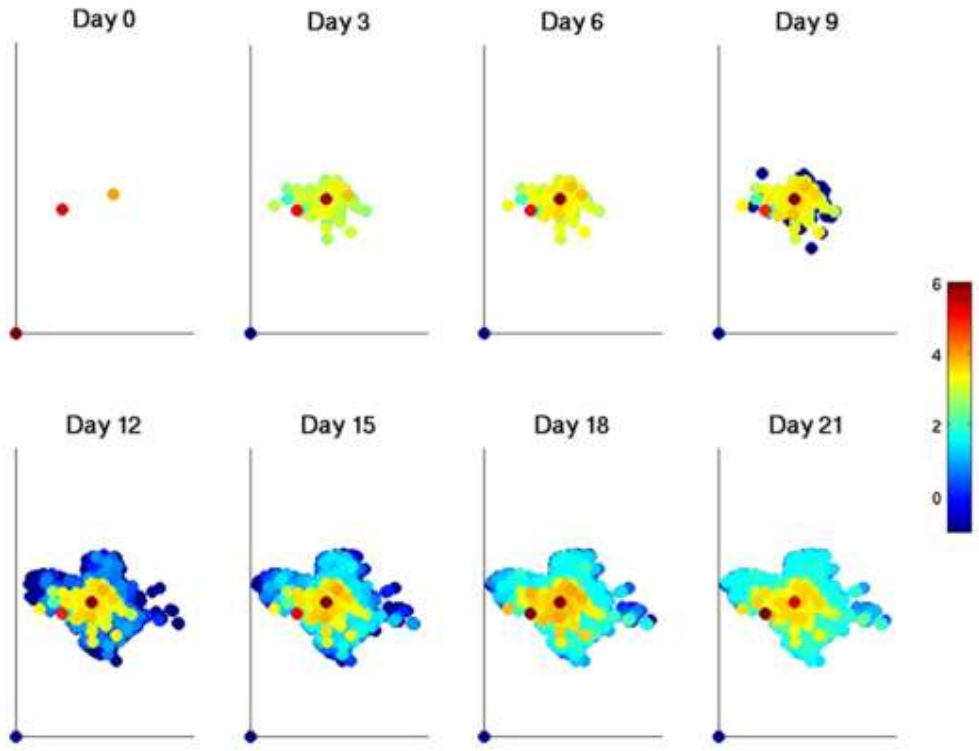

Figure 2. Representative spatial spread of smallpox in an urban setting over 21 days, using full-fidelity dynamic network, Data Set Version 1.0.

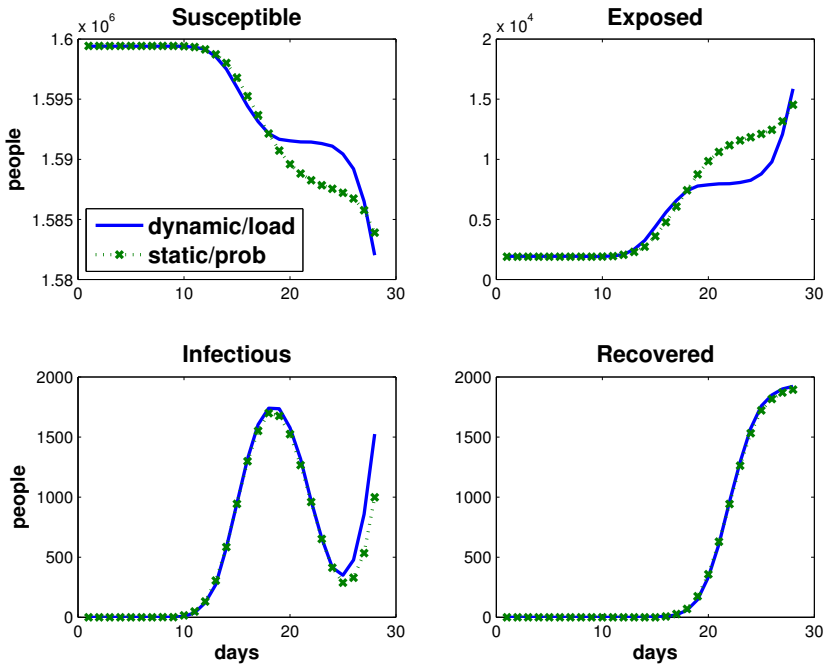

Figure 3. Comparison of summmary SEIR data for dynamic and static network approaches, one month horizon, Data Set Version 2.0.

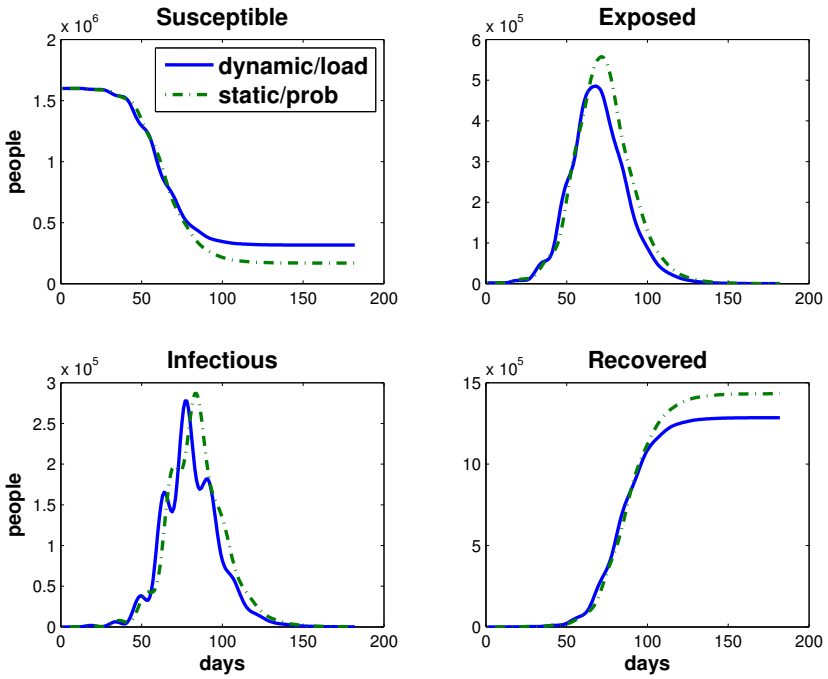

Figure 4. Comparison of summmary SEIR data for dynamic and static network approaches, six month horizon, Data Set Version 2.0.

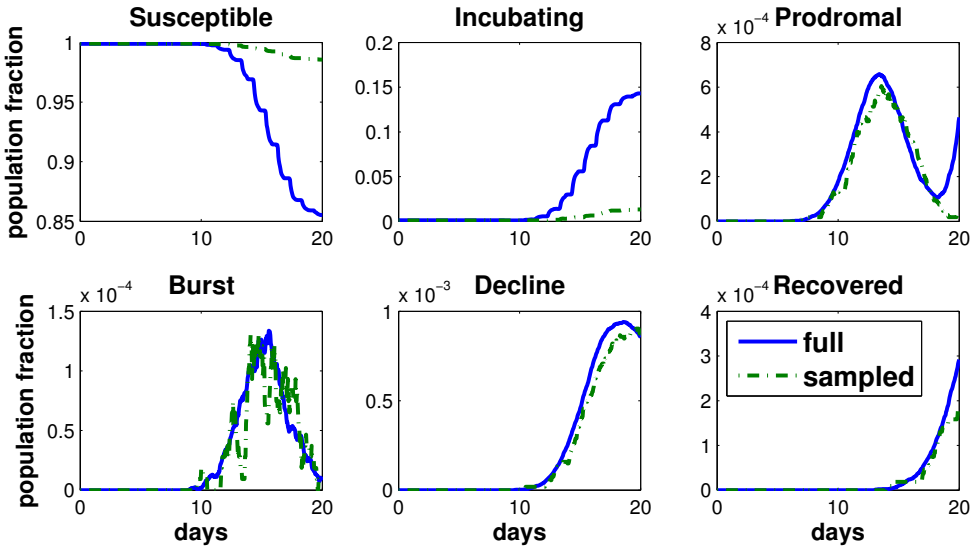

Figure 5. Smallpox epidemic: Population fraction in each disease stage, comparing full-fidelity dynamic network to reduced-order network obtained via person-to-location clustering, Data Set Version 1.0.

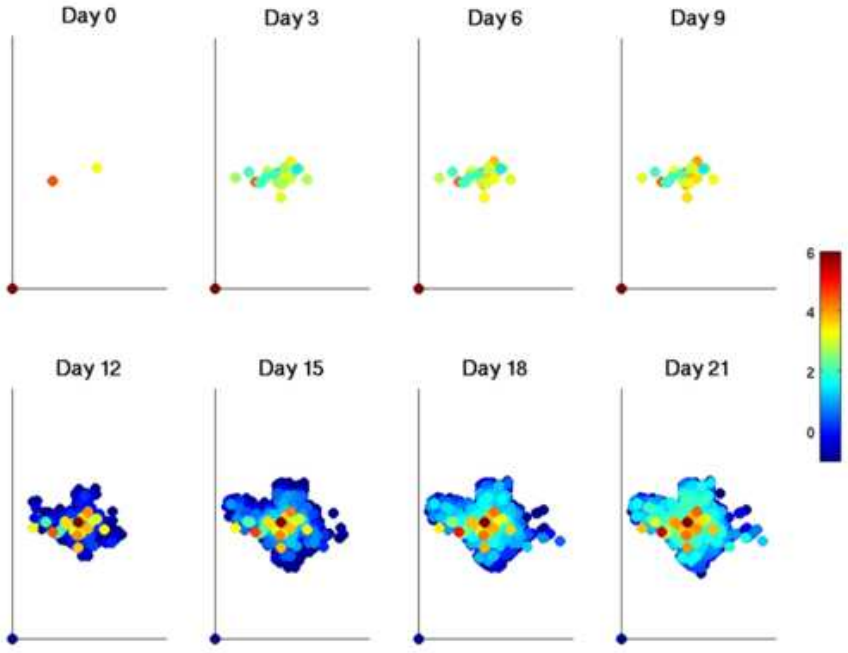

Figure 6. Representative spatial spread of smallpox in an urban setting over 21 days, using reduced-order dynamic network based on person-to-person clustering.

4 Characterizing outbreaks from partial observations: The case for non-contagious diseases

Non-contagious diseases (i.e. diseases that are not spread from human to human) usually refer to vector-borne diseases or zoonotics. Such diseases have been largely controlled by standard public-health policies. However, a few of them, most famously, anthrax, have been weaponised - i.e. turned into a form, by industrial means, into an aerosol that can be dispersed easily and efficiently (see [82, 4] for a review of various state-sponsored programs and their vicissitudes). Outbreaks caused by their release are not expected to bear any resemblance to commonly observed zoonotic outbreaks. In this section we identify what such outbreaks of non-contagious disease may look like, what the relevant issues are and mostly, how such issues may be resolved. For the purposes of our study, we will use anthrax as the pathogen.

The release of weaponised anthrax (similar to the Amerithrax attacks [3]) will probably occur in a bioterrorism or biological warfare setting. The aim will be to infect a large number of people, N, who then would have to be cared for, thus imposing a severe cost in terms of resources. From the victim's point of view, estimating N becomes critical, for the allocation of resources. Since the quantity of pathogen released, the time and location of release may not be known, a straightforward estimation (using models of aerosol dispersion) is impossible. N, therefore, may have to be estimated by indirect means.

N individuals infected with a spectrum of doses of anthrax will come out of incubation over a period of time, leading to a time-series, $n_i, i = 1 \ldots M$ of diagnosed people over M days. The time-series will provide definite proof of a biological attack and may allow one to infer N (and secondarily, the time of release τ and a representative dose D). Further, we will adopt a purely temporal approach which will not require the location of the diagnosed people at the time of infection.

In this section, we develop a Bayesian formulation for inferring BT attack characteristics in the form of probability distributions for N, τ, and D, using data from the first 3–5 days of an outbreak. All tests are performed with anthrax as the pathogen. Compared to [34] and [46], we introduce a new degree of realism to outbreak data and its analysis. Unlike [34], we consider dose-dependent incubation periods and populations infected by a broad range of doses, commensurate with atmospheric dispersion, and infer a representative dose for the population. Since aerosol releases in confined spaces can lead to high doses (comparable to or greater than ID_{50}), the inferred dose serves as a useful indicator of the indoor versus outdoor nature of the release. Unlike [46], model uncertainty—situations in which the disease model used in the inference procedure is not an accurate representation of the disease's behavior—is examined here in order to assess how large an error one might encounter under realistic conditions. We also explore how the accuracy and uncertainty of estimates are affected by the size of the outbreak, the dose received, and the frequency with which patient data is collected. Further, we identify correlations between the inferred parameters of the attack, demonstrating realistic cases in which scarce data might support multiple characterizations. These were not explored in [34, 46].

Since our analysis is strictly temporal, we do not rely on the geographical location of patients;

in a mobile population, locations can be a significant source of error, especially if the time of infection is not known and a detailed movement schedule of the infected patients is unavailable. For example, the night time locations of the Sverdlovsk victims shows no trend, while the day time locations lie approximately on a straight line bearing south-east of a military facility (see [82] for a discussion; the plots of the day time and night time populations are in [83]). We conclude with an application of this method to the Sverdlovsk outbreak of 1979 [7].

4.1 Formulation

We now formulate a Bayesian parameter estimation problem for the characteristics of a BT attack. Consider a time series of infected patients $\{t_i, n_i\}$, $i = 0 \ldots M$, where n_i is the number of people developing symptoms in the time interval $(t_{i-1}, t_i]$. For simplicity, we let the intervals be of uniform length $\Delta t = t_i - t_{i-1}$, typically 6 or 24 hours. t_0 marks the time at which the first symptomatic patient is identified; this patient may have developed symptoms anytime between t_0 and $t_{-1} = t_0 - \Delta t$. M is the total length of the time series and is expected to be small, e.g., 3–5 days. We seek a probabilistic model for these observables, conditioned on an attack that infects N people at time τ with a uniform dose of D spores. By convention, we set t_0 to zero, and thus τ, the time of infection, is always negative.

The dose-dependent incubation period is described by its cumulative distribution function (CDF) $C(T,D)$, where T, the incubation period, is the time elapsed since infection. The probability of an infected individual developing symptoms in the interval $(t_{i-1}, t_i]$ is thus $\{C(t_i - \tau, D) - C(t_{i-1} - \tau, D)\}$. Let $L = \sum_{i=0}^{M} n_i$ be the total number of people who have developed symptoms by the end of the observation period t_M. Then $N - L$ infected people are still asymptomatic; the probability of someone remaining asymptomatic at t_M is the survival probability, $P_{\text{surv}}(t_M - \tau, D) = 1 - C(t_M - \tau, D)$. Since the incubation times of each individual are conditionally independent given N, τ, D and the disease model, the probability of the entire time series $\{t_i, n_i\}$ obeys a multinomial distribution with $M + 2$ outcomes. One outcome corresponds to remaining asymptomatic at t_M; the $M + 1$ others correspond to developing symptoms in a preceding time interval. The resulting conditional probability distribution is given by the following expression:

$$
\begin{aligned}
& P\left(\{t_i, n_i\}_{i=0}^{M} | N, \tau, D\right) \\
&= \frac{N!}{(N-L)! \prod_{i=0}^{M} n_i!} \times \{P_{\text{surv}}(t_M - \tau, D)\}^{N-L} \\
&\quad \times \prod_{i=0}^{M} \left(C(t_i - \tau, D) - C(t_{i-1} - \tau, D)\right)^{n_i} \\
&\equiv \mathcal{L}(N, \tau, D)
\end{aligned}
\tag{3}
$$

In last line of this equation, we rewrite the probability of the observables as a likelihood function $\mathcal{L}(N, \tau, D)$. We then use Bayes rule to obtain the posterior probability of the attack parameters:

$$
p\left(N, \tau, \log_{10}(D) | \{t_i, n_i\}_{i=0}^{M}\right) \propto \mathcal{L}(N, \tau, D) \pi_N(N) \pi_\tau(\tau) \pi_D(\log_{10}(D)) \tag{4}
$$

Note that we have written the posterior density in terms of $\log_{10}(D)$ rather than D; this is in keeping with [43, 42], where response to infection is generally modeled as a function of the log-dose. Here π_N, π_τ, and π_D are prior densities on N, τ, and $\log_{10}(D)$. Presuming a lack of additional information, we use broad uniform priors on all three parameters. The joint posterior density can then be marginalized to obtain individual probability density functions (PDFs) for N, τ and $\log_{10}(D)$. Integrals yielding these marginal densities are evaluated using the VEGAS algorithm [84], an iterative adaptive Monte Carlo method implemented in the GNU Scientific Library [85].

4.2 Anthrax incubation models

This section briefly reviews mathematical models used to predict the onset of symptoms in a person exposed instantaneously to D anthrax spores. The onset time is a random variable, described by its cumulative distribution function (CDF). These models are from Wilkening [42]; details of their derivation can be found in [42, 86].

The CDF for Wilkening's Model D is given by [42, 86]

$$C_{\text{ModelD}}(T,D) = \int_0^T F(T-s;D,\lambda,\theta)g(s)ds, \tag{5}$$

which is a convolution of $F(T;D)$—the probability that at least one spore out of a dose of D spores will germinate into a vegetative anthrax cell by time t—and $g(s)$, which is the PDF of the time s taken, post-germination, to reach a bacterial load at which symptoms appear. F and g are defined as

$$\begin{aligned} F(T;D,\lambda,\theta) &= \frac{1}{p}\left(1-\exp\left(-\frac{D\lambda}{\lambda+\theta}Q(T)\right)\right), \quad \text{where} \\ Q(T) &= 1-\exp\left(-(\lambda+\theta)T\right), \end{aligned} \tag{6}$$

$$p = 1-\exp\left(-\frac{D\lambda}{\theta+\lambda}\right) \tag{7}$$

and

$$g(s) = \frac{1}{\sqrt{2\pi}\sigma_s s}\exp\left(-\frac{1}{2}\frac{\log^2(s/M_s)}{\sigma_s^2}\right). \tag{8}$$

p, the probability of showing symptoms in infinite time, is also called the attack rate. These distributions depend on a number of parameters:

- N_{thresh}, a threshold bacterial load in a person that causes symptoms

- t_2, the bacterial load doubling time in a given medium (e.g., mediastinal lymph nodes where the spores germinate), which can be obtained from *in vitro* laboratory experiments
- t_M, which is the time required to reach a bacterial load of N_{thresh}

$$t_M = t_{\text{lag}} + \frac{t_2}{\log(2)} \frac{N_{\text{thresh}}}{D}$$

- t_{lag}, a lag time in bacterial growth experiments (typically 1 hour)
- σ_s^2, the variance of the log of the time required to reach the symptomatic bacterial load
- θ, the probability rate of clearance of a spore (by the immune system), specified in terms of probability of clearance per spore per day
- λ, the probability rate of germination of a spore, specified in terms of probability of germination per spore per day

In the present models, M_s, the median time to symptoms, is set to t_M. The values of the parameters for Model D are $\theta = 0.109$ day^{-1}, $\lambda = 8.79 \times 10^{-6}$ day^{-1}, $t_{\text{lag}} = 1$ hour, $t_2 = 2.07$ hour, $N_{\text{thresh}} = 10^9$ and $\sigma_s = 0.544$ day^{-1}.

Sartwell [87] found that the incubation period for a number of diseases was log-normally distributed, which is at odds with Eq. 5. Wilkening's Model A2 captures this alternative by assuming a log-normal distribution,

$$C_{\text{ModelA2}}(T,D) = \frac{1}{2}\left[1 + \text{erf}\left(\frac{\ln(T/T_0)}{\sqrt{2}S}\right)\right], \quad S = 0.804 - 0.079\log_{10}(D) \quad . \tag{9}$$

where T_0, the median incubation time, is obtained by solving an integral equation derived from Eq. 5

$$0.5 = \int_0^{T_0} F(T_0 - s; D, \lambda, \theta) g(s) ds.$$

However, in solving for t_0, Wilkening used a slightly different set of parameters: $\theta = 0.11$ day^{-1}, $\lambda = 8.84 \times 10^{-6}$ day^{-1}, $t_{\text{lag}} = 1$ hour, $t_2 = 2.06$ hour, and $\sigma_s = 0.542$ day^{-1}. The reason for the slight change in parameters as well as the difference between Models A2 and D is discussed below.

Parameters in Eqs. 5 and 9 were obtained by fitting the models to the median incubation periods observed in experiments with non-human primates (performed by Henderson *et al.* [37] and Friedlander *et al.* [32]) and to the data from the Sverdlovsk anthrax outbreak. The average dose in the Sverdlovsk outbreak, however, had to be inferred from atmospheric dispersion models and the probability of exhibiting symptoms (in infinite time) given a dose of D spores. This is the procedure adopted by Wilkening [42]. Using Glassman's model [43] for the probability of infection, one obtains an average dose of 2.4 spores. Alternatively, if one employs Eq. 7 (which is similar in form

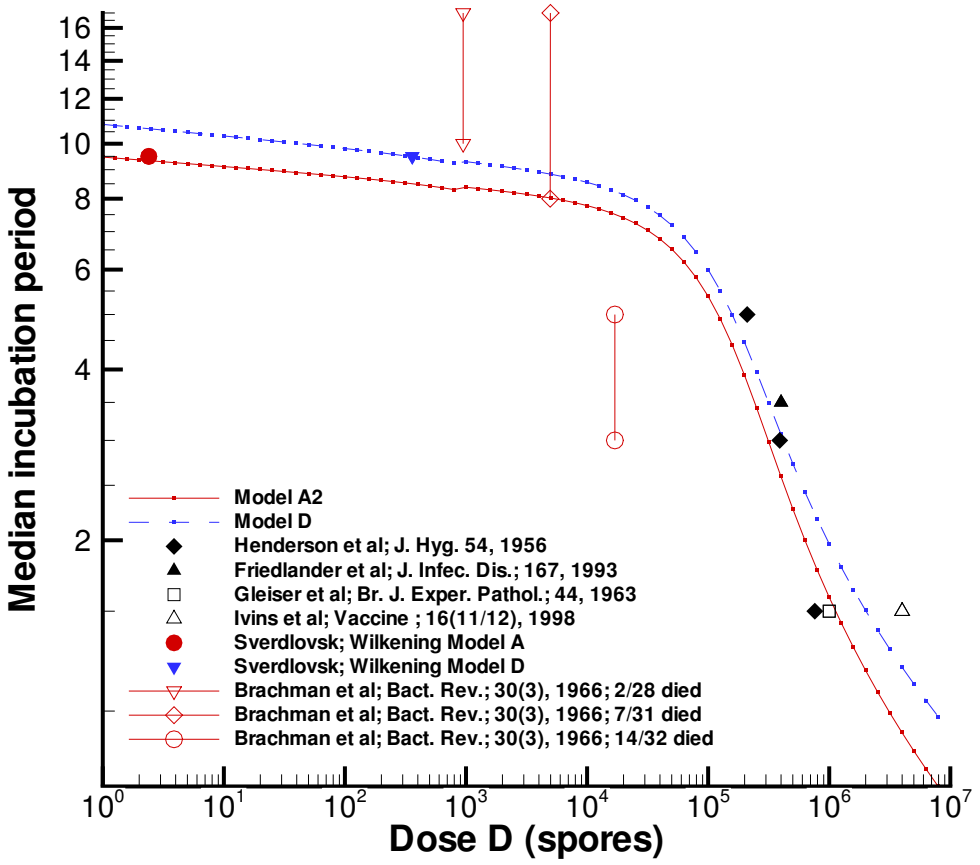

Figure 7. The median incubation period for anthrax as a function of dose D. The solid line is Model A2, which assumes a dose of 2.4 spores at Sverdlovsk; the dashed line is Model D, which assumes 300 spores. The solid symbols are median incubation periods obtained from experimental investigations or from Sverdlovsk data. The filled circle (Sverdlovsk; Wilkening Model A) refers to both Models A1 and A2, though only Model A2 is used in the current study. Symbols which are not filled denote experiments where the population of primates was too small to draw statistically meaningful results. The experiments by Brachman *et al.* [1] are shown by vertical lines between symbols. In these experiments, only the lower and upper bounds of the incubation period were provided. These ranges were not used for determining model parameters and are only provided for reference.

to Druett's [44] and was used by Brookmeyer in [41]) one obtains a dose of 300 spores. Wilkening retained both possibilities and incorporated them into separate models. Model D is based on a dose of 300 spores at Sverdlovsk while A2 assumes 2.4 spores.

In Figure 7, we plot the median incubation period predicted by Models A2 and D as a function of dosage D. The dosage at Sverdlovsk, estimated as 2.4 spores (represented by $\bullet$) is used to calculate parameters for Model A2 (solid line); the alternative estimate of 300 spores (represented by a filled ∇) is used for Model D (dashed line). Studies by Henderson [37] with 2.1×10^5, 3.9×10^5 and 7.6×10^5 spores (represented as filled $\diamond$) and Friedlander with 3.5×10^5 spores (represented by filled $\triangle$) were also used to calculate the parameters of both models. Studies by Ivins *et al.* [40] (unfilled $\triangle$) and Gleiser *et al.* [39] (unfilled $\square$) were conducted with very few primates and consequently are plotted only for reference. Primate experiments by Brachman [1] simulated the effect of prolonged regular exposure to low doses, as might be the case in a contaminated wool-sorting mill. The primates experienced extended periods during which they received no spores at all. The dose was defined as the total number of spores inhaled and was generally low, between 1000 and 10,000 spores. We plot the resulting ranges of incubation periods observed at various dosages, also for reference.

We see that the tests by Gleiser *et al.* and Ivins *et al.* agree with both models, which in turn agree with each other. However, significant differences arise when $D \lesssim 10^3$ spores. (Note that the vertical axis is logarithmic.) Brachman's tests show median incubation periods which are at odds with the models' predictions; however the mode of infection (a continuous low-level infection process spread over days or months) was very different from the rapid (timescale of an hour) challenge one would expect in a BT attack. Both models show a "kink" at $D \approx 10^3$; this is because they are evaluated with a lower value of λ (1.3×10^{-6} day^{-1}), corresponding to a primate ID_{50} of 55,000 spores, for comparison with primate results at the high dose limit, while the low dose predictions were developed with a human ID_{50} of 8600 spores for comparison with Sverdlovsk data. To the best of the authors' knowledge, this is the sum total of experimental data obtained from anthrax challenges of non-human primates where incubation times were measured. We have omitted a study by Klein *et al.* [88] in which an incubation period increase was observed with increasing doses, because only one primate was subjected to each dose, making the behavior statistically unreliable.

4.3 Inference of attack parameters with ideal cases

In this section we test the Bayesian estimation procedure described above. We use Wilkening's Model A2, described in Section 4.2, to simulate inhalational anthrax outbreaks of different sizes. The same model is used for inference; that is, there are no systematic errors between the inference and data-generation models. Thus, posterior uncertainties may be ascribed to (1) incomplete observation of the outbreak, specifically finite time resolution Δt and a short time series, and (2) the probabilistic character of disease incubation. We investigate how the quality of the inference varies with the size of the outbreak and the dose received. We also investigate whether a higher-resolution time series spanning a given observation period performs significantly better than a lower-resolution one.

Table 2. Time series obtained from six different outbreaks, simulated with the parameters $\{N, \tau, D\}$ as noted at the bottom of the table. The table has been divided into 24-hour sections, where the n_i in each section are summed to produce the low-resolution time series (24-hour resolution) used to investigate the effect of temporal resolution. Time is measured in days and dose in spores.

time	Case A	Case B	Case C	Case D	Case E	Case F
0.00	1	1	1	1	2	1
0.25	0	2	2	7	13	7
0.50	0	1	1	12	18	24
0.75	1	1	1	39	39	29
1.00	2	2	2	50	38	60
1.25	0	3	3	77	64	96
1.50	1	2	3	77	84	153
1.75	2	1	1	98	116	164
2.00	1	1	2	126	130	193
2.25	1	1	2	162	137	223
2.50	2	3	3	146	141	258
2.75	3	1	4	148	160	302
3.00	2	1	3	149	190	299
3.25	1	3	3	163	175	312
3.50	1	1	2	181	182	304
3.75	1	1	3	162	201	335
4.00	2	1	2	165	200	373
4.25	1	5	5	177	238	340
4.50	1	4	4	169	202	327
4.75	3	2	2	217	216	332
5.00	1	1	1	167	217	350
5.25	1	3	4	182	237	321
5.50	1	1	5	163	207	316
N	100	100	100	10,000	10,000	10,000
τ	-0.75	-2.25	-2.25	-0.5	-1.0	-1.25
D	1	100	10,000	1	100	10,000

In Table 2, we list time series at 6-hour resolution: the number of patients showing symptoms collected over 6-hour intervals obtained from 6 simulated outbreaks, henceforth called Cases A–F. Each infected patient received an identical dose D. N indicates the number of people infected and τ is the time of attack, measured in days prior to the exhibition of symptoms in the first diagnosed patient.

We use the procedure outlined in Section 4.1 to develop posterior PDFs for N, τ, and $\log_{10}(D)$ in Cases A–F. Figs. 8, 9, and 10 plot the resulting marginal densities for $\{N, \tau, \log_{10}(D)\}$. These are conditioned on the 6-hour resolution time series listed in Table 2. In Table 3, we summarize the maximum a posteriori (MAP) estimates and 90% posterior credibility intervals (CIs) for N, τ, and $\log_{10}(D)$ obtained with 5 days of data. We see that the marginal MAP estimate for N (the value of N corresponding to the peak of $p\left(N|\{t_i, n_i\}_{i=0}^{M}\right)$) is generally close to the correct value even with 3 days of data; increasing the length of the observation period to 5 days usually sharpens the PDF, reflecting a reduction in uncertainty. This trend holds true for small attacks ($N = 10^2$) as well as for large ones ($N = 10^4$). An exception is Case F, which will be discussed below. The time of attack τ is also identified quite readily, except for the small-N low-dose Case A. Larger attacks (Cases D, E, and F) have narrower PDFs for τ compared to Cases A, B, and C. Higher values of n_i in Eq. 3 (which generally result from large N attacks) provide structure in $\mathcal{L}$ and allow a more accurate estimation of the attack. The dose D is the most difficult parameter to infer. PDFs for Cases A, B and C in Figs. 8 and 9 show that it is virtually impossible to estimate the dose for small ($N = 10^2$) attacks; appreciable posterior probability is spread over 5 orders of magnitude. Table 3 confirms that MAP estimates of the dose in these small attacks are incorrect. Larger attacks ($N = 10^4$) yield more informative PDFs for D. Note that the sensitivity of $C(T, D)$ to D is rather small for Model A2 (see the expression for S in Eq. 9), suggesting that dependence of the likelihood function on D will be weak unless n_i or M is large.

Cases D, E, and F (Figs. 9 and 10) demonstrate how early observations of an outbreak may support multiple hypotheses, and at times favor a "wrong" hypothesis over the correct one. For instance, Case D exhibits peaks in $p(N)$ at $N \approx 4 \times 10^3$ and $N \approx 10^4$. Peaks in the PDF of $\log_{10}(D)$ occur at 1 spore and between 10^4 and 10^5 spores. For this case, both marginal PDFs overwhelmingly favor a large N, low-dose attack, which is the correct characterization. A similar minor ambiguity is observed in Case E. Marginal PDFs in Case F (Figure 10) are much more strongly bimodal, however. In Figure 11 we plot the joint posterior density $p(N, \log_{10}(D))$ to examine correlations among these parameters; it clearly shows two distinct islands—one corresponding to a large-N low-dose attack, and the other corresponding to a small-N high-dose attack. Up to Day 5, the data favors the wrong hypothesis (a larger, low-dose attack) over the correct one. Note also that the large low-dose attack corresponds to larger (i.e., later) values of τ, as evidenced by the posterior density $p(\tau)$ for Days 3–5 (Figure 10, right column). With more data (Day 6 and 7), the correct values for $\{N, \tau, \log_{10}(D)\}$ are recovered, with peaks at $N \approx 10^4$, $\tau \approx -1.2$, and $\log_{10}(D) \approx 4$. However, such a long observation period would not be relevant for consequence planning purposes. We stress that a Bayesian analysis is free to identify competing hypotheses, and that the degree of belief assigned to each is determined by the data and the prior information. In a partially observed attack, the MAP estimate may be erroneous, especially if data is scarce. One possible remedy is the use of informative priors for N, τ, and/or $\log_{10}(D)$ instead of the broad uniform priors used here. Otherwise, natural ambiguities may remain and should be accounted for in consequence management plans based on these inferences.

Table 3. Cases A–F; MAP estimates and 90% credibility intervals (in parentheses) for N, τ, and $\log_{10}(D)$, conditioned on the high-resolution time series at Day 5. The number in the curly brackets $\{\}$ is the correct value.

Case	N	τ	$\log_{10}(D)$
A	70, (39.45 – 123.3) $\{100\}$	-1.75, (-2.90 – -1.04) $\{-0.75\}$	0.0, (0.18 – 4.12) $\{0\}$
B	110, (65.7 – 148.4) $\{100\}$	-2.0, (-3.1 – -1.33) $\{-2.25\}$	0.00, (0.14 – 3.97) $\{2\}$
C	150, (88.78 – 194.7) $\{100\}$	-1.75, (-2.85 – -1.22) $\{-2.25\}$	0.0, (0.153 – 4.13) $\{4\}$
D	9800, (9439 – 10,350) $\{10,000\}$	-0.50, (-0.85 – -0.44) $\{-0.50\}$	0.00, (0.024 – 1.03) $\{0\}$
E	10,200, (8396 – 10,890) $\{10,000\}$	-0.9, (-1.41 – -0.67) $\{-1.00\}$	1.75, (0.87– 3.23) $\{2\}$
F	18,500, (10,500 – 19,290) $\{10,000\}$	-0.5, (-0.99 – -0.34) $\{-1.25\}$	0.75, (0.16 – 3.84) $\{4\}$

Coarser time resolution ($\Delta t = 24$ hours instead of 6 hours) was investigated in [86] and generally yielded only a mild degradation in the smoothness of the PDFs. In cases where a multimodal PDF evolves into a unimodal PDF over time (e.g., Cases D, E, and F), evolution is more rapid when the observations are collected in 6-hour intervals.

To summarize, solution of the inference problem successfully provides N and τ for small and large attacks. D can be estimated for large attacks. Posterior PDFs are sharper for large attacks and for high-dose attacks. Higher temporal resolution may smoothen the PDFs slightly. When conditioning on a short time series, the Bayesian method may suggest multiple hypotheses, supported to differing degrees by the data. In some cases, e.g., Case F, the data might initially support the wrong hypothesis, but the correct characterization is recovered as more data becomes available.

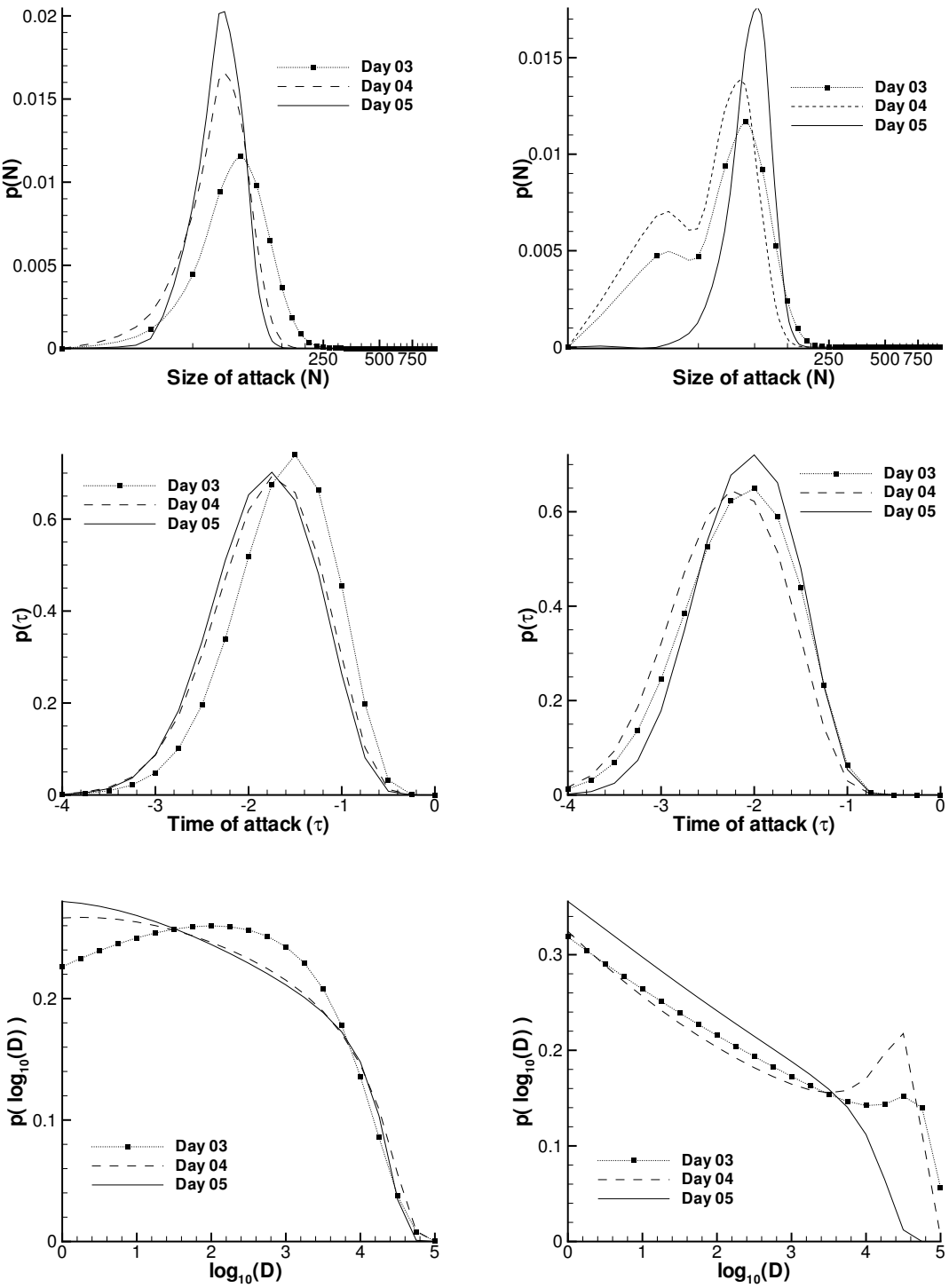

Figure 8. Posterior PDFs for N (top), τ (middle), and $\log D$ (bottom) based on the time series for Case A (left) and Case B (right), as tabulated in Table 2. Data are collected at 6-hour intervals in both cases. The correct values for $\{N, \tau, \log_{10}(D)\}$ in Case A are $\{10^2, -0.75, 10^0\}$; in Case B they are $\{10^2, -2.25, 10^2\}$. In both cases, PDFs are reported after 3-, 4- and 5-day observational periods (dotted, dashed, and solid lines respectively).

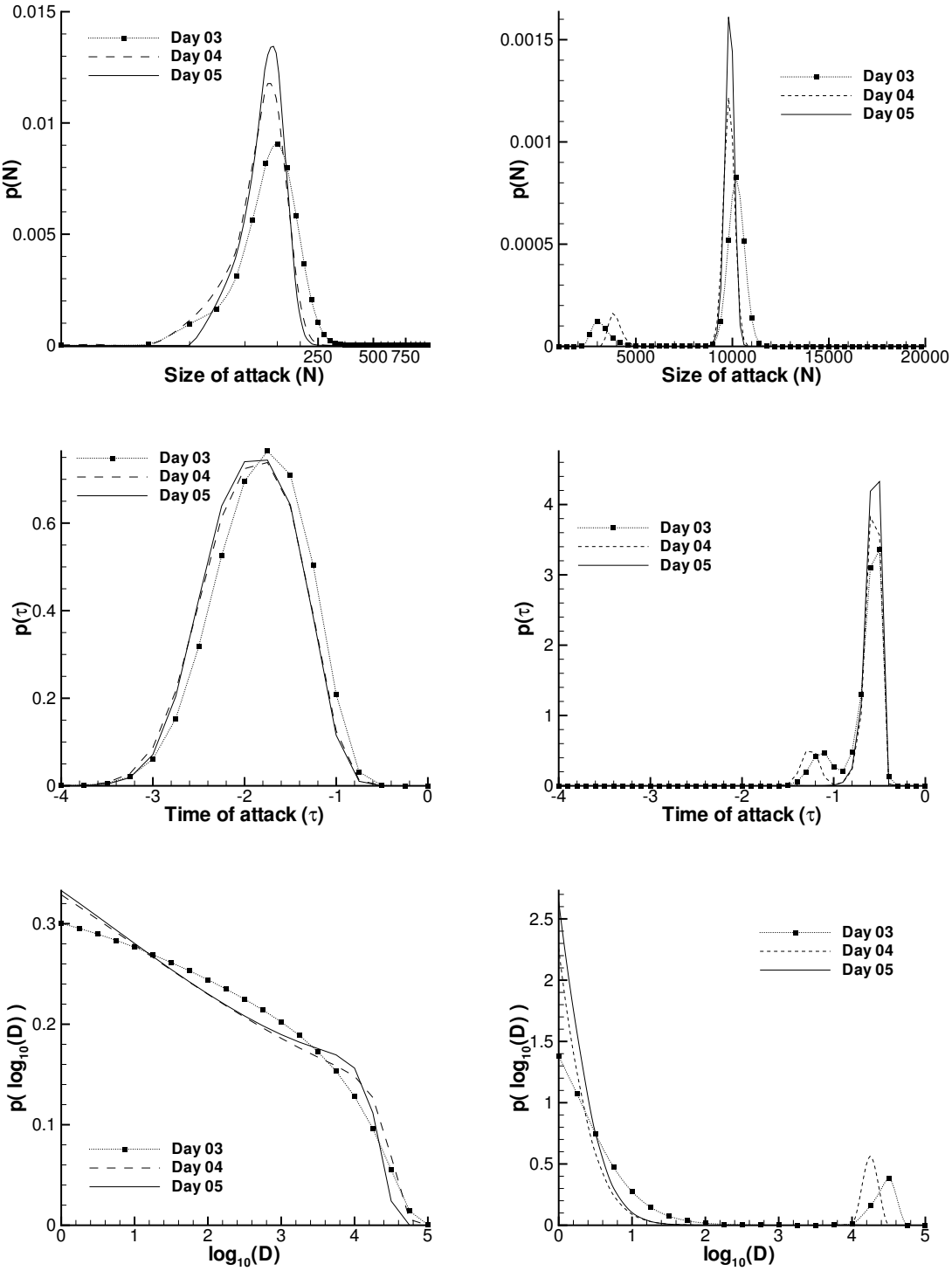

Figure 9. Posterior PDFs for N (top), τ (middle), and $\log D$ (bottom) based on the time series for Case C (left) and Case D (right), as tabulated in Table 2. Data are collected at 6-hour intervals in both cases. The correct values for $\{N, \tau, \log_{10}(D)\}$ in Case C are $\{10^2, -2.25, 10^4\}$; in Case D they are $\{10^4, -0.05, 10^0\}$. In both cases, PDFs are reported after 3-, 4- and 5-day observational periods (dotted, dashed, and solid lines respectively).

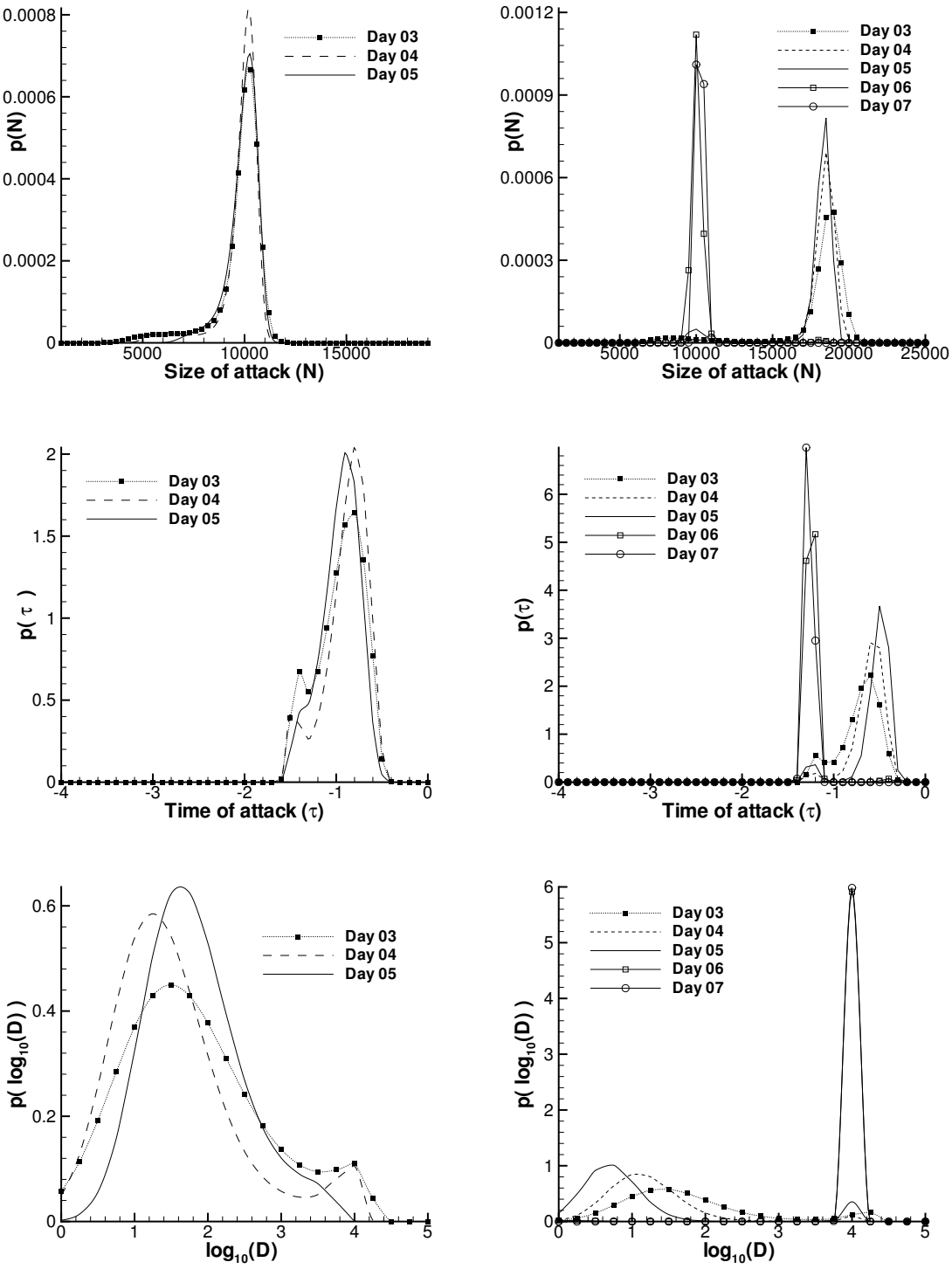

Figure 10. Posterior PDFs for N (top), τ (middle), and $\log D$ (bottom) based on the time series for Case E (left) and Case F (right), as tabulated in Table 2. Data are collected at 6-hour intervals in both cases. The correct values for $\{N, \tau, \log_{10}(D)\}$ in Case E are $\{10^4, -1.0, 10^2\}$; in Case F they are $\{10^4, -1.25, 10^4\}$. In both cases, PDFs are reported after 3-, 4- and 5-day observational periods (dotted, dashed, and solid lines respectively), but Case F also includes PDFs at Day 6 (solid lines with filled squares) and at Day 7 (solid lines with filled circles).

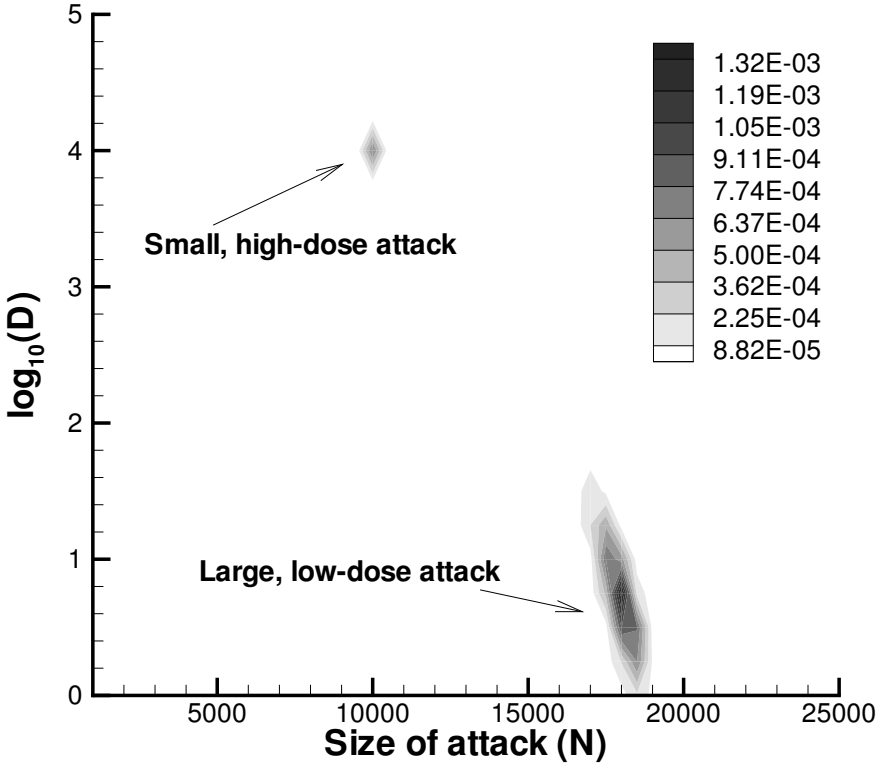

Figure 11. The joint probability density $p(N, \log_{10}(D))$ obtained after 5 days of data for Case F. We clearly see a dual characterization—a larger low-dose attack and a smaller high-dose attack.

4.4 Inference of attack parameters under variable doses

In this section we conduct eight tests corresponding to more realistic conditions. In the first four (Cases I, Ia, II, and IIa) we relax the assumption of a constant dose D; instead, the infected people receive a range of doses commensurate with atmospheric dispersion. However, the disease is still assumed to evolve per Wilkening's Model A2, with the same model providing $C(T,D)$ to the inference procedure. In the second set of tests (Cases III, IIIa, IV, and IVa), we retain distributed doses and additionally relax the second assumption: data is generated with Model D, while the inference procedure still uses Model A2 to evaluate the incubation period distribution. This mismatch introduces a degree of realism into the inference process since the host-pathogen interaction for humans and anthrax will seldom be characterized accurately.

In order to obtain a realistic distribution of doses in a geographically distributed population, we first simulate an explosive point release of spores at a height of 100 meters with a Gaussian plume model, thus exposing different numbers of people to varying doses as described in Appendix A. We see from Figure A.1 that given a quantity of spores, the number of people infected depends on the total population in the domain, the orientation of the plume, and the population distribution. A release does not lead to many infected people if the high concentration isopleths of the plume miss the localized regions of high population density.

Tables 4 and 5 list the time series obtained from all eight cases. The time series have a resolution of 6 hours, with successive 24-hour intervals indicated in the tables. As noted in Appendix A, these cases correspond to two choices of population size ($p_{\text{exposed}} = 10^3$ for Cases Ia, II, IIIa and IV; $p_{\text{exposed}} = 10^4$ for Cases I, IIa, III and IVa) combined with two choices of plume orientation ($\theta = 170^\circ$ for Cases I, Ia, III and IIa; $\theta = 125^\circ$ for Cases II, IIa, IV and IVa). The latter orientation directs the plume over a more population-dense region. Tables 4 and 5 also report quantiles of the dose distribution $D_1, D_{25}, D_{50}, D_{75}$, and D_{99}. That is, 1% of the population receives a dose of D_1 spores or less, 75% of the population receives less than D_{75} spores, and D_{99} is near the maximum dose. In Figure A.2 (Appendix A), we plot dose distributions corresponding to the cases given in Tables 4 and 5. Note that while the doses may easily span two orders of magnitude, about 80% of the infected people lie within a one-decade range of doses. We will essentially try to estimate a representative dose D for this range, by fitting the probabilistic model developed in Section 4.1, which assumes a constant dose. This is a source of model error, adding to the uncertainty caused by incomplete observations and the inherent stochasticity of the data. This model error is not expected to diminish with additional data, and thus one of the aims of this investigation is to quantify it.

4.4.1 Inference of attack parameters without incubation model mismatch

We begin with results from Cases Ia, I, II, and IIa—i.e., eliminating the assumption that each infected person receives the same dose of anthrax spores, but simulating and inferring disease progression with Wilkening's Model A2.

Figs. 12, 13, 14, and 15 show posterior PDFs for $\{N, \tau, \log_{10}(D)\}$ conditioned on the time series in Tables 4 and 5. Table 6 reports the MAP estimates and the 90% CIs for $\{N, \tau, \log_{10}(D)\}$ after 5

Table 4. Time series obtained from eight simulated outbreaks with variable doses. Cases I, Ia, II, and IIa are simulated using Wilkening's Model A2, with the attack parameters—N, τ, and the dose distribution—indicated at the bottom of the table. Cases III, IIIa, IV and IV are simulated using Wilkening's Model D. $\bar{D}$ is the average dose for the N infected individuals. The table has been divided into 24-hour sections, where the values n_i in each section can be summed to produce the low-resolution time series used to investigate the effect of temporal resolution. The dose distribution is represented by its quantiles D_1, D_{25}, D_{50}, D_{75}, and D_{99}; x% of the population receives a dose of D_x or less. Table 5 continues the time series from Day 5 to Day 8.

Time (days)	Case Ia	Case I	Case II	Case IIa	Case IIIa	Case III	Case IV	Case IVa
0.0	1	3	2	5	1	1	1	3
0.25	2	3	2	8	1	8	5	14
0.50	0	6	1	8	0	20	6	36
0.75	4	12	5	27	1	16	13	81
1.00	1	14	7	46	3	9	12	77
1.25	2	26	12	57	2	18	14	94
1.50	2	28	9	85	2	28	13	123
1.75	6	49	16	94	1	30	13	132
2.0	6	57	9	133	2	37	17	129
2.25	5	65	20	134	2	27	15	159
2.50	7	68	12	139	4	41	17	126
2.75	6	53	18	163	2	39	14	149
3.0	11	80	15	138	3	34	9	131
3.25	8	62	15	180	2	32	14	129
3.50	9	89	21	140	3	25	16	136
3.75	8	106	16	164	6	33	12	100
4.00	17	70	20	180	4	27	14	125
4.25	12	65	21	136	5	33	11	104
4.50	9	87	8	147	3	33	6	110
4.75	3	87	8	151	5	23	11	106
5.0	6	76	7	127	6	23	15	90
N	318	2989	454	4537	161	1453	453	4453
τ	-1.5	-1.5	-1.5	-1.25	-0.75	-0.75	-0.75	-0.5
$\bar{D}$	2912.8	2776.8	13,870.5	13,150.4	3603.5	3660.77	16,941	16,532
$D_1 \times 10^{-2}$	0.53	0.65	1.39	1.32	3.41	2.65	3.1	3.0
$D_{25} \times 10^{-3}$	1.23	1.15	3.96	3.47	1.99	2.13	9.8	9.45
$D_{50} \times 10^{-4}$	0.29	0.26	1.34	1.24	0.33	0.35	1.65	1.57
$D_{75} \times 10^{-4}$	0.41	0.39	1.91	1.87	0.48	0.48	2.09	2.07
$D_{99} \times 10^{-4}$	0.83	0.87	5.79	5.91	0.92	0.95	6.74	6.52

Table 5. Continuation of Table 4 beyond Day 5. Time series obtained from 4 simulated outbreaks with variable doses. Cases I, Ia, II, and IIa are simulated using Wilkening's Model A2, with the attack parameters—N, τ, and the dose distribution—indicated at the bottom of the table. $\bar{D}$ is the average dose for the N infected individuals. The table has been divided into 24-hour sections, where the values n_i in each section can be summed to produce the low-resolution time series used to investigate the effect of temporal resolution. The dose distribution is represented by its quantiles D_1, D_{25}, D_{50}, D_{75}, and D_{99}; $x\%$ of the population receives a dose of D_x or less.

Time (days)	Case Ia	Case I	Case II	Case IIa
5.25	9	70	16	129
5.50	8	91	8	109
5.75	10	79	9	147
6.00	9	86	12	126
6.25	8	82	13	108
6.50	7	55	9	114
6.75	7	69	7	90
7.0	6	75	8	96
7.25	8	61	6	88
7.50	4	67	6	77
7.75	6	65	8	75
8.00	2	62	6	69
N	318	2989	454	4537
τ	-1.5	-1.5	-1.5	-1.25
$\bar{D}$	2912.8	2776.8	13,870.5	13,150.4
$D_1 \times 10^{-2}$	0.53	0.65	1.39	1.32
$D_{25} \times 10^{-3}$	1.23	1.15	3.96	3.47
$D_{50} \times 10^{-4}$	0.29	0.26	1.34	1.24
$D_{75} \times 10^{-4}$	0.41	0.39	1.91	1.87
$D_{99} \times 10^{-4}$	0.83	0.87	5.79	5.91

days of data. Since the true doses are distributed, we use the log of the median dose, $\log_{10}(D_{50})$, as a reasonable value for comparison to the posterior $\log_{10}(D)$.

First consider Figs. 12 and 13, corresponding to Cases Ia and I. These attacks have similar dose distributions but differ by an order of magnitude in N. In both cases, the MAP estimate of τ nearly coincides with the true value after only 3 days of data. In Case Ia, the MAP estimate of N deviates from the true value by approximately 20%, but the 90% CIs bracket the correct N quite easily. In Case I, the PDF for N initially favors an inaccurate characterization (a peak at $N \approx 4000$) but by Day 5, assumes a bimodal shape with a peak close to the correct characterization. Dose is the most difficult parameter to estimate in Case Ia—the marginal PDF of $\log_{10}(D)$ remains rather broad at all times. In the larger-N Case I, however, the posterior on $\log_{10}(D)$ at least indicates that the attack is not a low dose (i.e., $D_{50} \leq \mathrm{ID}_{25}$) event. Also in Case I, conditioning on the high resolution time-series provides more structure to the PDF; the posterior densities on $\log_{10}(D)$ and even on τ are more prominently bimodal, indicating that inference is inconclusive, and more observations will be required to obtain a unique characterization. For reference, both Figs. 12 and 13 include a further set of PDFs conditioned on data through Day 7; MAP estimates from these posteriors generally show even closer agreement with the true values of $\log_{10}(D_{50})$ and N.

Inference is considerably less challenging in Cases II and IIa, corresponding to Figs. 14 and 15. Because the doses are higher ($D_{50} > \mathrm{ID}_{50}$), the variance of the incubation period distribution is smaller. The time of attack τ is captured with only 3 days of data, as is a representative $\log_{10}(D)$ for the large N attack (Case IIa). With 5 days of data, MAP estimates for N are close to the correct values in both cases, as is the MAP estimate of $\log_{10} D$ in Case II. Here, conditioning on the higher-resolution time series yielded little gain over the lower-resolution time series. In Case II, MAP estimates of τ based on 6-hour data are in fact inaccurate on Days 3 and 4, recovering the correct characterization after 5 days of data.

In general, therefore, many of the behaviors discussed in Section 4.3 are repeated in the present cases. Dose D is difficult to estimate for small N attacks, while the time τ is always easy to infer. We can bound the size N of the attack quite accurately for all cases. MAP estimates of N obtained from 5 days of data are always within 20% of the correct value. Further, the 90% CIs at Day 5 for N, τ, and $\log_{10}(D)$ almost always bracket the true attack parameters. Finer temporal resolution Δt may better capture the evolution of the outbreak, but has a relatively minor impact on summaries of the posterior; MAP estimates obtained from the low and high-resolution time series are similar, as are the 90% CIs. Thus, while the errors incurred in fitting variable-dose data to a constant-dose inference model are not negligible, the current formulation provides a reasonable and useful characterization of the BT attack.

4.4.2 Inference of variable dose attack with incubation model mismatch

We now proceed to Cases III, IIIa, IV, and IVa. As noted above, these cases introduce a systematic difference between the simulated evolution of the disease in infected persons and the model used to interpret the observed data. We simulate BT attacks using Wilkening's Model D (i.e., sampling the incubation period distribution in Eq. 5), but infer the attack parameters using Model A2. As in Section 4.4.1, the infected population receives a distribution of doses (see Appendix A) but the

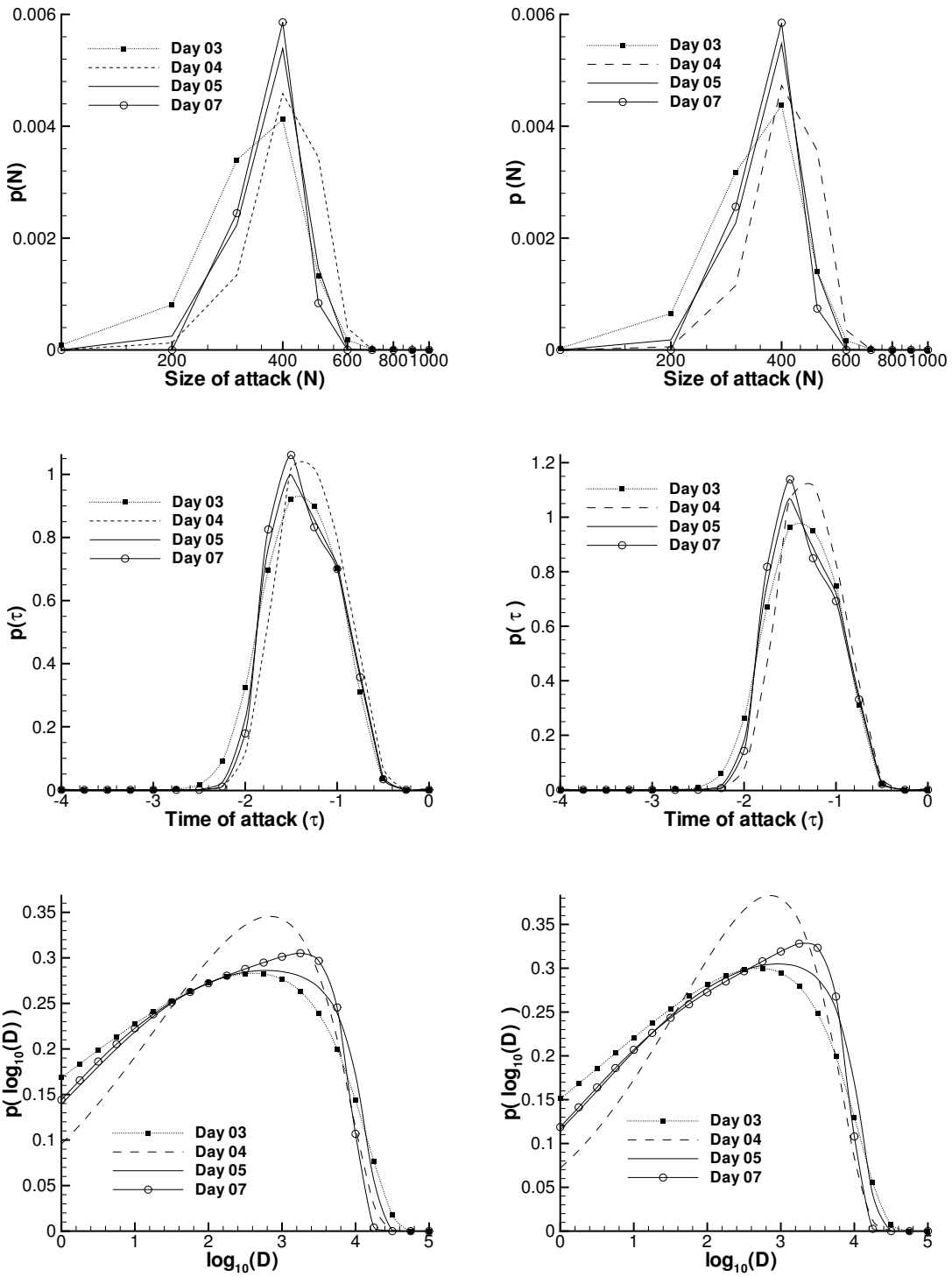

Figure 12. Posterior PDFs for N (top), τ (middle), and $\log_{10} D$ (bottom) based on the time series for Case Ia, as tabulated in Tables 4 and 5. Lower-resolution data (collected in 24-hour intervals) yields the PDFs on the left, while higher-resolution data yields the PDFs on the right. Correct values for $\{N, \tau, \log_{10}(D)\}$ are $\{318, -1.5, 3.46\}$, where the "correct" representative dose is taken to be $\log_{10}(D_{50})$. In both cases, PDFs are reported after 3-, 4-, 5-, and 7-day observational periods.

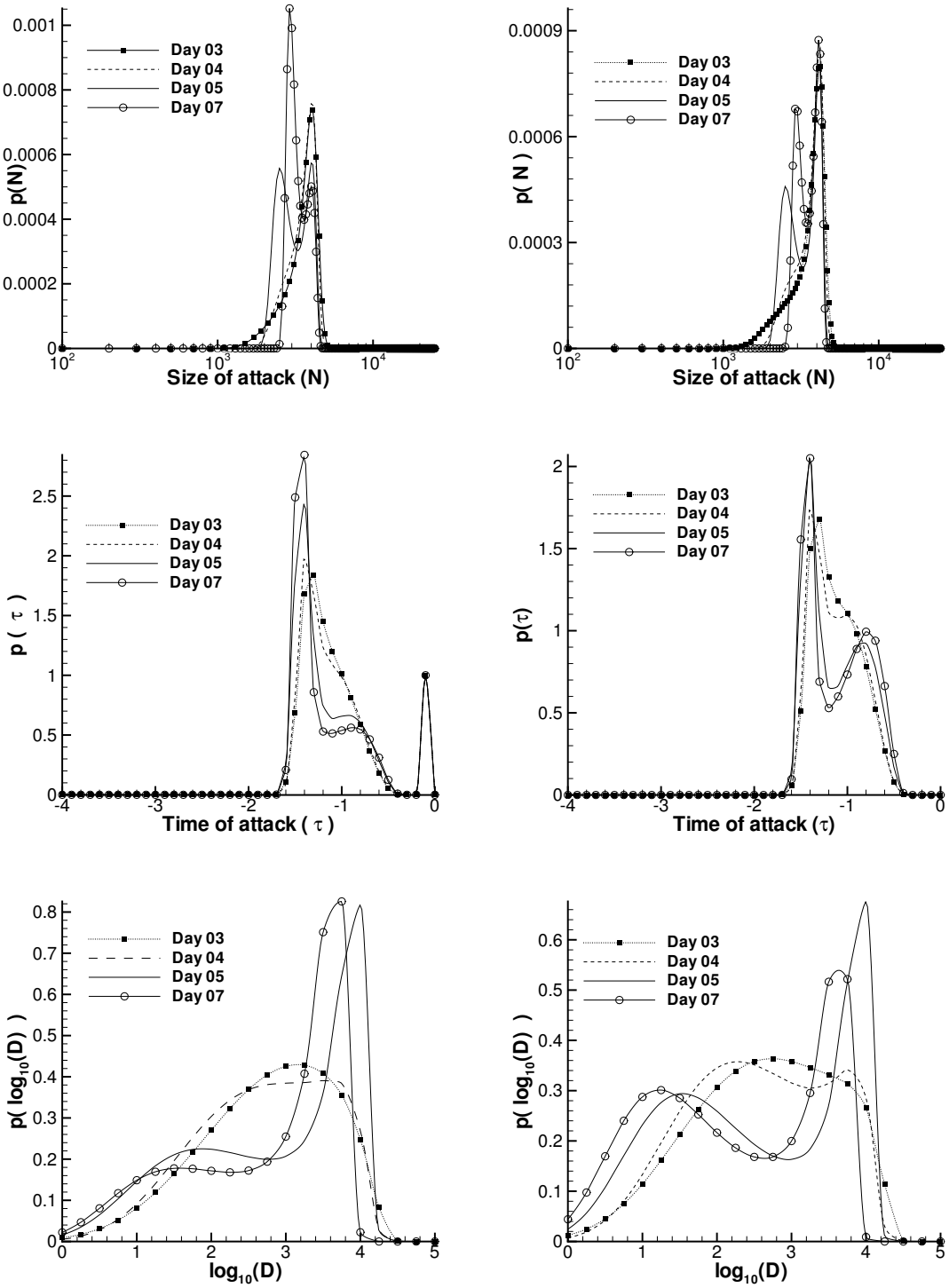

Figure 13. Posterior PDFs for N (top), τ (middle), and $\log_{10} D$ (bottom) based on the time series for Case I, as tabulated in Tables 4 and 5. Lower-resolution data (collected in 24-hour intervals) yields the PDFs on the left, while higher-resolution data yields the PDFs on the right. Correct values for $\{N, \tau, \log_{10}(D)\}$ are $\{2989, -1.5, 3.41\}$, where the "correct" representative dose is taken to be $\log_{10}(D_{50})$. In both cases, PDFs are reported after 3-, 4-, 5-, and 7-day observational periods.

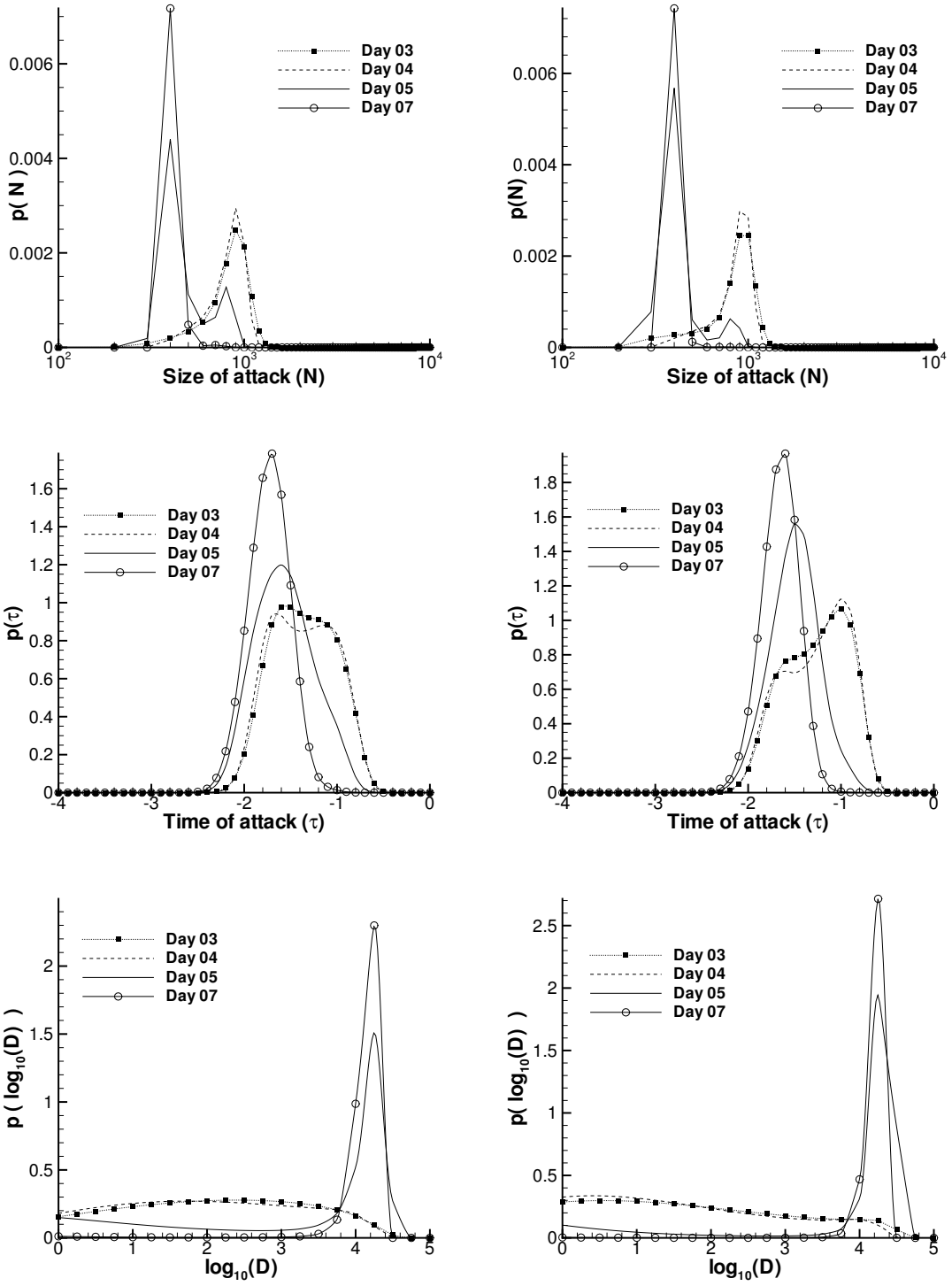

Figure 14. Posterior PDFs for N (top), τ (middle), and $\log_{10} D$ (bottom) based on the time series for Case II, as tabulated in Tables 4 and 5. Lower-resolution data (collected in 24-hour intervals) yields the PDFs on the left, while higher-resolution data yields the PDFs on the right. Correct values for $\{N, \tau, \log_{10}(D)\}$ are $\{454, -1.5, 4.13\}$, where the "correct" representative dose is taken to be $\log_{10}(D_{50})$. In both cases, PDFs are reported after 3-, 4-, 5-, and 7-day observational periods.

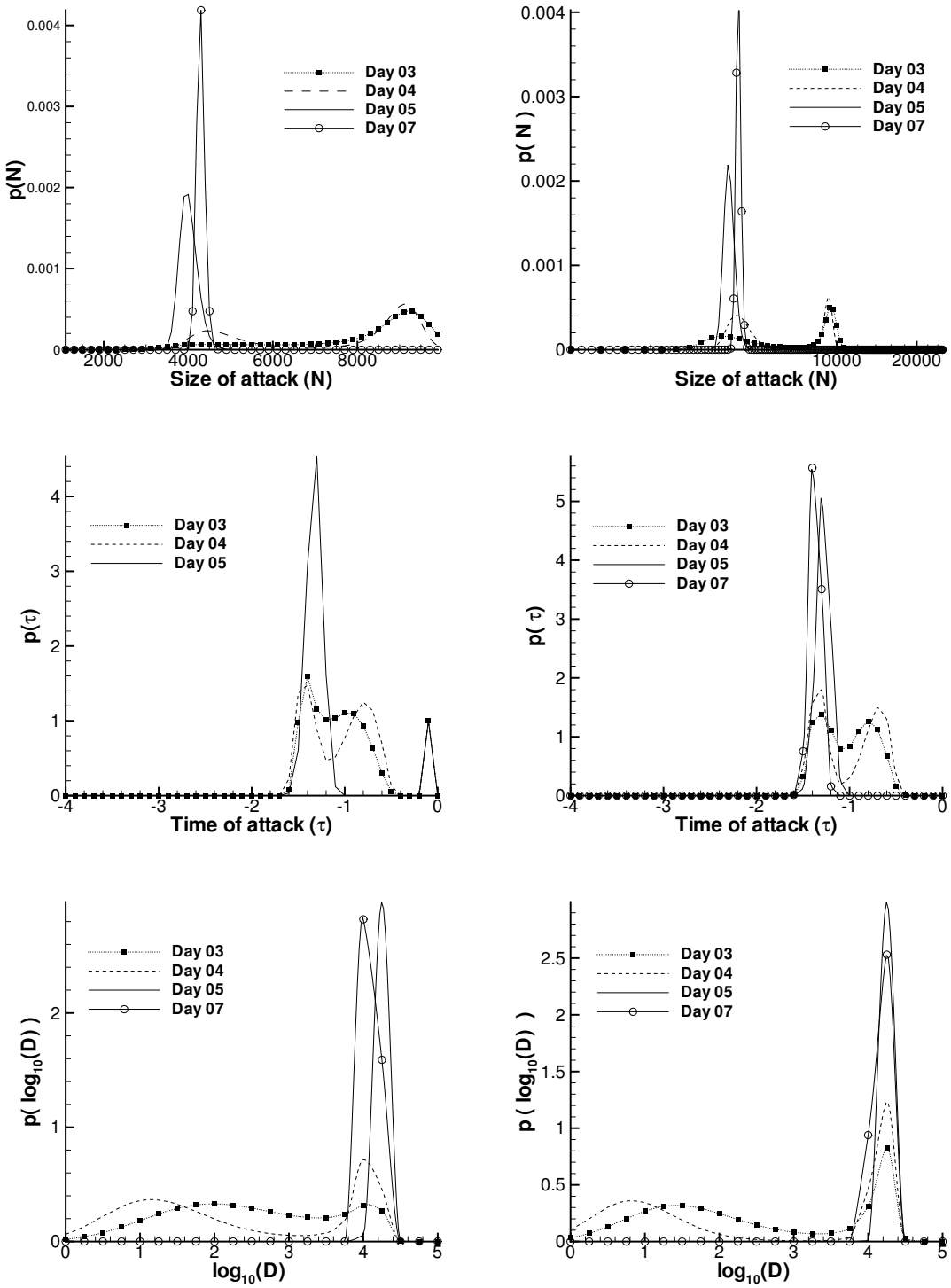

Figure 15. Posterior PDFs for N (top), τ (middle), and $\log_{10} D$ (bottom) based on the time series for Case IIa, as tabulated in Tables 4 and 5. Lower-resolution data (collected in 24-hour intervals) yields the PDFs on the left, while higher-resolution data yields the PDFs on the right. Correct values for $\{N, \tau, \log_{10}(D)\}$ are $\{4537, -1.25, 4.09\}$, where the "correct" representative dose is taken to be $\log_{10}(D_{50})$. In both cases, PDFs are reported after 3-, 4-, 5-, and 7-day observational periods.

Case	N	τ	$\log_{10}(D)$
Ia (6-hr resolution)	400, (233.6 – 581.9)	-1.5, (-2.00 – -0.795)	3.0, (0.37 – 3.99)
Ia (24-hr resolution)	400, (230.4 – 582.2)	-1.5, (-2.04 – -0.78)	2.75, (0.32 – 4.00)
	{318}	{−1.5}	{3.46}
I (6-hr resolution)	4100, (2334 – 4439)	-1.4, (-1.57 – -0.64)	4.00, (0.715 – 4.147)
I (24-hr resolution)	4000, (2281 – 4358)	-1.4, (-1.59 – -0.70)	4.00, (0.91 – 4.173)
	{2989}	{−1.5}	{3.41}
II (6-hr resolution)	400, (305.5 – 981.6)	-1.5, (-1.98 – -1.08)	4.25, (0.68 – 4.72)
II (24-hr resolution)	400, (327.0 – 984.7)	-1.6, (-2.10 – -1.03)	4.25, (0.36 – 4.69)
	{454}	{−1.5}	{4.13}
IIa (6-hr resolution)	3900, (3686 – 4340)	-1.3, (-1.48 – -1.14)	4.25, (4.05 – 4.72)
IIa (24-hr resolution)	4000, (3709 – 4433)	-1.5, (-1.55 – -1.18)	4.25, (4.04 – 4.72)
	{4537}	{−1.25}	{4.09}

Table 6. Cases I, Ia, II, IIa; MAP estimates and 90% credibility intervals (in parentheses) for N, τ, and $log_{10}(D)$ conditioned on data through Day 5. Correct values for N and τ are in { }. The "correct" representative dose is taken to be $\log_{10}(D_{50})$, also in { }.

model used in the inference process assumes a constant dose.

Figs. 16 and 17 show posterior PDFs for $\{N, \tau, \log_{10}(D)\}$ conditioned on the time series in Table 4. As described in the preceding section, finer resolution in the time series does not have a great impact on the posterior, and hence we only plot PDFs resulting from daily observations in each case. Several features are worth highlighting. First, the dose is identified much more closely in Cases IV and IVa, where both N and $\log_{10}(D)$ are higher, than in Cases III and IIIa. Indeed, $p(\log_{10}(D))$ in the low-dose small-N Case IIIa remains broad at all times. In Case III, after only 3 days of data, we observe a dual characterization of the outbreak: $N \approx 700$ and, to a larger extent, $N \approx 2000$. However, $p(N)$ becomes unimodal as additional data become available. In fact, PDFs for all three parameters in all four cases are unimodal by Day 5. The resulting MAP estimates and 90% CIs for $\{N, \tau, \log_{10}(D)\}$ are reported in Table 7. In contrast to Section 4.4.1, MAP estimates for N and τ are not within 20% of the true values. With the exception of Case IIIa, N is smaller than it should be, and in all cases τ is more negative than it should be.

A qualitative explanation for these discrepancies is advanced as follows. Since Model A2 predicts shorter incubation periods than Model D (recall Figure7), the epidemic curve as simulated with Model D will rise more slowly that predicted by Model A2. When this data is interpreted using Model A2, it is reasonable to expect the posterior to compensate for the slower rise by underestimating N, i.e., by suggesting a smaller outbreak. Simultaneous estimation of D and τ raises a few additional complications, however. Recall that the posterior of D is centered quite close to its true value in Cases IV and IVa, and to a lesser extent in Case III. But in the likelihood function, this dose enters the wrong model. Using a "correct" dose in Model A2 is akin to using a much larger dose in Model D; both situations yield shorter incubation periods. Now draw a parallel with Case F in Section 4.3. There, we found that a large-dose small-N attack and a small-dose large-N attack gave rise to very similar patient data during the first five days of an outbreak. Moreover, we found that N and τ were positively correlated (and that both were negatively correlated with D): the small-N mode of the posterior also favored more negative τ, i.e., attacks that occurred approximately one day earlier. The very same correlations affect inference in the present cases. Incubation model mismatch is roughly equivalent to an overestimation of D, which is compensated for by underestimating N and τ.

In summary, Table 7 shows that MAP estimates for N are typically within a factor of two below the true result and that τ is estimated roughly a day too early.

4.5 The Sverdlovsk anthrax outbreak of 1979

We now address the characterization of the Sverdlovsk anthrax outbreak. It is suspected that on 2 April 1979, a high-grade anthrax formulation was accidentally released from a military facility in Sverdlovsk (today, Yekaterinburg), Russia. The resulting outbreak lasted 42 days, and patient data was collected on a daily basis [7]. Characterizing the Sverdlovsk case presents significant challenges. It corresponds to a low-dose "attack" infecting fewer than 100 people. Wilkening [42] estimates that the average dose was either around 2–3 spores, based on his Model A, or around

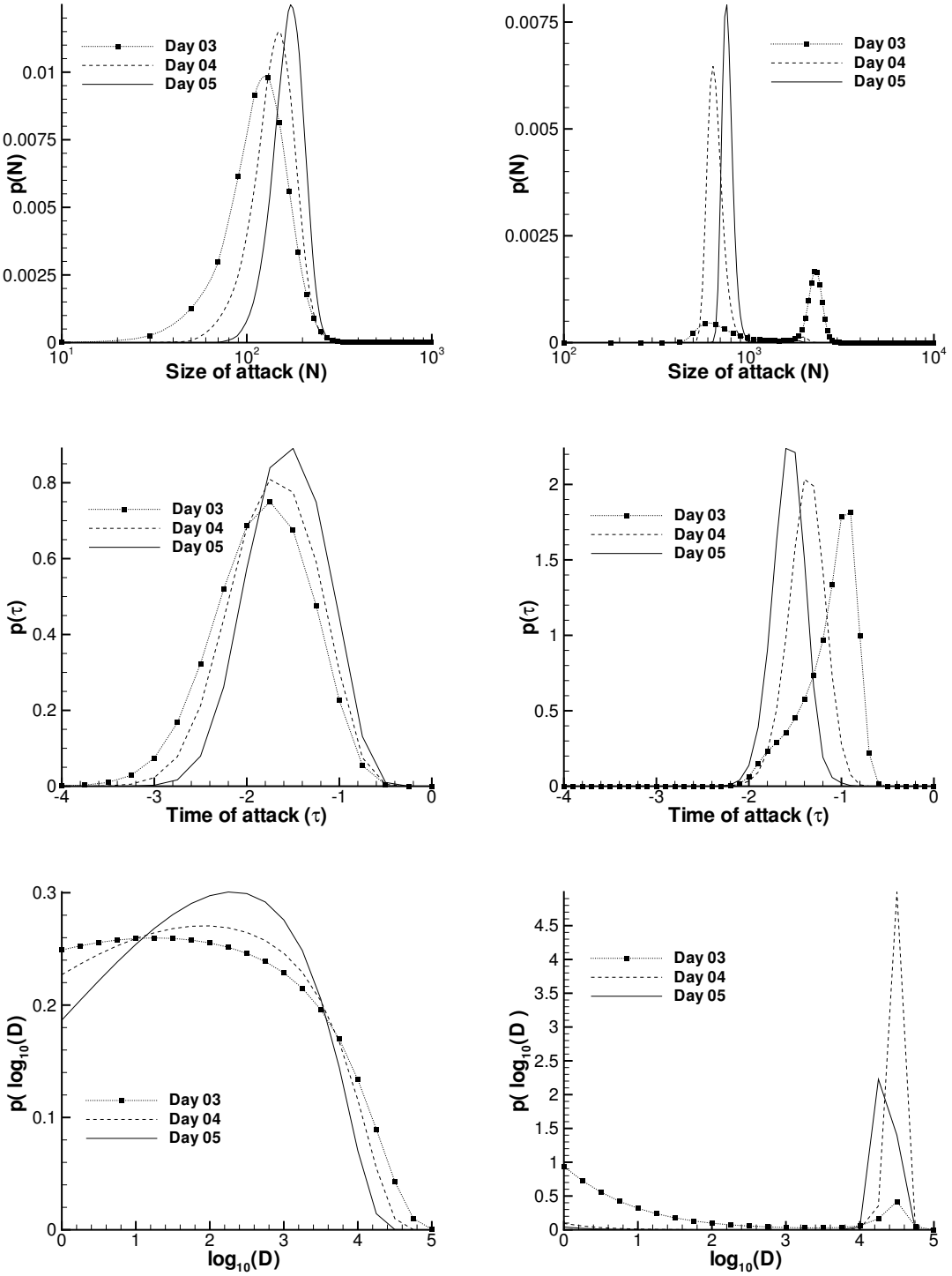

Figure 16. Posterior PDFs for N (top), τ (middle), and $\log_{10} D$ (bottom) based on daily time series for Case IIIa (left) and Case III (right). Correct values for $\{N, \tau, \log_{10}(D)\}$ are $\{161, -0.75, 3.52\}$ (Case IIIa) and $\{1453, -0.75, 3.54\}$ (Case III), where the "correct" representative dose is taken to be $\log_{10}(D_{50})$. In both cases, PDFs are reported after 3-, 4-, and 5-day observational periods (dotted, dashed and solid lines respectively).

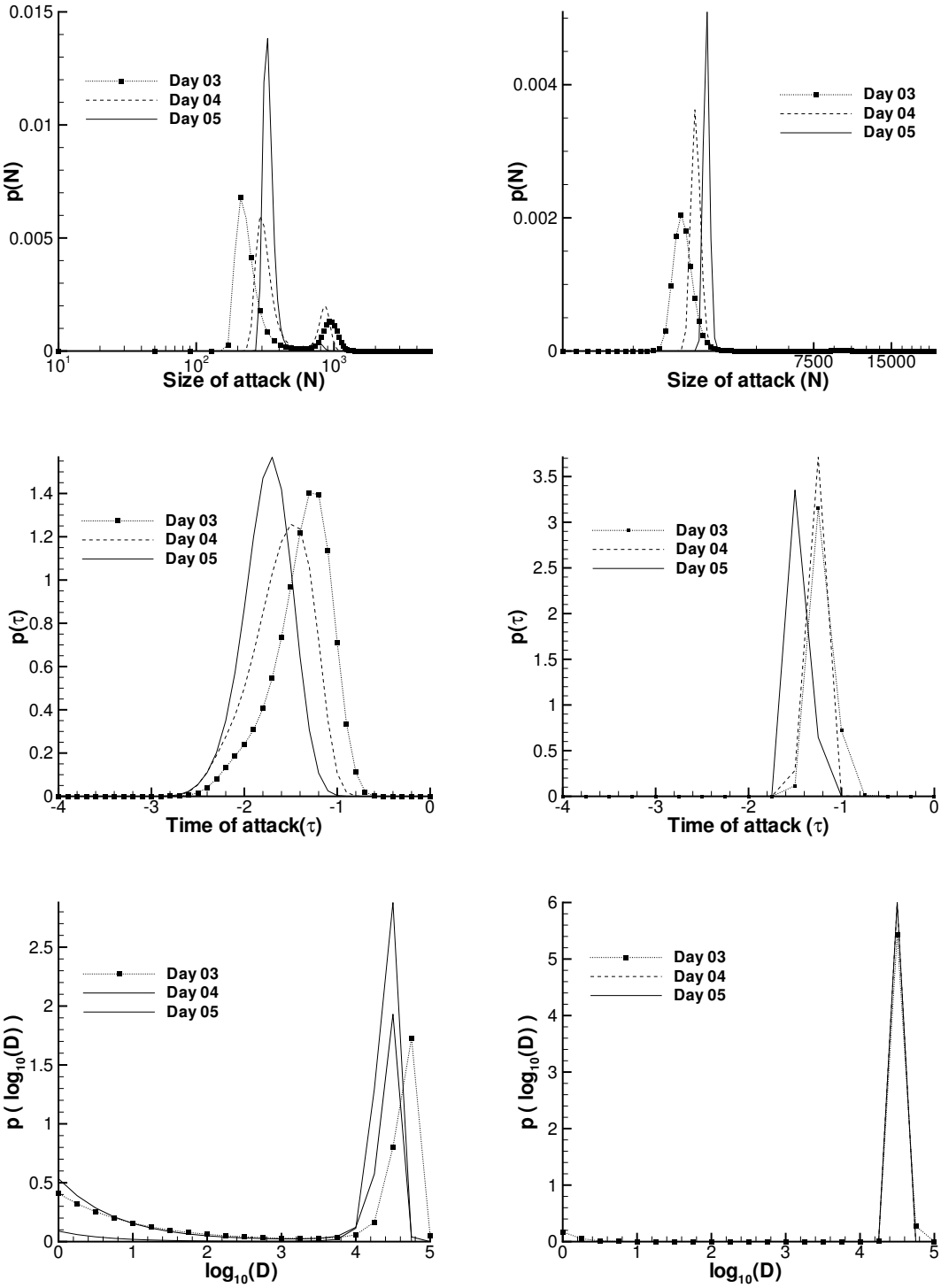

Figure 17. Posterior PDFs for N (top), τ (middle), and $\log_{10} D$ (bottom) based on daily time series for Case IV (left) and Case IVa (right). Correct values for $\{N, \tau, \log_{10}(D)\}$ are $\{453, -0.75, 4.22\}$ (Case IV) and $\{4453, -0.5, 4.20\}$ (Case IVa), where the "correct" representative dose is taken to be $\log_{10}(D_{50})$. In both cases, PDFs are reported after 3-, 4-, and 5-day observational periods (dotted, dashed and solid lines respectively).

Case	N	τ	$\log_{10}(D)$
IIIa (6-hr resolution)	170, (130.1 – 243.6)	-1.5, (-2.3 – -0.86)	2.0, (0.23 – 3.74)
IIIa (24-hr resolution	170, (125.1 – 238.8)	-1.5, (-2.4 – -0.94)	2.5, (0.255 – 3.78)
	$\{161\}$	$\{-0.75\}$	$\{3.52\}$
III (6-hr resolution)	780, (722 – 945.5)	-1.7, (-2.03 – -1.42)	4.25, (4.02 – 4.723)
III (24-hr resolution	760, (701 – 891.7)	-1.6, (-1.91 – -1.31)	4.25, (4.04 – 4.724)
	$\{1453\}$	$\{-0.75\}$	$\{3.54\}$
IV (6-hr resolution)	330, (297.2 – 668.6)	-1.7, (-2.23 – -1.40)	4.5, (1.4 – 4.72)
IV (24-hr resolution)	330, (296.3 – 705.3)	-1.7, (-2.26 – -1.38)	4.5, (1.45 – 4.72)
	$\{453\}$	$\{-0.75\}$	$\{4.22\}$
IVa (6-hr resolution)	2900, (2728 – 3056)	-1.5, (-1.90 – -1.1)	4.5, (4.275 – 4.725)
IVa (24-hr resolution)	2900, (2741 – 3064)	-1.5, (-1.97 – -1.26)	4.5, (4.275 – 4.725)
	$\{4453\}$	$\{-0.5\}$	$\{4.20\}$

Table 7. Cases III, IIIa, IV, and IVa: MAP estimates and the 90% credibility intervals (in parentheses) for N, τ, and $log_{10}(D)$ conditioned on data through Day 5. Correct values for N and τ are in $\{\ \}$. The "correct" representative dose is taken to be $\log_{10}(D_{50})$, also in $\{\ \}$.

300 spores based on his Model D; Meselson [7] estimates 100–2000 spores as the likely dose. The first patient presented symptoms on 4 April 1979. Around 12 April, tetracycline was administered around Sverdlovsk; around 15 April, people were vaccinated. These prophylactic measures probably cured a few people and increased the incubation period in others. Further, the available data almost certainly contains some recording errors. Errors in the data, the effect of prophylaxis (which is not modeled in our likelihood function), and the small size of the infected population are expected to stress our inference process.

In Figure 18 we plot the posterior densities of N and τ based on the data in [7]. Model A2 is used for inference. After 9 days of data, the MAP estimate for N centers around 50, though the earlier PDFs underestimate N. The 90% CI for N is $[41.15, 66.49]$. In comparison, 70 people are believed to have died [7, 35] and 80 are believed to have been infected [35]. The time of release was easy to infer (the MAP estimate of τ is -2, i.e., 2 April 1979); the 90% CI for τ is $[-3.22, -1.38]$. PDFs for the dose (omitted here) were indeterminate (the 90% CI for $\log_{10}(D)$ spans $[0.18, 3.5]$); further the average dose at Sverdlovsk is unknown. However, by 13 April (i.e., Day 9, the day of the start of the prophylaxis campaign and 2 days before the vaccination campaign), our estimate for N is certainly within a factor of two of the correct value, and it is clear that the outbreak will

affect fewer than 200 people. However, approximately 59,000 people in the Chkalovskiy *raion* were impacted by the medical measures; 80% were vaccinated at least once [7].

Guillemin [82] documents the public health response undertaken by the Soviet authorities once the Sverdlovsk epidemic was detected. In conjunction with the quantitative analysis presented in this section, [82] illustrates the difficulties and pitfalls faced by medical responders when the origin and the extent of an epidemic are unknown. Without a model for inhalational anthrax, observations of symptomatic patients could not be used to prove (or disprove) any hypothesis regarding the epidemic's genesis. Indeed, Soviet authorities held that the epidemic was caused by infected meat and spent considerable effort searching for it. The response also prompted a significant external component, engaging many medical personnel and officials from outside Sverdlovsk; yet by Day 4 (8 April 1979) it was clear that the epidemic was small (Figure 18, left) and could be handled by local authorities. (Sverdlovsk was a military-industrial city with a population of 1.2 million [7].) Guillemin [82] also describes efforts to disinfect buildings and trees by hosing them down with disinfectants; yet with knowledge of τ (Figure 18, right) and meteorological conditions, the bounds of the affected region could have been established (as Meselson did in 1994 [7]) and the public health response suitably targeted. Therefore, a quantitative model and an inferential capability could have been of assistance in 1979. The lessons are equally applicable in contemporary bioterror scenarios.

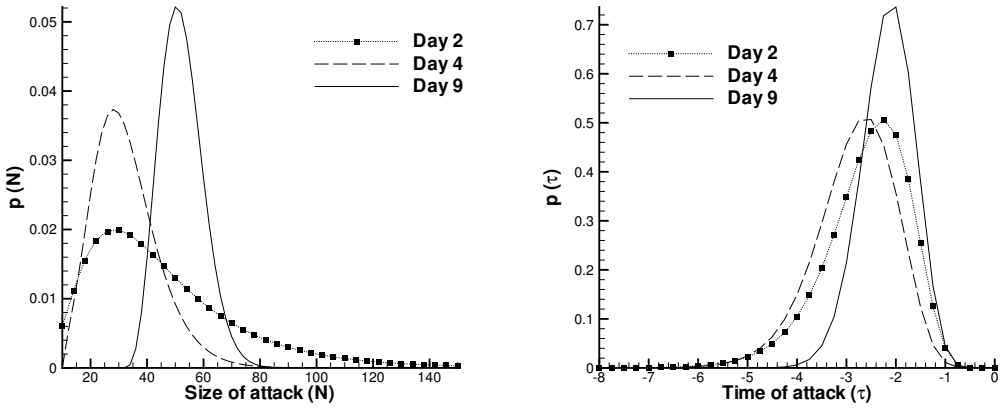

Figure 18. PDFs of N (left) and τ (right) for the Sverdlovsk outbreak.

5 Characterizing outbreaks from partial observations: The case for contagious diseases

In Section 4, we described how one may estimate the characteristics of an outbreak caused by a (nearly) instantaneous infection of N individuals from partial observations. While the analysis assumed that the disease was non-contagious, the same approach could also be applied quite easily to a contagious disease *provided the incubation period of the disease was larger than the period of observation / data collection*; in such a case, the time-series would consist solely of the index cases and consequently, the problem would be no different from the one for non-contagious diseases.

The difference between an outbreak of a non-contagious disease and a contagious one is, of course, the ability of the disease to spread. The spread of a disease depends on the intrinsic transmissibility of the pathogen (for example, the common cold spreads far more easily than tuberculosis) as well as the mixing patterns of the population that allow the transmission of the pathogen. The mixing pattern can be enormously influential in deciding the spreading rate - smallpox epidemics have exhibited reproductive ratios R_0 varying between 3 and 17 [9] and consequently, when characterizing an outbreak of a contagious disease, it becomes necessary to infer both the transmissibility of the pathogen and the social mixing pattern. This is particularly true if the pathogen is an emerging one (e.g. avian flu) which may be expected to undergo evolutionary changes (affecting transmissibility) from outbreak to outbreak, over time.

Epidemics of contagious diseases are conventionally simulated with SEIR models. An individual in a population is expected to be in one of the four states (Susceptible–Exposed–Infectious–Removed). Removal can mean death or recovery. The time spent by an individual in the exposed and infectious states are assumed to be random variables. Often they are modeled as exponential distributions, though realistic (observed) distributions tend to be better modeled with log-normal or Γ distributions. The evolution of the fractions of a population in these categories are governed by a system of ordinary differential equations [12], where the effect of pathogenic transmissibility and mixing in a population in determining the spread are "lumped together" in a basic reproductive number R_0. These models are generally applicable in large epidemics where homogeneous mixing of a population can be safely assumed. Network models(e.g. [89]; also see Section 2.3), on the other hand, consist of individual members of a population linked into social network which reflects their mixing pattern; transmission occurs along the links in a probabilistic manner. Individual-based models [25] are even more detailed, involving pathogen-load dynamics inside each individual and transmission of the disease via contamination of shared locations.

In this section, we will formulate a statistical inverse problem to infer the transmissibility of a pathogen as well as the mixing pattern. We will adopt a network model of the epidemic, since it is the simplest one that explicitly separates transmissibility from mixing. This formulation is based on the one in [60], but extends it in the following ways:

1. We will use SEIR models, instead of SIR.
2. We will model the disease realistically, using Γ distributions for the incubation and removal periods (i.e. the exposed and infectious states respectively, rather than exponential distribu-

tions.

3. We will incorporate a structured population.

Below, we describe the context from which the data (observations of an epidemic) were extracted. Thereafter, we formulate the inverse problem.

5.1 The data: The Abakaliki smallpox outbreak of 1967

Between April and June, 1967, a smallpox outbreak with 32 cases occurred in the town of Abakaliki (pop. 31,200) in Nigeria [90]. 30 of the 32 cases belonged to the Faith Tabernacle Church (FTC), a religious group that refused vaccination and medical treatment. The FTC members (a total of 74) lived in 7 large "compounds", often with many non-FTC members (total over 7 compounds: 92); however, social contacts between them were rare. The FTC members were closely related and visited each other; four times a week they gathered at the church. The index case (Case 0) in Abakaliki came to live in Compound 1 on April 2nd, 1967; by April 5th, (Day 0) she had developed a macular rash. Thereafter, smallpox spread in Compound 1. On Day 25, a family seemingly free of smallpox moved to Compound 2, where on day 26 and 30, the children developed clinical signs of smallpox. Thereafter, smallpox spread in Compound 2. The last case occurred on June 20, Day 76. No medical interventions were instituted till Case 11; thereafter, smallpox cases were somewhat tardily isolated at the hospital. This was the only concession FTC members made to the health authorities; they steadfastly refused to be vaccinated. Many of the FTC members were vaccinated (35 out of 74), but the smallpox cases generally were not. Only the dates of appearance of symptoms were recorded and the fate of the smallpox cases are unknown (no deaths were reported (!)). A tabulation of the dates of appearance of symptoms, as well as the possible duration of the contagious period (since we do not know of the cases' fates) are available in [2]. The 2 non-FTC smallpox cases were associated with FTC members; one (Case 20) operated a booth in the market opposite Case 1, while the second, Case 27, washed clothes for the people in Compound 1.

To test our technique, we will slightly idealize the problem. We will ignore the non-FTC population (including the 2 cases) and treat the FTC community as a closed population. We will assume that a social network, modeled as a binomial random graph with a link probability of p_{in}, exists within each compound; cross-compound social links exist, also as a binomial random graph, with a link probability of p_{out}. The spread of infection along a social link is modeled as a Poisson process, with rates β_{in}, β_{out} for in-compound and out-of-compound social links respectively. Each of 30 cases have unknown dates of infection and removal $I_i, R_i, i = 1 \ldots 30$, while their dates of exhibition of symptoms S_i forms the evidence on which the inverse problem is predicated. The other (unobserved) parameters of this stochastic epidemic model are the social network $\mathcal{G}$ and the infection pathway $\mathcal{P}$. We model the behavior of smallpox as "latent" period (non-contagious) which corresponds to the sum of the incubation and prodromal periods and the contagious period. The incubation, prodromal and contagious/symptomatic periods are assumed to be Γ distributed, with the means and standard deviations in Table 8; these were obtained from [2].

Table 8. Means and standard deviations of the incubation, prodromal and contagious/symptomatic periods for smallpox. These were obtained from [2].

Disease state	Mean	Std. Dev.
Incubation period, t_I	11.6	1.9
Prodromal period, t_P	2.49	0.88
Contagious period, t_C	16.0	2.83

5.2 Formulation of the inverse problem

Let $\mathbf{I}$ be the set of infection dates I_i for the 30 cases. Let $\mathbf{R}$ be the set of dates of removal and $\mathbf{S}$, the dates when the cases showed symptoms (the evidence). Then, if two cases j and k were known to have a social link between them, then the likelihood that j showed symptoms on S_j, conditioned on the infection date I_k and transmission rate β_{jk} is

$$\mathcal{L}^{(A)}_{(j,k)}(S_j|\beta_{jk},I_k) = \beta_{jk}\exp(-\beta_{jk}(I_k - S_j))$$

where $\beta_{jk} = \beta_{in}$ or β_{out} depending upon whether j and k belong to the same of different compound. Thus the likelihood of observing $\mathbf{S}$, given the infection pathway (the directed graph of links $j \to k$ over which the disease has been transmitted) $\mathcal{P}$

$$\mathcal{L}^{(A)}(\mathbf{S}|\mathbf{I},\vec{\beta},\mathcal{P}) = \prod_{(j,k)\in\mathcal{P}} \exp(-\beta_{jk}(I_k - S_j)) \tag{10}$$

where $\vec{\beta} = \{\beta_{in},\beta_{out}\}$ and (j,k) denotes the directed link $j \to k$.

$\mathcal{P}$ contains a subset of the links in $\mathcal{G}$, the undirected graph of social contacts, over which disease transmission occurred; the other links in $\mathcal{G}$ did not transmit the disease. This includes links (j,k) between two nodes (people) who never contracted the disease, links where $j \in \mathcal{P}$ but k is not (and thus transmission over (j,k) never occurred) and $j,k \in \mathcal{P}$ but link $(j,k) \notin \mathcal{P}$. Given our Poisson process model of transmission over links, the probability of escaping infection, having been in contact for time τ_{jk}, is $\exp(-\beta_{jk}\tau_{jk})$. Thus the likelihood of observing $\mathbf{S}$ given $\mathbf{I},\mathbf{R},\vec{\beta},\mathcal{P},\mathcal{G}$ is

$$\begin{aligned}\mathcal{L}^{(B)}(\mathbf{S}|\mathbf{I},\mathbf{R},\vec{\beta},\mathcal{P},\mathcal{G}) &= \prod_{(j,k)\in\mathcal{G}\setminus\mathcal{P},\, j\in\mathcal{P}} \exp(-\beta_{jk}\tau_{jk}),\\ \tau_{jk} &= \max(\min(I_k,R_j) - S_j, 0)\end{aligned} \tag{11}$$

where $(j,k) \in \mathcal{G} \setminus \mathcal{P}$ are the set of links that exist in $\mathcal{G}$ but not in $\mathcal{P}$. For j who never contracted the disease, $I_j = \infty$ in Eq. 11.

Let a social network $\mathcal{G}$ contain $\mathbf{n} = \{n_1, n_2, \ldots n_7\}$ in-compound links in the 7 compounds and $\mathbf{m} = \{m_1, m_2, \ldots m_7\}$ out-of-compound links. Then given $\mathbf{p} = \{p_{in}, p_{out}\}$, the probability of observing $\mathcal{G}$ is

$$\mathcal{L}^{(E)}(\mathcal{G}\,|\mathbf{p}) = \prod_{i=1}^{7} p_{in}^{n_i}(1-p_{in})^{C(N_i)-n_i} p_{out}^{m_i/2}(1-p_{out})^{D(N_i,m_i)} \tag{12}$$

where

$$\begin{aligned} C(N_i) &= \left(\begin{array}{c} N_i \\ 2 \end{array}\right), \\ D(N_i, m_i) &= \frac{N_{tot} - N_i)N_i - m_i}{2}, \end{aligned} \tag{13}$$

N_i is the FTC population in Compound i and $N_{tot} = 74$ is the total FTC population.

Since an infection pathway $\mathcal{P}$ is always contained in $\mathcal{G}$, we set the probability of observing $\mathcal{P}$, conditioned on $\mathcal{G}$ i.e $\mathcal{L}^{(F)}(\mathcal{P}\,|\mathcal{G}) = 1$.

The probability of observing a set of symptom time $\mathbf{S}$, conditioned on the set of infection and removal times, is given by

$$\mathcal{L}^{(D)}(\mathbf{S}|\mathbf{I},\mathbf{R}) = \prod_{i=1}^{30} p_I(S_i - I_i)p_R(R_i - S_i) \tag{14}$$

where $p_I(t)$ and the $p_R(t)$ are the PDFs of the latent and contagious periods of smallpox.

Thus the probability of observing $\mathbf{S}$ is

$$\mathcal{L}(\mathbf{S}|\mathbf{I},\mathbf{R},\vec{\beta},\mathbf{p},\mathcal{P},\mathcal{G}) = \mathcal{L}^{(A)}\mathcal{L}^{(B)}\mathcal{L}^{(D)}\mathcal{L}^{(E)}) \tag{15}$$

Using Bayes theorem, we can derive an expression for the joint posterior probability for $\mathbf{I}, \mathbf{R}, \vec{\beta}, \mathbf{p}, \mathcal{P}, \mathcal{G}$ conditioned on $\mathbf{S}$ is

$$p(\mathbf{I},\mathbf{R},\vec{\beta},\mathbf{p},\mathcal{P},\mathcal{G}\,|\mathbf{S}) \propto \mathcal{L}(\mathbf{S}|\mathbf{I},\mathbf{R},\vec{\beta},\mathbf{p},\mathcal{P},\mathcal{G})\pi^I(\mathbf{I})\pi^R(\mathbf{R})\pi^{\beta}(\vec{\beta})\pi^p(\mathbf{p}) \tag{16}$$

where the various $\pi(\ldots)$ are PDFs representing prior beliefs regarding the various parameters.

Given the data (in Section 5.1), it is realistic to assume that $\beta_{in} \geq \beta_{out}$ and $p_{in} \geq p_{out}$. We express these relations as

$$\begin{aligned} \beta_{in} &= (1+r)\beta_{out}, \quad r \geq 0, \beta_{out} \geq 0 \\ p_{out} &= \rho p_{in}, \quad \rho \leq 1, p_{in} \leq 1 \end{aligned} \tag{17}$$

and introduce them into Eq. 16 to derive an expression for the posterior distribution $p(\mathbf{I},\mathbf{R},r,\beta_{out},p_{in},\rho,\mathcal{P},\mathcal{G}\,|\mathbf{S})$ as

$$\begin{aligned} &p(\mathbf{I},\mathbf{R},r,\beta_{out},p_{in},\rho,\mathcal{P},\mathcal{G}\,|\mathbf{S}) \propto \\ &\quad \mathcal{L}\,(\mathbf{S}|\mathbf{I},\mathbf{R},r,\beta_{out},p_{in},\rho,\mathcal{P},\mathcal{G}\,)\pi^{I}(\mathbf{I})\pi^{R}(\mathbf{R})\pi^{r}(r)\pi^{\beta_{out}}(\beta_{out})\pi^{p_{in}}(p_{in})\pi^{\rho}(\rho) \end{aligned} \tag{18}$$

We employ uniform distributions for all the priors. Incubation periods $\mathbf{I}$ are believed to lie uniformly between 0 and 30 days, while the removal period $\mathbf{R}$, between 0 and 40. The prior for r is a $\mathcal{U}(0,40)$, while that for β_{out} is $\mathcal{U}(0,10)$. The priors for both ρ and p_{in} are $\mathcal{U}(0,1)$.

The solution to the inverse problem lies in evaluating $p(\mathbf{I},\mathbf{R},r,\beta_{out},p_{in},\rho,\mathcal{P},\mathcal{G}\,|\mathbf{S})$ in Eq. 18 for various values of the independent variables. Given its high dimension, this is intractable; instead we draw samples from the distribution $p(\mathbf{I},\mathbf{R},r,\beta_{out},p_{in},\rho,\mathcal{P},\mathcal{G}\,|\mathbf{S})$ using a simple Markov chain Monte Carlo method and histogram them to develop PDFs for $\beta_{in},\beta_{out},p_{in}$ and p_{out}.

Given our choice of priors, the posterior distribution in Eq. 18 cannot be easily expressed in terms of canonical distributions and hence we use a Metropolis-Hastings sampler. We use an independence sampler for the duration of incubation and removal periods as well as for $\mathcal{P}$ and $\mathcal{G}$, in a manner similar to [60]. We also carry out the following transformations:

$$\begin{aligned} \xi_{in} &= \log(r), & -\infty < \xi_{in} < \infty \\ \xi_{out} &= \log(\beta_{out}), & -\infty < \xi_{in} < \infty \\ \eta_{in} &= \text{logit}(p_{in}), & -\infty < \eta_{in} < \infty \\ \eta_{out} &= \text{logit}(\rho), & -\infty < \eta_{out} < \infty \end{aligned}$$

and use a normal distribution as a proposal for the remaining parameters $(\vec{\xi},\vec{\eta})$. The problem is thus solved using a single-component MH sampler, in terms of $\mathbf{I},\mathbf{R},\vec{\xi},\vec{\eta},\mathcal{P}$, and $\mathcal{G}$, using a combination of random-walk and independence samplers. The parameters of interest are subsequently reconstructed from them. Note that the priors are expressed in terms distributions of r,β_{out},p_{in} and ρ and are transformed to equivalent distributions in $\vec{\xi}$ and $\vec{\eta}$ in the actual computations.

5.3 Results

The model and the inverse problem described above in Section 5.2 cannot be solved to generate useful information, given the data, as described in Section 5.1; there are simply too many parameters/variables to be inferred from the data. However, the problem can be significantly simplified by making a few fairly realistic assumptions. These assumptions results in different, simpler problems which can, thereafter, be solved to extract useful information from the data. These simpler problems are described below.

Problem I: In this case, we simplify the problem by assuming that we have a fully connected population (i.e. $\rho = 1, p_{in} = 1$). However, in-compound and cross-compound social links differ

in strength and the observed differential rates of spread are accommodated by having two distinct spread rates i.e. β_{in}, β_{out}, on in-compound and cross-compound links respectively. This requires that we estimate r and β_{out} from the data. Further, while there is a unique $\mathcal{G}$ (given full connectivity), there may be many possible $\mathcal{P}$. In this model, we will develop estimates for $\mathbf{I}, \mathbf{R}, \beta_{in}, \beta_{out}$ and the expected value of the disease transmission chain $< \mathcal{P} >$.

In Figure 19, above, we plot the variation of four parameters as the MCMC chain progresses. The MCMC chain was run for 20,000 iterations, with a sample being saved every 10 iterations. It is clear that the last 15,000 iterations (i.e. the last 1,500 saved samples) show proper mixing of the chain and that the samples can be used for drawing inferences. In Figure 19, below, we plot the samples in a $\beta_{in} - \beta_{out}$ space to show the correlation structure between the two spread rates - we see very little of it. In Figure 20 (above), we plot the inferred PDFs for the dates of infection and removal of 3 different Cases. While the base of the PDFs are rather wide (10 days for the dates of infection and 15 for dates of removal), they are roughly symmetric, unlike the skewed Γ-distributions one observes for incubation and removal periods i.e. the PDFs are informative. However, both the dates of infection and removal are largely nuisance variables and the specificity with which we can infer a value for them is not very important. In Figure 20 (below), we plot the PDFs for both β_{in} and β_{out} (lines with symbols). The MAP (maximum a posteriori) estimate of β_{in} is roughly four times larger that the MAP value for β_{out}, corroborating the faster spread of smallpox inside a compound that was qualitatively observed in the data. However, somewhat non-intuitively, the PDF for β_{in} is much wider than that of β_{out} - given the preponderance on in-compound transmission (as indicated by the larger MAP value of β_{in}), one would have expected the credibility interval on the value of β_{in} to be much narrower. This issue is being investigated further. In Figure 21, we plot the expected infection pathway $< \mathcal{P} >$ as a directed graph. The 30 Cases (the nodes in the graph) are colored per their compound affiliations. The color of the links indicates the probability of their existence - all links with a probability of 0.3 or higher are in red and those between 0.1 and 0.3 are in blue. We see a clear transmission of the disease from the index case (Node_000) to other members of her compound (most of the links originating from Node_000 are red and connect to other members of her compound); further most red links exists between nodes of the same color, indicating cohabitation in the same compound. Cross-compound links are generally blue, indicating less certainty regarding their existence; however, this lack of certainty is compensated by their larger number, which enable inter-compound transmission. Also, higher numbered nodes (which were infected and showed symptoms later) have a larger number of blue links connected to them (e.g. Node_000 has only one, while Node_029 has four), indicating an inability to "pin down" the source of their infection as well as the identity of the individuals they infected. This is expected - as the number of infected individuals increased and became a substantial fraction of the total population of 74, the infection mechanism approached those that would be observed in a homogeneously mixed population i.e. where one has an equal probability of being infected by any contagious individual.

The results presented above were obtained with data collected during the entire 90-day duration of the epidemic. We now consider partial observations. By Day 40 of the outbreak, the disease had spread outside Compound 1 and thus there was some (slight) evidence of inter-compound transmission. Using the data collected by Day 40, we infer the same parameters. In Figure 20(below), we plot the PDFs for β_{in}, β_{out} so inferred using lines without symbols. While the widths of the PDFs are about the same, the MAP value of β_{in} is about a third smaller, as is the MAP value for

β_{out}. The cause for this underestimation is being investigated, but may be due to errors in the data - while the epidemic started in early April, 1967, it was reported to the WHO (which gathered the data) in the later part of May [90]. Further, the date of appearance of symptoms were obtained by interviewing the families of the Cases rather than by any documentary proof - thus, there was ample scope for the introduction of "measurement" errors due to faulty memories. In Figure 22, we plot the expected infection pathway, as obtained from the first 40 days of data.

Problem II: In this case, we investigate a partially connected society. We assume, realistically, that the rate of spread along a social link is the same i.e $\beta_{in} = \beta_{out}$ (or alternatively, $r = 0$), and the slower spread of the disease across compounds is due to the paucity of "strong" social links across compounds. We assume that in-compound mixing is strong - we have a fully connected social network inside a compound (i.e. $p_{in} = 1$) and a sparse social network exists between individuals across compounds (i.e. $\rho < 1$). In this case we infer β_{out} and $p_{out} = \rho$ as PDFs. We also generate samples of $\mathcal{P}$ as in the previous case, as well as $\mathcal{G}$. We perform the analysis with the first 40 days of data. For the purposes of the MCMC chain, the prior on ρ was $\mathcal{U}(0.25, 0.75)$; this was done to rule out a fully connected social network.

We performed an examination of MCMC chain, in a manner similar to Problem I, to ensure that the chain was mixing and the samples could be used for analysis. As above, the first 500 saved samples were discarded, to account for the "burn-in" period. In Figure 23(above), we plot the PDFs for the dates of infection and recovery of Cases 5 and 10 (Case 30 had not showed symptoms in the first 40 days). Comparing these PDFs with those obtained in Problem I, we see that the shape of the PDFs is similar, and the MAP values are for Problem II are shifted to the right by about 1 day (Problem II infers that infections and removals happened a day later, compared to Problem I). Thus the two models show very little difference in the values of the infection and removal dates. In Figure 23(below), we plot the PDF for $\beta_{in} = \beta_{out}$. Comparing with Figure 20(below), we see that the width of the PDF for Problem II is similar to that for β_{in} in Problem I, while the MAP value lies in between the MAP values for β_{in} and β_{out}, though closer to β_{in} if one considers the median value ($3.37 \times 10^{-3}\text{day}^{-1}$, as opposed to $4.76 \times 10^{-3}\text{day}^{-1}$ for Problem I). This is intuitively correct, since the bulk of the transmission was in-compound. In Figure 24 we plot the expected infection pathway $< \mathcal{P} >$. Comparing to Figure 22, we see that the two are very similar, but not identical (for example, Node_008 has four in-links in Problem II, but 3 in Problem I). Again, transmission links can be inferred with high probability only for the first few Cases; the links connected to the latter Cases are generally blue. Thus the inferences drawn by the two models are quite comparable. In Figure 25 we plot the expected social network $< \mathcal{G} >$. As in Problem I, the 74 nodes (individuals in the entire population) are colored by their compound affiliations. Only cross-compound links which have appeared in 50% (or higher) of the samples are plotted. A subset of these links, which were instrumental in disease transmission, appear in Figure 24.

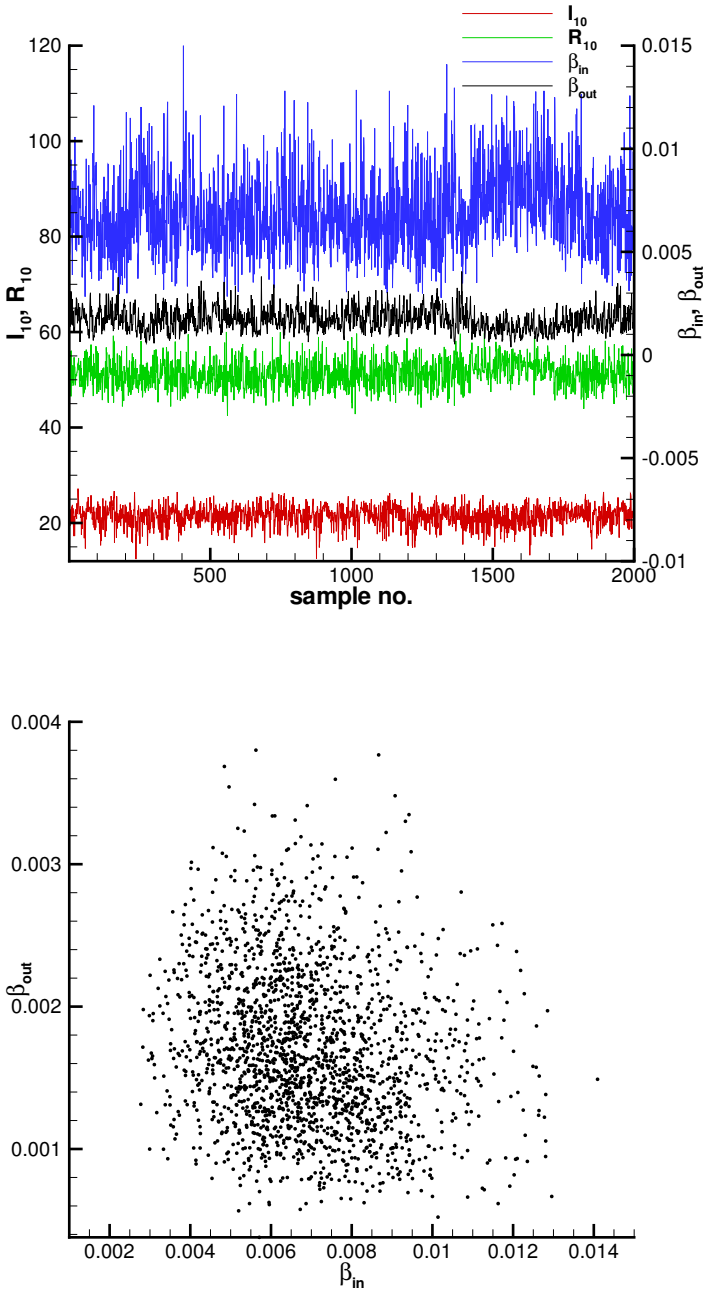

Figure 19. Above: Samples of $I_{10}, R_{10}, \beta_{in}$ and β_{out} for Problem I, as the MCMC chain progressed. We see little autocorrelation and the chains are seen to be mixing well. Below: Scatter plot of samples of β_{in} and β_{out}. We see very little correlation between the two.

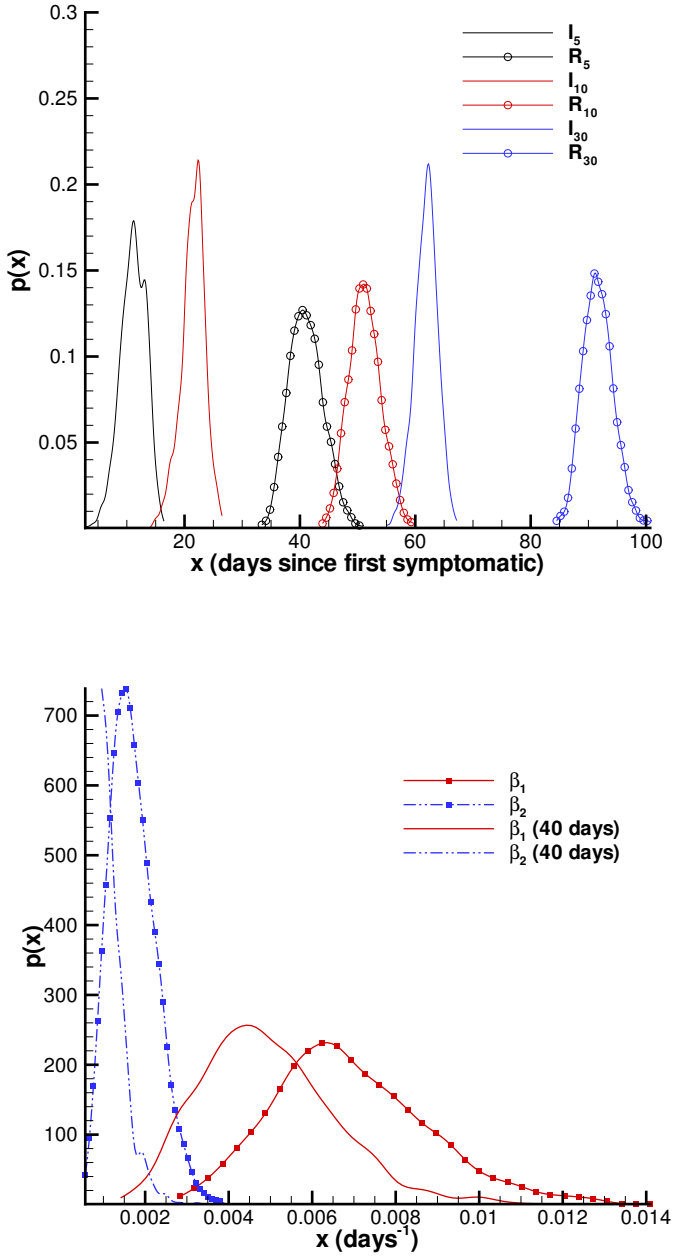

Figure 20. Above: PDFs for the dates of infection and removal for Cases number 5, 10 and 30 for Problem I. The Cases are denoted by separate colors; plots of removal dates contain a symbol. While the bases of the PDFs are almost as wide as those of the incubation and removal periods, the shapes of the PDFs are quite different from the skewed Γ-distributions which are used to model incubation and removal durations for smallpox. Below: The PDFs for the rates of spread β_{in} and β_{out} developed from data collected during the entire epidemic (plots with symbols) and from the first 40 days of data (plots without symbols). The MAP estimate of the spread rates drawn from the first 40 days may be affected by measurement error in the data.

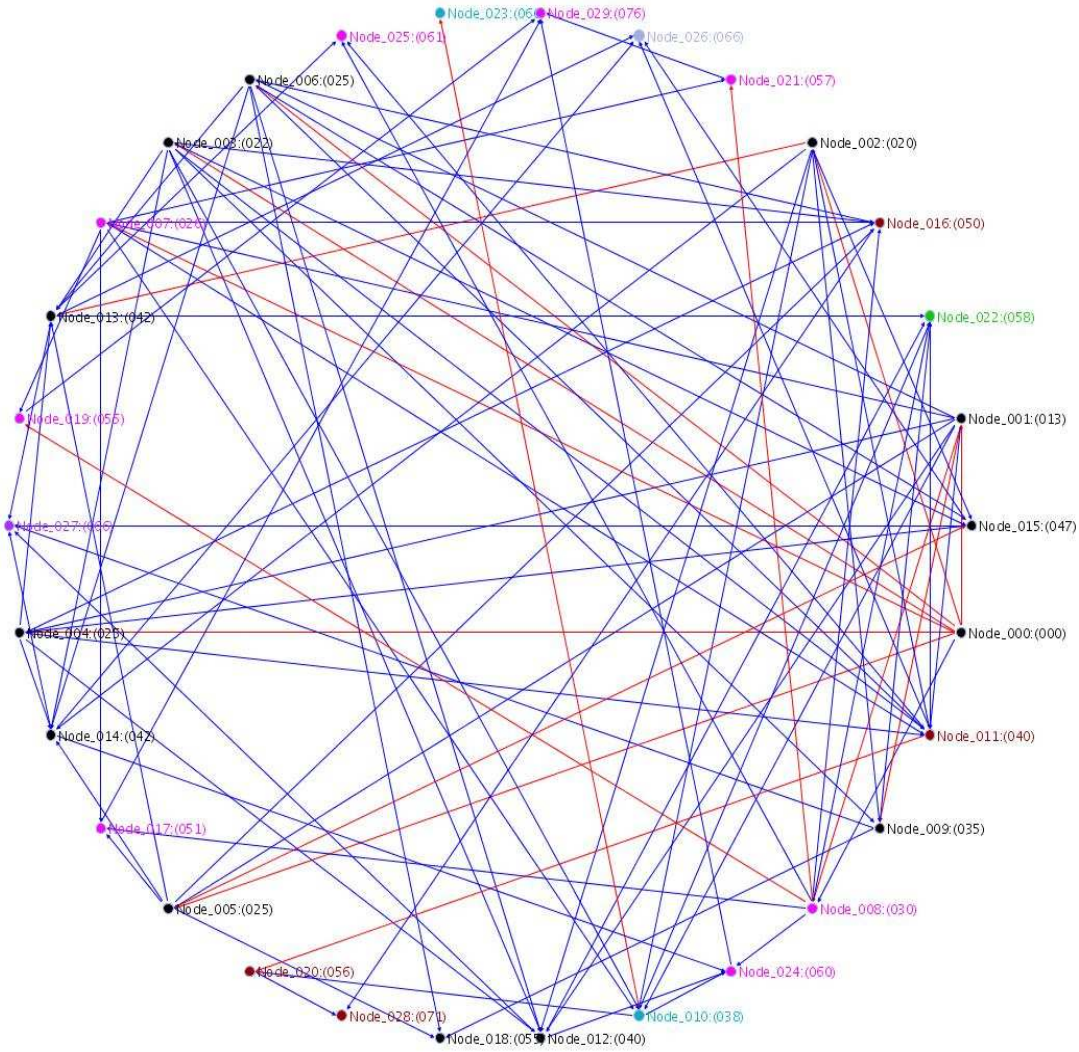

Figure 21. The expected infection pathway $< \mathcal{P} >$, drawn from data collected during the entire epidemic for Problem I. Nodes represent the 30 Cases of the outbreak and are colored by their compound affiliations. The links in the graphs are colored by their probability of existence - links with probability of 30% or higher are in red while those between 10% and 30% are in blue. Most of the transmission from the index case (Node_000) are in red, and connect individuals in the same compound. Infection transmission between the later Cases are almost completely in blue, indicating the reduction of heterogeneity in the transmission mechanism as a large fraction of the population became infected.

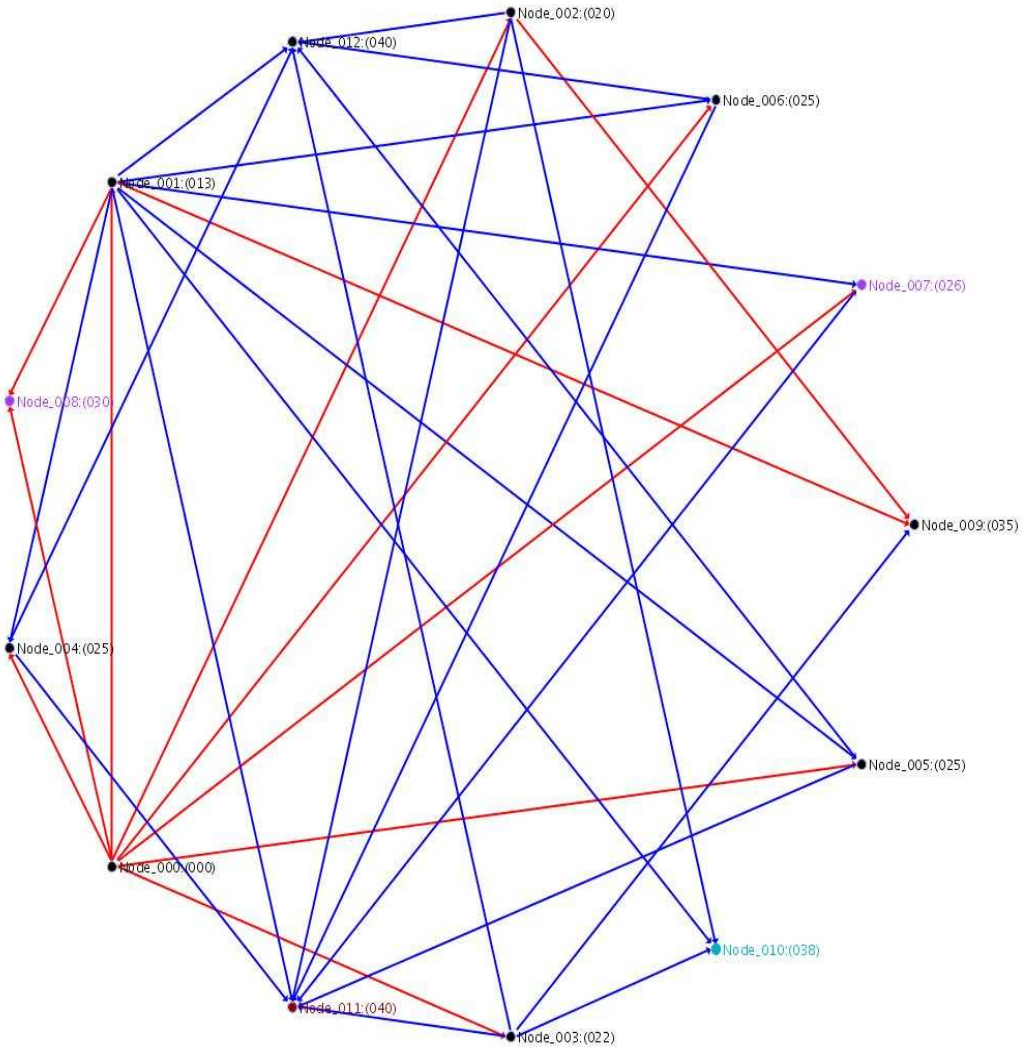

Figure 22. The expected infection pathway $< \mathcal{P} >$, drawn from data collected during the first 40 days of the epidemic, for Problem I. Nodes represent the 30 Cases of the outbreak and are colored by their compound affiliations. The links in the graphs are colored by their probability of existence - links with probability of 30% or higher are in red while those between 10% and 30% are in blue. Most of the transmission from the index case (Node_000) are in red, and connect individuals in the same compound. A few red cross-compound links are also evident at this point in the epidemic.

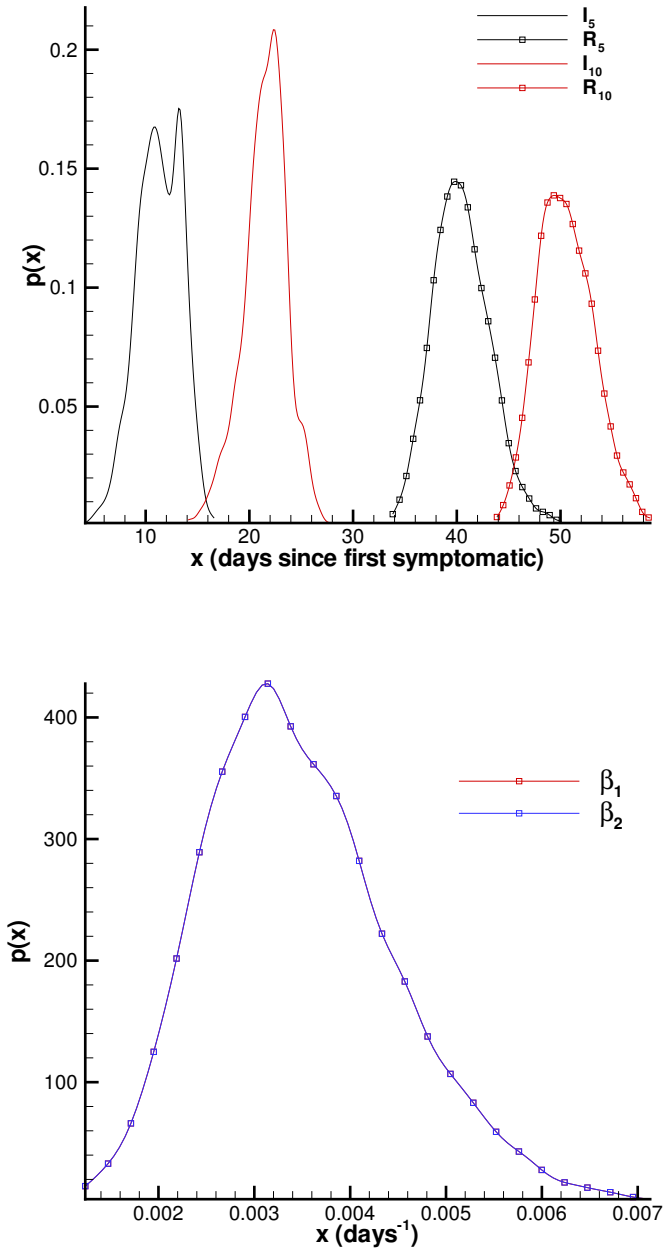

Figure 23. Above: PDFs for the dates of infection and removal for Case numbers 5 and 10, for Problem II. The cases are denoted by separate colors; plots of removal dates contain a symbol. While the bases of the PDFs are almost as wide as those of the incubation and removal periods, the shapes of the PDFs are quite different from the skewed Γ-distributions which are used to model incubation and removal durations for smallpox. Below: The PDFs for the rate of spread $\beta_{in} = \beta_{out}$.

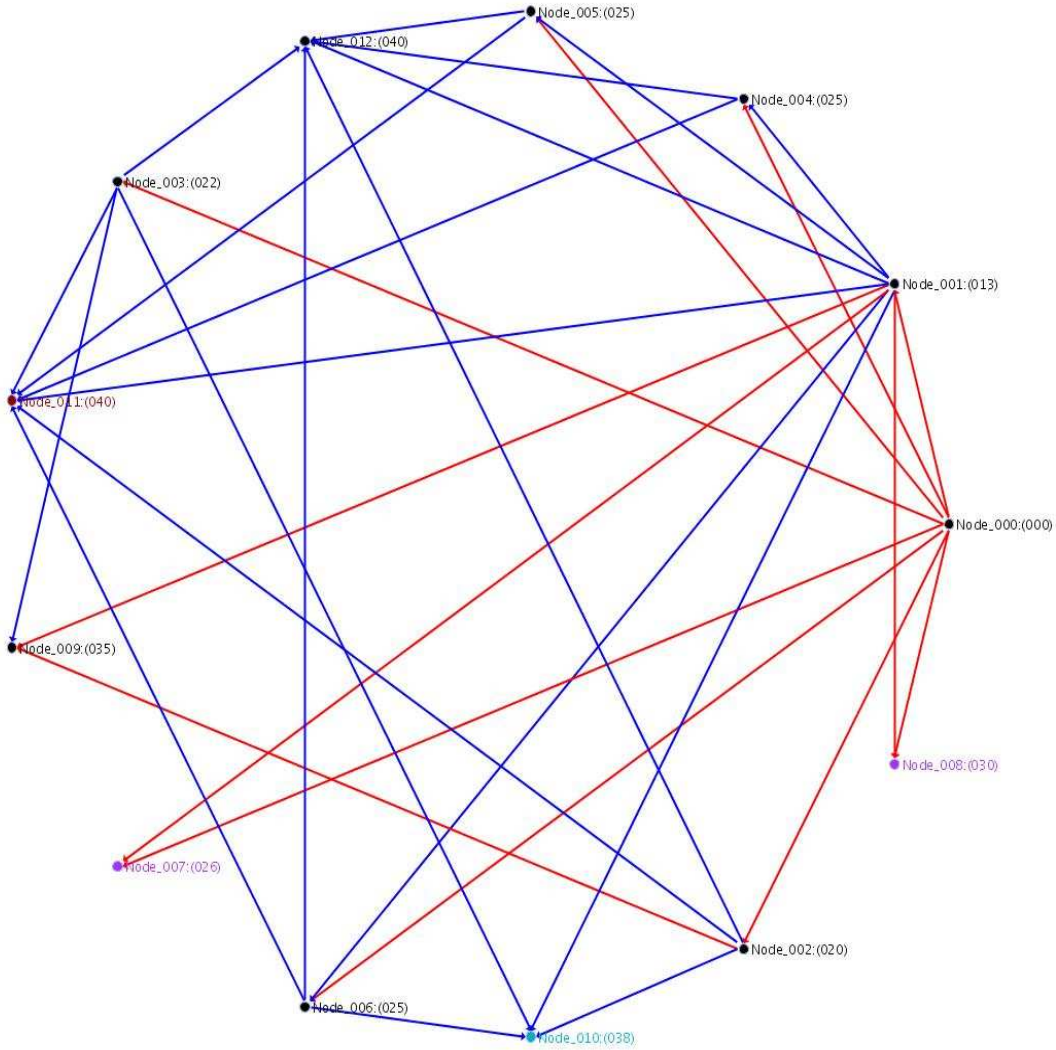

Figure 24. The expected infection pathway $< \mathcal{P} >$, drawn from data collected during the first 40 days of the epidemic, for Problem II. Nodes represent the Cases of the outbreak and are colored by their compound affiliations. The links in the graphs are colored by their probability of existence - links with probability of 30% or higher are in red while those between 10% and 30% are in blue. Most of the transmission from the index case (Node_000) are in red, and connect individuals in the same compound. Infection transmission between the later Cases are almost completely in blue, indicating the reduction of heterogeneity in the transmission mechanism as a large fraction of the population became infected.

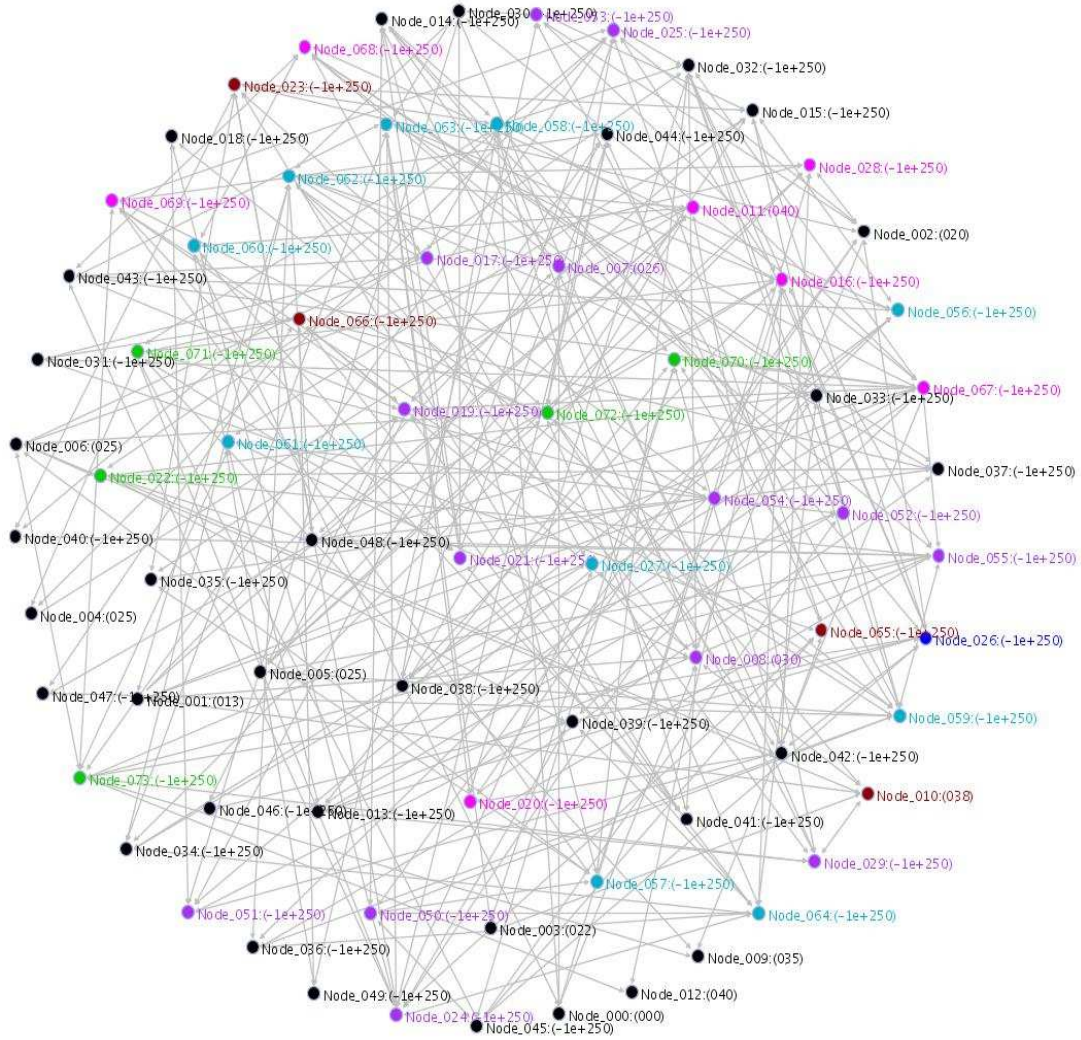

Figure 25. The expected social network $< \mathcal{G} >$, drawn from data collected during the first 40 day, for Problem II. Nodes represent the 74 members of the population and are colored by their compound affiliations. Links in the graphs represent the *cross-compound* social links seen in 50% (or higher) of the samples.

6 Conclusions

Simulations with individual-based models yield detailed predictions of where on the social network disease will spread as well as the transmission chain information and structure necessary to perform source inversion problems. They are potentially valid when the well-mixed assumptions of (deterministic) differential equations are inappropriate. In Section 3 we present smallpox simulation results from our individual-based model, which we designed to be parallel scalable and to investigate the effect of social contact network structures and reduced-order modeling. A static unipartite graph formulation demonstrated comparable performance to a more computationally costly dynamic bipartiite formulation, though with the loss of geographic location information. Static person-to-person contact graphs with basic probability of transmission models still offer individual resolution and contact tracing ability. Results for network sampling (model reduction) indicate that qualitatively similar predictions are possible for both geographic spread and population count epidemic spread. However, appropriately scaling the simulation to the reduced network remains a challenge. Our simulation demonstrates that individual-based models can scale well in parallel, if appropriate communication patterns are used.

We favor the term "individual-based" for our models since, with the exception of choosing when to seek medical care or self-quarantine, disease progression is uninfluenced by autonomous decision, and we include no cognitive models. However, we offer some concluding remarks on agent-based models in general. Agent-based models typically implement local learning, adaptation, and interaction rules, but can exhibit global emergent (complex adaptive) behavior. They are useful for modeling social systems and networks, traffic and transit systems, economies and governments, supply chains and logistics, and even systems of systems, e.g., the nuclear fuel cycle or operational fleets. Agent-based models closely align with structures present in these kinds of systems.

While often intuitive, agent-based models can be complex, and all but the simplest transition models are challenging to analyze. Other criticisms and potential limitations include:

- lack of detailed quantitative information needed to define agent behavior (especially for human decision-making models);
- stochastic nature mandates simulation ensembles to characterize outcomes instead of single calculations;
- computational intensity; and
- not thought of as prescriptive models in the typical engineering sense and there is no clear way to invert to identify parameters.

Perhaps the greatest challenge with agent-based models is performing verification and validation to ensure credibility for their intended use. Verification assesses whether the "right" solution to the mathematical formulation implemented has been found. Code verification can largely (though not entirely) be addressed through good software development practices. Solution verification however presents a problem. There is no clear analog to order of convergence or method of manufactured

solutions. While not rigorous, steps toward verification might include: testing network models by creating carefully structured networks, where transmission paths and incubation rates would be appropriately modeled by an ODE model; or testing cases where analytic or well-established solutions apply. These, however, will not exercise the full extent of an agent-based model (stochasticity, dynamic contacts, emergent behavior, etc.).

Credibility checks on models require data for both individual agent, connectivity, and system-level model validation. This is rare in the context of eradicated or emergent diseases, though there are some exceptions [91]. At least for social networks, data might be gathered from studies in a controlled population, perhaps in observational studies of rumors or other information dissemination. Engineering hierarchical validation approaches may offer some help. Also, methods for validating stochastic simulations are needed: one can only examine output summary statistics or distributions, and must therefore employ appropriate statistical measures.

Applications of social network-based simulations reach far beyond disease modeling; developing verified and validated predictive capability is crucial. Ultimately, practitioners will demand best estimate predictions with quantified uncertainty, especially when model details are not well known.

In Section 4 we have developed a Bayesian approach to characterize bioterror (BT) attacks from a time series of diagnosed patients. Our tests with anthrax show that an observation period of 3–5 days may be sufficient to estimate the number of asymptomatic infected people, the time of infection, and a representative dose, and to provide quantified uncertainty intervals around these estimates; in the absence of an accurate disease model, we may arrive within a factor of two of the size of the attack. The resolution of the time series of diagnosed patients has a small impact if the disease model is accurate; otherwise, model errors dominate.

This Bayesian approach is amenable to many potential extensions and improvements. Informative prior distributions for N and τ, drawn from syndromic surveillance data, may increase the efficiency of the inference process. The ability to "fuse" disparate sources of data via prior distributions contributes significantly to the robustness of Bayesian inference in data-starved environments. One could also incorporate atmospheric transport processes into the likelihood function, thus using the spatial locations of diagnosed patients to guide posterior estimates, though for urban terrains this could lead to very involved computations. Also, the present approach can immediately be applied to other noncontagious diseases, as well as to contagious diseases with long incubation periods, such as smallpox, where secondary cases do not appear in the early time series of patient data.

The Bayesian inference method developed in Section 5 is novel, in the sense it explicitly accounts for the role of pathogenic transmissibility and mixing in the spread of a disease and infers them from the data. These two variables could also be used, independently - the link probabilities inferred in Section 5 (Case II) could easily be used in a network epidemic model for any other disease (e.g., influenza). The inference of the transmissibility, β, can be easily used to track the evolution (of the transmissibility of) emerging infectious diseases; in fact, such efforts have already commenced [51, 52, 53, 54]. Further, the inference of a transmission chain can often be helpful in identifying the mechanism of transmission. For example, in-family transmission is not very strong for Ebola, since it mainly spreads via contaminated bodily fluids, most commonly when preparing a contaminated corpse for final disposal. This is generally done by the women of a village [92],

leading to strong *cross*-family transmission. Such a mechanism would be reflected in both $\vec{\beta}$ and $< \mathcal{P} >$ and would be helpful in detecting the actual mechanism. Thus, while our intention was to apply our approach to counter bioterrorism, there is little to prevent its use with emerging infectious diseases.

The importance of quantitatively characterizing a BT attack was explicitly identified in the "Dark Winter" exercise [8]. Participants sought the ability "...to immediately predict the likely size of the epidemic on the basis of the initial cases; to know how many people were exposed." Thus the primary utility of our inference procedure is within a response plan framework. Response to a BT attack would typically involve confirmatory testing and logistics (the transport of medical materials and personnel), both of which could be better targeted by a quantitative characterization of the attack. By placing the estimated origin of an attack in the context of a transportation network, one could predict the location of future patients and guide prophylactic activities accordingly. The probabilistic characterizations developed here, along with resource hedging for risk-mitigation, support a *measured* approach to addressing BT attacks. In addition to being more sustainable, measured responses may introduce fewer undesirable side effects and be less susceptible to feints.

References

[1] Philip S. Brachman, Arnold F. Kaufmann, and Frederic G. Dalldorf. Industrial inhalational anthrax. *Bacteriological Reviews*, 30(3):646–657, 1966.

[2] Martin Eichner and Klaus Dietz. Transmission potential of smallpox: Estimates based on detailed data from an outbreak. *American Journal of Epidemiology*, 158(2):110–117, 2003.

[3] John A. Jernigan et al. Bioterrorism-related innhalational anthrax: The first 10 cases reported in the United States. *Emerging Infectious Diseases*, 7(6):933–944, 2001.

[4] Jeanne Guillemin. *Biological Weapons: From the Invention of State-Sponsored Programs to Contemporary Bioterrorism*. Columbia University Press, 2006.

[5] P. Williams and D. Wallace. *Unit 731: Japan's Secret Biological Warfare in World War II*. The Free Press, New York, 1989.

[6] Ken Alibek and Stephen Handelman. *Biohazard*. Delta, New York, NY, USA, 2000.

[7] Matthew Meselson, Jeanne Guillemin, Martin Hugh-Jones, Alexander Langmuir, Ilone Popova, Alexis Shelokov, and Olga Yampolskaya. The Sverdlovsk anthrax outbreak of 1979. *Science*, 266:1202–1208, 1994.

[8] Tara O'Toole, Michael Mair, and Thomas V. Inglesby. Shining light on Dark Winter. *Clinical Infectious Diseases*, 34:972–983, 2002.

[9] R. Gani and S. Leach. Transmission potential of smallpox in contemporary populations. *Nature*, 414, December 2001.

[10] P. F. Wehrle, J. Posch, K. H. Richter, and D. A. Henderson. An airborne outbreak of smallpox in a German hospital and its significance with respect to other recent outbreaks in Europe. *Bulletin of the World Health Organization*, 43:669–679, 1970.

[11] F. Fenner, D. A. Henderson, I. Arita, Z. Jezek, and I. D. Ladnyi. *Smallpox and Its Eradication*. World Health Organization, Geneva, 1988.

[12] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: Dynamics and Control*. Oxford University Press, New York, 1992.

[13] Henry C. Tuckwell, Laurent Toubiana, and Jean-Francois Vibert. Spatial epidemic network models with viral dynamics. *Physical Review E*, 57(2), February 1998.

[14] A.L. Lloyd and S. Valeika. Network models in epidemiology: An overview. In B. Blasius, J. Kurths, and L. Stone, editors, *Complex Population Dynamics: Nonlinear Modeling in Ecology, Epidemiology and Genetics*. World Scientific, 2007.

[15] A.L. Lloyd, S. Valeika, and A. Cintron-Arias. Infection dynamics on small-world networks. In A.B. Gumel, C. Castillo-Chavez, and D.P. Clemence, editors, *Modeling the Dynamics of Human Diseases: Emerging Paradigms and Challenges*. AMS, 2006.

[16] R. M. May and A. L. Lloyd. Infection dynamics on scale-free networks. *Physical Review E.*, 64(066112), 2001.

[17] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. *European Physics Journal B*, 26:521–529, 2002.

[18] Marc Barthelemy, Alain Barrat, Romualdo Pastor-Satorras, and Alessandro Vespignani. Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology*, 235(2):275–288, July 2005.

[19] Lauren Ancel Meyers, M.E.J. Newman, and Babak Pourbohloul. Predicting epidemics on directed contact networks. *Journal of Theoretical Biology*, 240(400418), 2006.

[20] Lauren Ancel Meyers, Babak Pourbohloul, M.E.J. Newman, Danuta M. Skowronski, and Robert C. Brunham. Network theory and SARS: predicting outbreak diversity. *Journal of Theoretical Biology*, 232(7181), 2005.

[21] M. E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45(2):167–256, 2003.

[22] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. Villas Boas. Characterization of complex networks: A survey of measurements. *Advances in Physics*, 56(1):167–242, 2007.

[23] Stephen Eubank, V.S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasany, and Nan Wangz. Structural and algorithmic aspects of massive social networks. In *Proceedings of the fifteenth annual ACM-SIAM symposium on discrete algorithms*, pages 718–727, New Orleans, LA, 2004.

[24] Hazhir Rahmandad and John Sterman. Heterogeneity and network structure in the dynamics of diffusion: Comparing agent-based and differential equation models. Technical Report (Working Paper Series) ESD-WP-2004-05, Engineering Systems Division, Massachusetts Institute of Technology, November 2004.

[25] Stephen Eubank, Hasan Guclu, V. S. Anil Kumar, Madhav V. Marathe, Aravind Srinivasan, Zoltan Toroczkai, and Nan Wang. Modelling disease outbreaks in realistic urban social networks. *Nature*, 429:180–184, 2004.

[26] Timothy C. Germann, Kai Kadau, Ira M. Longini, Jr., and Catherine A. Macken. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Science*, 103(15):5935–5940, April 2006.

[27] Donald S. Burke, Joshua M. Epstein, Derek A. T. Cummings, Jon I. Parker, Kenneth C. Cline, Ramesh M. Singa, and Shubha Chakravarty. Individual-based computational modeling of smallpox epidemic control strategies. *Academic Emergency Medicine*, 13:1142–1149, 2006.

[28] Ira M. Longini Jr., M. Elizabeth Halloran, Azhar Nizam, Yang Yang, Shufu Xu, Donald S. Burke, Derek A.T. Cummings, and Joshua M. Epstein. Containing a large bioterrorist smallpox attack: a computer simulation approach. *International. Journal of Infectious Diseases*, 11:98–108, 2007.

[29] Thomas V. Inglesby, Donald A. Henderson, John G. Barteltt, Michael S. Ascher, Edward Eitzen, Arthur M. Friedlander, Jerome Hauer, Joseph McDade, Michael T. Osterholm, Tara O'Toole, Gerald Parker, Trish M. Perl, Philip K. Russel, and Kevin Tomcat. Anthrax as a biological weapon - Medical and public health management. *Journal of the American Medical Association.*, 281(18):1735–1745, 1999.

[30] R. W. Titball, P. C. Turnbull, and R. A. Hutson. The monitoring and detection of *Bacillus anthracis* in the environment. *Journal of Applied Bacteriology*, 70(suppl):9S–18S, 1991.

[31] R. P. Williams. *Bacillus anthracis* and other spore forming bacilli. In A. Braude, L. E. Davis, and J. Fierer, editors, *Infectious Disease and Medical Microbiology*, pages 270–278. W. B. Saunders Co., Philadelphia, PS, 1986.

[32] M. M. Friedlander, S. L. Welkos, M. L. Pitt, J. W. Ezzell, P. L. Worsham, K. J. Rose, B. E. Ivins, J. R. Lowe, G. B Howe, P. Mikesell, and Wade B. Lawrence. Postexposure prophylaxis against experimental inhalation anthrax. *Journal of Infectious Disease*, 167(5):1239–1243, 1993.

[33] J. M. Ross. The pathogenesis of anthrax following the administration of spores by the respiratory route. *Journal of Pathology and Bacteriology*, 73:485–495, 1957.

[34] J. Walden and E. H. Kaplan. Estimating time and size of bioterror attack. *Emerging Infectious Diseases*, 10(7):1202–1205, 2004.

[35] R. Brookmeyer, N. Blades, M. Hugh-Jones, and D. A. Henderson. The statistical analysis of truncated data: application to the Sverdlovsk anthrax outbreak. *Biostatistics*, 2:233–247, 2001.

[36] Ron Brookmeyer and Natalie Blades. Statistical models and bioterrorism: Application to the U.S. anthrax attacks. *Journal of the American Statistical Association*, 98(464):781–788, 2003.

[37] D. W. Henderson, S. Peacock, and F. C. Belton. Observations on the prophylaxis of experimental pulmonary anthrax in the monkey. *Journal of Hygiene*, 54:28–36, 1956.

[38] Wilhelm S. Albrink and Robert J. Goodlow. Experimental inhalational anthrax in the chimpanzee. *The American Journal of Pathology*, 35(5):1055–1065, 1959.

[39] C. A. Gleiser, C. C. Berdjis, H. A. Hartman, and W. S. Gochenour. Pathology of experimental respiratory anthrax in *Macaca Mulatta*. *British Journal of Experimental Pathology*, 44:416–426, 1963.

[40] B. E. Ivins, M. L. M. Pitt, P. F. Fellows, J. W. Farchaus, G. E. Benner, D. M. Waag, S. F. Little, G. W. Anderson, P. H. Gibbs, and A. M. Friedlander. Comparative efficacy of experimental anthrax vaccine candidates against inhalational anthrax in rhesus macaques. *Vaccine*, 16(11/12):1141–1148, 1998.

[41] R. Brookmeyer, E. Johnson, and S. Barry. Modelling the incubation period of anthrax. *Statistics in Medicine*, 24:531–542, 2005.

[42] D. Wilkening. Sverdlovsk revisted: Modeling human inhalational anthrax. *Proceedings of the National Academy of Science*, 103(20):7589–7594, May 2006.

[43] H. N. Glassman. Discussion on industrial inhalational anthrax. *Bacteriological Review*, 30:657–659, 1966.

[44] H. A. Druett, D. W. Henderson, L. Packman, and S. Peacock. Studies on respiratory infection. I.The influence of particle size on respiratory infection with anthrax spores. *Journal of Hygiene*, 51:359–371, 1953.

[45] C. N. Haas. On the risk of mortality to primates exposed to anthrax spores. *Risk Analysis*, 22(2):189–193.

[46] W. R. Hogan, G. Cooper, G. L Wallstrom, and M. Wagner. An Improved Bayesian Aerosol Release Detector. Technical report, The RODS Laboratory, 550 Cellomics Building, 100 Technology Drive, Pittsburgh, PA 15219, 2005.

[47] Fu-chang Tsui, Jeremy U. Espino, Virginia M. Dato, Per H. Gesteland, Judith Hutman, and Michael M. Wagner. Technical Description of RODS : A Real-time public health surveillance system. *Journal of the Americal Medical Informatics Association*, 10(5):399–408, 2003.

[48] D. Bruce Turner. *Workbook of Atmospheric Dispersion Estimates: An Introduction to Dispersion Modeling*. Lewis Publishers, CRC Press LLC, 2000 N.W. Corporate Blvd. Boca Raton, FL 33431, 1994.

[49] A. R. Rao. *Smallpox*. The Kothari Book Depot, Bombay, 1972.

[50] M. Elizabeth Halloran, Ira M. Longini, Azhar Nizam, and Yang Yang. Containing bioterrorist smallpox. *Science*, 298:1428–1432, November 2002.

[51] Luis M. A. Bettencourt and Ruy M. Ribeiro. Real-time Bayesian estimation of the epidemic potential of emerging infectious diseases. *Public Library of Science - One*, 3(5), 2008.

[52] Gerardo Chowell, Hiroshi Nishiura, and Luis M. A. Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society - Interface*, 4, 2007.

[53] Yang Yang, M. Elizabeth Halloran, Johathan D. Sugimota, and Ira M. Longini. Detecting human-to-human transmission of avian influenza A (H5N1). *Emerging Infectious Diseases*, 13(9):1348–1353, 2007.

[54] Michiel van Boven, Marion Koopmans, Mirna Du Ry van Beest Holle, Adam Meijer, Don Klinkenberg, Christl A Donnelly, and Hans (J. A. P.) Heesterbeek. Detecting emerging transmissibility of avian influenza virus in human households. *Public Library of Science, Computational Biology*, 3(7):e145, Jul 2007.

[55] S. Cauchemez amd F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boelle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23:3469–3487, 2004.

[56] Simon Cauchemez, Alain-Jacques Valleron, Pierre-Yves Boelle, Antoine Flahault, and Neill M. Ferguson. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, 452(10):750–755.

[57] Phillip O'Neill, David J Balding, Niels G. Becker, Mervi Eerola, and Denis Mollison. Analyses of infectious disease data from household outbreaks by MCMC methods. *Applied Statistics*, 49(4):517–542.

[58] P. D. O'Neill and G. O. Roberts. Bayesian inference of partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series A*, 162:121–129, 1999.

[59] Michael Hohle, Erik Jorgensen, and Philip D. O'Neill. Inference in disease transmission experiments by using stochastic epidemic models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):349–366, 2005.

[60] Tom Britton and P. O'Neill. Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29:375–390, 2002.

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

[62] G. Chowell, J. M. Hyman, S. Eubank, and C. Castillo-Chavez. Scaling laws for the movement of people between locations in a large city. *Physical Review E*, 68(066102):1–7, 2003.

[63] Tom Britton, Maria Deijfen, and Anders Martin-Lof. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6):1377–1397, 2006.

[64] J. Kaipio and E. Somersalo. *Statistical and Computational Inverse Problems*. Springer, New York, 2004.

[65] D. S. Sivia. *Data Analysis: A Bayesian Tutorial*. Oxford University Press, Inc, New York, 2004.

[66] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, 1996.

[67] C. P. Robert. Mixtures of distributions: inference and estimation. In W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, editors, *Markov Chain Monte Carlo in Practice*, chapter 24, pages 441–464. Chapman & Hall, 1996.

[68] Andrew Gelman, John B. Carlin, Hal S. Stern, and Donal B. Rubin. *Bayesian Data Analysis*. Chapman & Hall / CRC Press, Boca Raton, FL, 2004.

[69] Hakon Tjelmeland and Jo Eidsvik. On the use of local optimizations within Metropolis-Hastings updates. *Journal of the Royal Statistical Society, B*, 66(2):411–427, 2004.

[70] Hakon Tjelmeland and Bjorn Kare Hegstad. Mode jumping proposals in MCMC. *Scandinavian Journal of Statistics*, 28:201–223, 2001.

[71] Cristian Sminchisescu, Max Welling, and Geoffrey Hinton. A mode-hopping MCMC sampler. Technical Report CSRG-478, University of Toronto, 6 King's College Road, Pratt Building, Toronto, Ontario, CANADA, M5S 3G4, 2003. http://www.cs.toronto.edu/ crismin/.

[72] Heikki Haario, Eero Saksman, and Johanna Tamminen. An adaptive Metropolis algorithm. *Bernoulli*, 7(2):223–242, 2001.

[73] Heikki Haario, Eero Saksman, and Johanna Tamminen. Component-wise adaptation for high dimensional MCMC. *Computational Statistics*, 20:265–273, 2005.

[74] J. A. Vrugt, H. V. Gupta, W. Bouten, and S. Sorooshian. A shuffled complex evolution Metropolis algorithm for optimization and uncertainty assessment of hydrologic model parameters. *Water Resources Research*, 39(8), 2003.

[75] J. A. Vrugt, B. Ó. Nualláin, B. A. Robinson, Willem Bouten, Stefan C. Dekker, and Peter M. A. Sloot. Application of parallel computing to stochastic parameter estimation in environmental models. *Computers and Geosciences*, 32:1139–1155, 2006.

[76] Synthetic data products for societal infrastructures and proto-populations: Data set 1.0. Technical Report NDSSL-TR-06-006, Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State University, 1880 Pratt Dr, Building XV, Blacksburg, VA, 24061, 2006. http://ndssl.vbi.vt.edu/Publications/ndssl-tr-06-006.pdf.

[77] Synthetic data products for societal infrastructures and proto-populations: Data set 2.0. Technical Report NDSSL-TR-07-003, Network Dynamics and Simulation Science Laboratory, Virginia Polytechnic Institute and State University, 1880 Pratt Dr, Building XV, Blacksburg, VA, 24061, 2007. http://ndssl.vbi.vt.edu/Publications/ndssl-tr-07-003.pdf.

[78] C. Barrett, R. Beckman, K. Berkbigler, K. Bisset, B. Bush, K. Campbell, S. Eubank, K. Henson, J. Hurford, D. Kubicek, M. Marathe, P. Romero, J. Smith, L. Smith, P. Speckman, P. Stretz, G. Thayer, E. Eeckhout, and M.D. Williams. TRANSIMS: Transportation analysis simulation system. Technical Report Technical Report LA-UR-00-1725, Los Alamos National Laboratory, 2001. Unclassified Report.

[79] K.D. Devine, E.G. Boman, R.T. Heaphy, R.H. Bisseling, and U.V. Catalyurek. Parallel hypergraph partitioning for scientific computing. In *Proc. of 20th International Parallel and Distributed Processing Symposium (IPDPS'06)*. IEEE, 2006.

[80] J. Ray, Y. M. Marzouk, H. N. Najm, M. Kraus, and P. Fast. A Bayesian method for characterizing distributed micro-releases: I. The single-source case for non-contagious diseases. SAND Report SAND2006-1491, Sandia National Laboratories, Livermore, CA 94551-0969, March 2006. Unclassified unlimited release.

[81] Erik Boman, Karen Devine, Lee Ann Fisk, Robert Heaphy, Bruce Hendrickson, Courtenay Vaughan, Umit Catalyurek, Doruk Bozdag, William Mitchell, and James Teresco. *Zoltan 3.0: Parallel Partitioning, Load-balancing, and Data Management Services; User's Guide*. Sandia National Laboratories, Albuquerque, NM, 2007. Tech. Report SAND2007-4748W http://www.cs.sandia.gov/Zoltan/u_html/ug.html.

[82] Jeanne Guillemin. *Anthrax: The Investigation of a Deadly Outbreak*. University of California Press, 2001.

[83] Jeanne Guillemin. The political determinants of delayed diagnosis: The 1979 Sverdlovsk anthrax outbreak and the 1972 Yugoslavian smallpox epidemics. In Rajan Gupta and Mario R. Perez, editors, *Confronting Terrorism – 2002*. Los Alamos National Laboratory, March 25–29 2002. http://library.lanl.gov/cgi-bin/getdoc?event=CT2002&document=20.

[84] G. Peter Lepage. A new algorithm for adaptive multidimensional integration. *Journal of Computational Physics*, 27:192–203, 1978.

[85] GNU Scientific Library. http://www.gnu.org/software/gsl/.

[86] J. Ray, Y. M. Marzouk, M. Kraus, and P. Fast. A Bayesian method for characterizing distributed micro-releases: II. Inference under model uncertainty with short time-series data. SAND Report SAND2006-7568, Sandia National Laboratories, Livermore, CA 94551-0969, December 2006. Unclassified unlimited release.

[87] P. E. Sartwell. The distribution of incubation periods of infectious diseases. *American Journal of Hygiene*, 51:310–318, 1950.

[88] Frederick Klein, Jerry S. Walker, David F. Fitzpatrick, Ralph E. Lincoln, Bill G. Mahlandt, William L. Jones, James P. Dobbs, and Kenneth J. Hendrix. Pathophysiology of anthrax. *Journal of Infectious Diseases*, 116(2):123–138, 1966.

[89] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical Review Letters*, 86(14):3200–3203, 2001.

[90] D. Thompson and William Foege. Faith Tabernacle smallpox epidemic. Technical Report WHO/SE/68.3, World Health Organization, 1968.

[91] Jonathan B. Tucker and Raymond A. Zilinskas. The 1971 smallpox epidemic in Aralsk, Kazakhstan, and the Soviet biological warfare program. Technical Report Occasional Paper No. 9, Center for Nonproliferation Studies, Monterey Institute of International Studies, Monterey, California, 2002.

[92] G. Chowell, N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. M. Hyman. The basic reproductive number of Ebola and the effects of public health measures: The case of Congo and Uganda. Technical Report LA-UR-03-8189, Los Alamos National Laboratory, Los Alamos, NM, 2003.

A Methodology for obtaining a dose distribution consistent with atmospheric dispersion over a geographically distributed population

The spatial distribution of dosages due to atmospheric release of an aerosol can be modeled using a simple Gaussian plume model [48]. An atmospheric release typically occurs over a domain with a non-uniform population distribution; we can combine the plume model with the population distribution to calculate the number of people exposed to a given dose. In this section, we describe a simple way to obtain such a population-dosage distribution.

We consider a square domain, L km on each side; in this study, $L = 10$ km. The domain is divided into N blocks per side; here $N = 100$. 25 population clusters are chosen in the form of Gaussian kernels $A\exp(-r^2/R^2)$, where $r^2 = |\mathbf{x} - \mathbf{x_0}|^2$. The strength of the kernel A, its center $\mathbf{x_0}$, and its length scale R are randomly sampled from independent uniform distributions. The population density in any block, with its center at $\mathbf{x}$, is a sum of the strengths of all the 25 kernels. The strengths of the kernels are scaled to obtain a total population (in the domain) of P_{domain}. The population in a given block is obtained by multiplying the population density with the block area. This creates a geographically distributed population.

The number of people exposed (i.e., who inhaled the aerosol, but may or may not develop symptoms) and infected (i.e., who will develop symptoms) is dependent on the location and size of the release and direction of the wind. We release 10^{13} spores at the origin, at a height of 100 meters. A wind speed of 4 m/s and a Pasquill stability class of "B" are assumed. Pasquill stability classes indicate atmospheric stability; class B indicates a moderately unstable atmosphere with strong daytime insolation. Details of Pasquill stability classes and atmospheric dispersion are in [48]. In our study, wind directions are measured in degrees from due north; that is, a wind direction of zero degrees is a wind from due north, 90 degrees is a westerly wind, and a direction of 180 degrees is a wind from due south. The release is assumed to be an explosive point release, and the concentration of the aerosol at any point (x,y) on the ground and any time t is given by [48]

$$\chi(x,y,t) = \frac{2Q_T}{(2\pi)^{3/2}\sigma_{x'}\sigma_{y'}\sigma_{z'}} \exp\left(-\frac{(x'-ut)^2}{2\sigma_{x'}^2}\right) \exp\left(-\frac{(y')^2}{2\sigma_{y'}^2}\right) \exp\left(-\frac{(H')^2}{2\sigma_{z'}^2}\right) \tag{19}$$

where (x',y') are Cartesian coordinates in a frame of reference where the x'-axis is aligned with the wind. $\sigma_{x'}, \sigma_{y'}$ and $\sigma_{z'}$ are coefficients dependent on x' and on the Pasquill stability class. H is the height of release and χ is the concentration of the aerosol in spores per unit volume. u is the wind velocity. Q_T is the total number of spores released. The relation between $\mathbf{x}'$ and $\mathbf{x}$ is given by

$$\begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} \cos(\pi-\theta) & -\sin(\pi-\theta) \\ \sin(\pi-\theta) & \cos(\pi-\theta) \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}$$

where θ is the wind direction. Assuming a breathing rate β of 30 liters a minute, one can obtain an

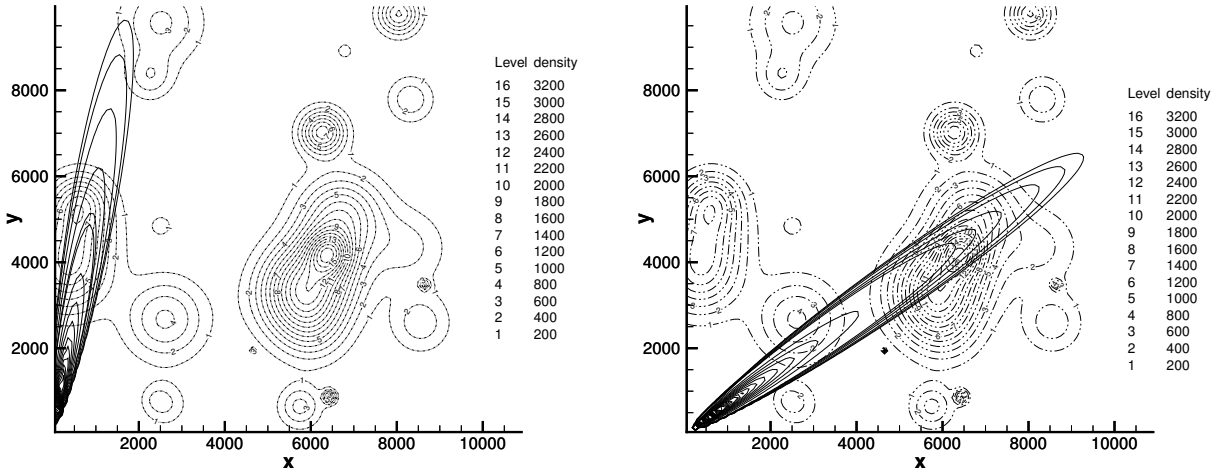

Figure A.1. Dosage plumes plotted over the population distribution for $\theta = 170°$ (left) and $125°$ (right). We see on the right that the extremities of the plume extend into a high population density region. Population density is measured in number of people per square kilometer. Thus we may expect a substantial number of high-dosage cases, resulting in a higher average dosage D.

expression for the number of spores inhaled per unit time. Integrating to infinite time, one obtains the total number of spores D inhaled by a person positioned at (x, y) (or at (x', y')):

$$D = \frac{Q_T \beta}{2\pi\sigma_{x'}\sigma_{y'}\sigma_{z'}} \exp\left(-\frac{(y')^2}{2\sigma_{y'}^2}\right) \exp\left(-\frac{(H')^2}{2\sigma_{z'}^2}\right) \left(1 + \mathrm{erf}(\mathrm{x}')\right).$$

The dosage assigned to a given block is decided by the location of its center. If we choose Model A2 to simulate the BT attack, we use Glassman's formula to model the probability a of showing symptoms (in infinite time) given a dosage D [43]:

$$a(D) = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{\ln(\mathrm{D}/\mathrm{D}_0)}{\mathrm{S}\sqrt{2}}\right)\right] \tag{20}$$

where $D_0 = 8600$ spores and $S = 3.44$. These correspond to a human ID_{50} of 8600 spores and a probit slope of 0.67 [42, 43]. If Model D is chosen instead, we employ Eq. 7 to determine the probability of infection given a dose D. Since the population in a block is known, we can then use the probability of infection to calculate the number of people in the block who will proceed to develop symptoms over time, per the incubation period model.

In this study, we use $P_{\mathrm{domain}} = 3 \times 10^6$ and two plume directions, $\theta = 170°$ and $125°$. The two releases result in, respectively, 686,068 and 1,869,741 *exposed* individuals, i.e., individuals who have received a dose of one spore or more. The maximum doses observed in the two cases are 30,877 and 314,053 respectively. The dose range is divided into 100 equal bins and a histogram

Table A.1. The wind direction, θ, and the size of the exposed population, p_{exposed}, used to generate the infected population in various attacks. For Cases I, Ia, II, and IIa, Eq. 20 is used for the probability of infection, while for Cases III, IIIa, IV, and IVa, Eq. 7 is used.

	$p_{\text{exposed}} = 10^3$	$p_{\text{exposed}} = 10^4$
$\theta = 170°$	Case Ia, Case IIIa	Case I, Case III
$\theta = 125°$	Case II, Case IV	Case IIa, Case IVa

of the number of people in each bin is developed for each of the cases. The histogram is then normalized to obtain the "exposure" PDF, i.e, the PDF of the dose received by an individual in the exposed population. Given the large population ($P_{\text{domain}} = 3 \times 10^6$), the PDF developed from a histogram with 100 bins is quite smooth. Note that only a fraction of the exposed population will develop symptoms, with an individual's probability of being infected (and subsequently developing symptoms) being given by Glassman's relation (Eq. 20) or Eq. 7.

The "exposure" PDFs developed for $\theta = 170°$ and $125°$ are then used to sample from a smaller exposed population of p_{exposed} for each of the tests. Values of p_{exposed} and θ used for the different cases are in Table A.1. Each exposed individual is then allowed to become infected with a dose-dependent probability. The resulting infected sub-population yields the final dose distribution.

Dose distributions resulting from this process, for all the cases (viz. Cases Ia, I, II, IIa, IIIa, III, IVa and IV) are depicted in Figure A.2. We plot the inverse CDF of doses—i.e., the abscissa is the fraction of the infected population which receives a dose less than or equal to the ordinate. In each inset, we also plot a histogram of the dose distribution. Note that while the doses may easily span two orders of magnitude, about 80% of the infected people lie within a one-decade range.

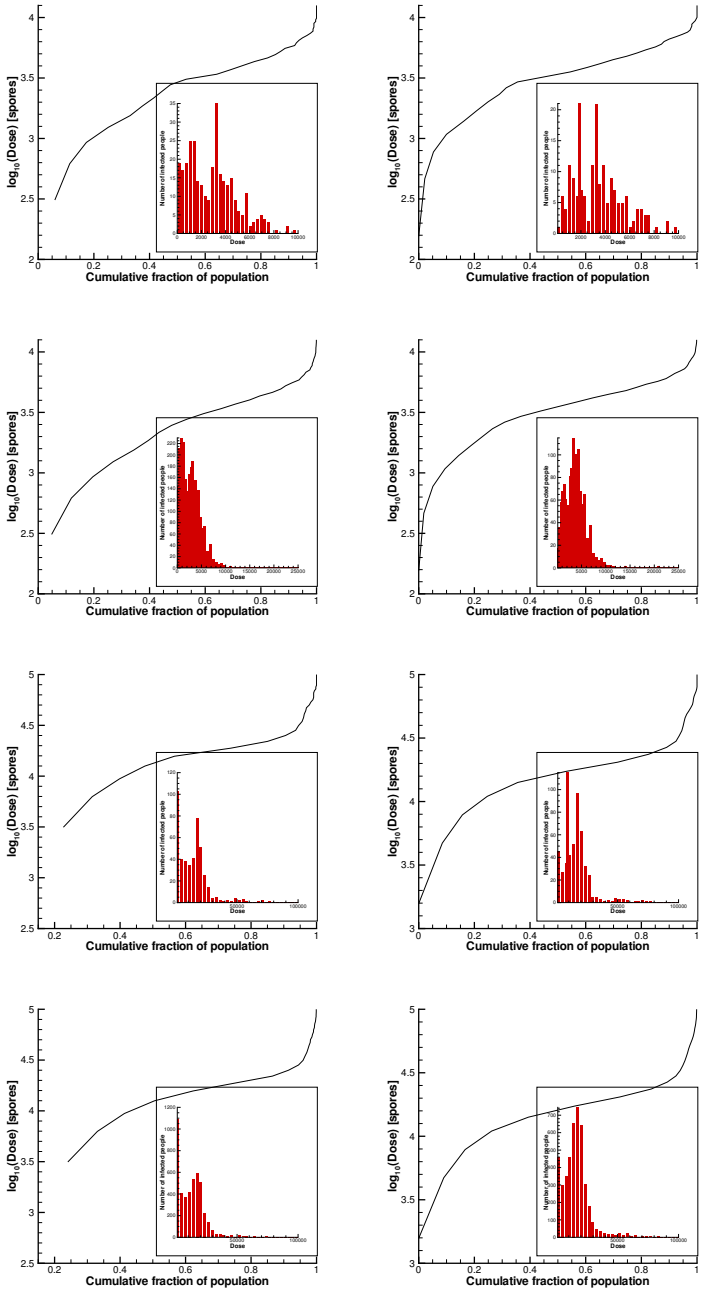

Figure A.2. The inverse cumulative distribution of doses for Cases Ia, I, II, and IIa (left column) and Cases IIIa, III, IV, IVa (right column). The abscissa is the fraction of the infected population which receives a dose less than or equal to the ordinate. Inset: we plot histograms containing the number of infected people in each dose bin. While the histograms have long tails, the bulk of the population receives doses spanning one order of magnitude.

.

Distribution:

1	Michael M. Wolf 4332 Siebel Center, MC-258 201 N. Goodwin Urbana, IL 61801	
1	Youssef Marzouk, 0851	MS 9056
1	Habib Najm, 08351	MS 9056
1	Brian Adams, 01411	MS 1318
1	Karen D. Devine, 01416	MS 1318
2	Jaideep Ray, 08964	MS 9159
1	Donna Chavez, LDRD Office, 01011	MS 0123
2	Central Technical Files 08945-1	MS 9018
1	Technical Library,99536	MS 0899