Bayesian Modeling of Space-Time Properties of Infectious Disease in a College Student Population

[1]Z. Xing, [5]B. Nicholson, [5]M. Jimenez, [3]T. Veldman, [3]L. Hudson, [3,4]J. Lucas, [4]D. Dunson
[2,3]A.K. Zaas, [2,3,5]C.W. Woods, [1,3]G.S. Ginsburg and [1,3]L. Carin

[1]Electrical & Computer Engineering Department
[2]Duke University Medical Center
[3]Duke Institute for Genomic Sciences & Policy
[4]Statistical Sciences Department
Duke University
Durham, NC, USA

[5]Durham Veterans Affairs Medical Center
Durham, NC, USA

Abstract - A Bayesian statistical model is developed for analysis of the time-evolving properties of infectious disease, with a particular focus on viruses. The model employs a latent semi-Markovian state process, and the state-transition statistics are driven by three terms: (i) a general time-evolving trend of the overall population, (ii) a semi-periodic term that accounts for effects caused by the days of the week, and (iii) a regression term that relates the probability of infection to covariates (here, specifically, to the Google Flu Trends data). Computations are performed using Markov Chain Monte Carlo sampling. Results are presented using a novel data set: daily self-reported symptom scores from hundreds of Duke University undergraduate students, collected over three academic years. The illnesses associated with these students are (imperfectly) labeled using real-time (RT) polymerase chain reaction (PCR) testing for several viruses, and gene-expression data were also analyzed. The statistical analysis is performed on the daily, self-reported symptom scores, and the RT PCR and gene-expression data are employed for analysis and interpretation of the model results.

Keywords: Infectious disease, Bayesian, semi-Markov

I. INTRODUCTION

There has been significant interest in the analysis of community-to-individual and individual-to-individual transfer of virus [1], [2], [3], [4], [5], [6]. Many of these studies have been concerned with infection transfer in households [2], [3], [5], in confined spaces like elementary schools [6], as well as transfer among domestic animal populations [4]. The modeling may take different forms, depending upon the

data considered and questions being asked. There has been a significant interest in influenza, and in such studies the interest is typically on influenza-like illness (ILI). For example, based on the incidences of ILI, one may be interested in the analysis of large-scale dynamics of epidemic propagation [1], [7], [8], where in this case the data may be counts of space-time ILI events. There are other studies for which the modeling is performed at or near the level of the symptoms or biomarkers, which are noisy and often imperfect [4]. The studies are complicated by missing and incomplete data, and an unknown number of competing pathogens [6].

Infection dynamics are complex, and therefore the power and flexibility of Bayesian models are attractive [1], [2], [3], [6]; we employ such a modeling approach in this study. Most of these models assume that a given individual is in a particular state of health, such as susceptible (S), exposed (E), infective (I) and recovered (R); an individual in state I is infectious, in that they are capable of transmitting the virus. The sequence of states considered in such a model defines its character, for example susceptible-infective-recovered (SIR) models are widely considered [7], and once in the R state individuals are often assumed removed from the population from the standpoint of infection transfer (because of acquired immunity, or because of death; in some settings R represents "removed" rather than recovered).

When dealing with many competing pathogens [6], such as distinct viruses characteristic of common/typical colds, even after an individual recovers from one virus, they may soon be susceptible to another. That is the case considered in this paper, which motivates a SIS model. We effectively ignore the short time period that may exist between state R and the return to S; during this short time period there may be some cross-immunity between pathogens [6]. Additionally, we do not explicitly model the distinction between states E and I, as these states are not distinguishable with the data considered. With the observed data under consideration, only symptoms allow distinction of states, and therefore we assume state S is one characterized by no or minimal symptoms, and state I is one in which symptoms are observed (there are complications with this definition of state I, as discussed below).

Propagation of influenza and influenza-like viruses has been considered within school settings, for example the Pittsburgh Influenza Prevention Project (PIPP) considered data from ten public elementary schools in the city of Pittsburgh [9]. Studies have also been conducted concerning influenza propagation in families, including data from France from over 300 families [10]. In analyzing such data [6], [2]

state-based models like those discussed above are typically employed, and two forms of dynamics are often considered for the probability of transitioning from S to E, or directly from S to I. One is based upon community-to-person contacts, associated with interactions outside close contacts, and the other is associated with person-to-person transfer among the close contacts. The person-to-person transfer is employed to model interactions between individuals in confined or intimate settings, such as the aforementioned elementary schools or households. In these settings the symptoms themselves are not modeled. Rather, it is assumed that some other mechanism is available to determine an individuals health state, and that this is done separately. For example, in [2] clinical influenza was defined as the presence of fever or feverishness, or at least two of the following signs: sore throat, headache, stiffness or myalgias, fatigue, cough, nasal congestion or rhinorrhea or sneezing. Similarly, in [1], the data modeled were defined cases of illness, with symptoms themselves not modeled.

There are potential pitfalls associated with attempting to model person-to-person transfer, when this mechanism is tied to symptoms. It has been observed that individuals may be infected with a virus but display no symptoms [11]; additionally, for those who do ultimately have symptoms, pathogen transfer may occur in the presymptomatic state [11]. In other words, even though the absence of symptoms from an individual may indicate that she is in state S, she in fact may be in state I, and she may transfer/shed virus. The effectiveness of pathogen transfer from asymptomatic shedders is not well understood. Additionally, the data of interest for person-to-person transfer may be incomplete, in that it only accounts for a subset of close contacts. In [2] the authors modeled infection transfer within elementary schools, but not within the households of the students; in [6] the authors modeled person-to-person transfer within households, but not among other close contacts outside the households. For these reasons, and because of the characteristics of the data considered here (detailed in Section II), we only model community-to-person transfer. This allows for the possible transfer of pathogen from an asymptomatic (but virus shedding) member of the community to a member of our study. However, an asymptomatic member of our study will still be deemed in state S by our model, based on symptoms, even though they may be in state I and asymptomatically shedding virus; this issue is examined in detail when presenting results.

Within the proposed SIS model, we assume that the probability of transferring from state S to state I is time-dependent. Further, we assume that different individuals have distinct degrees of susceptibility to common viruses, and this is modeled as well ($i.e.$, there is a person-dependent character to the degree of susceptibility, and hence to the characteristics of state S). A unique aspect of this study is that the

modeling is performed directly at the level of observed symptoms, rather than using pre-specified means of defining whether one is in the S or I state. Specifically, in most of the above studies the state S/I of the individual was assumed observed, and the goal was to infer the statistics of the state dynamics ($e.g.$, the probability of transiting from S to I, and the duration of being in state I). In this study the symptoms are the observed data, and the state S/I is treated as being *latent*, and to be inferred. As discussed in Section II, we also have access to (imperfect) labels on the health of the individual at a given time, based upon real-time (RT) polymerase chain reaction (PCR) testing and gene-expression data. We compare the model-inferred state of the individual to the state based upon RT PCR and gene expression. This comparison provides insights into such mechanisms as the aforementioned asymptomatic virus shedders, as well as individuals who are symptomatic but not in state I as defined by RT PCR.

The time dependence in the probability of transiting from S to I captures time variation in the viruses present at a given time, as well as time-dependent dynamics of human interaction ($e.g.$, the mixing of a new set of people may increase virus transfer [12], [13]). A unique characteristic of the data considered here is that it is collected daily, for an entire university academic year; further, we have data from three full academic years. By comparison, the data in [2] only existed for 15 days after a household index case. The long time scale, and the daily sampling, introduce interesting phenomena that have not been investigated previously, to our knowledge (in [1] weekly sampling of symptoms was performed). Specifically, how one reports symptoms may be linked to their mood, which may vary with the day of the week. For example, it has been demonstrated that the way in which individuals rate music is linked to the day (and even time) of reporting [14], [15]. Since we are analyzing symptom data, we must consider biases in data reporting that may occur based upon the day of reporting. The degree of missingness tends to also be linked to the day of the week. This therefore motivates employing a semi-periodic, or weekly effect in the probability of transiting from state S to I. This is a novel characteristic of the model developed here, in which we generalize the use of models that employ seasonal terms [16] (here they become weekly, and they are modeling a latent process).

Covariates may be available that can be employed to impact the probability of transiting from state S to I. Given the very long longitudinal length of our data, we may consider new forms of covariates, becoming available from a web-centric world. We consider the Google Flu Trends data [17]. These covariates are constituted by region-specific web searches of words linked to ILI, and specifically here we employ the Google Flu Trends for Durham, NC, the city in which Duke University resides. Other

recent statistical analyses have modeled the space-time properties of such Google data alone [18], [19], but here we are focused on observed symptoms and the time-dependent Google data is a covariate. In [18] the focus is on modeling an epidemic like influenza, and therefore they employ a susceptible-exposed-infected-recovered (SEIR) model; once in the recovered state, an individual is effectively removed from the pandemic dynamics because of immunity. Here we are interested in modeling long-term illness dynamics from common viruses (producing ILI), in addition to influenza, and therefore removing an individual from the population upon recovery is not appropriate (in the data we observe some people with repeated ILI). This may motivate a SEIS model, but for simplicity we consider a SIS model [7].

II. Motivating Data and Questions

A. *Self-reported daily symptom data*

Self-reported symptom-score data were collected from undergraduate students at Duke University, following guidelines specified by the Duke Institutional Review Board (IRB). Data were collected daily during the 2009-2010, 2010-2011 and 2011-2012 academic school years; in each year, data were collected from the beginning of September until May, using a web-based tool. For each of these collection periods, respectively 246, 378 and 242 students participated. The total number of days in which data were recorded were respectively 222, 214, and 227 over each collection period. The 2009-2010 collection period coincided with the novel H1N1 pandemic [20].

The students' reported symptoms were routinely monitored by Duke University health professionals. When a student was deemed – from reported symptoms – to likely be sick with an infectious disease ($e.g.$, virus), the student was contacted and nasal and blood samples were taken. These are termed index cases. Further, each student provided a list of close contacts (other students they interacted with frequently). Blood samples were then collected daily for a week on these close contacts, with the hope that we may observe the transfer of infectious disease (and to analyze that in the context of the blood samples).

B. *Virus identification and gene-expression data*

For the students from whom samples were collected, RT PCR testing was done for a set of viruses. The particular viruses for which a RT PCR test was available were: Rhinovirus, Coxsackie, Echovirus, Coronavirus (229E, HKU1, NL63, OC43), Parainfluenzavirus (1, 2, 3, 4), RSV A/B, Influenza A, Influenza B, Metapneumovirus (A & B), and Adenovirus E & B, and Bocavirus (platform used: Qiagen ResPlex II V2.0). Therefore, *if* one of these viruses was responsible for the student's illness, and *if* the virus was

present in the collected sample, and *if* the RT PCR test worked properly, then the virus type responsible for illness can be detected. However, a negative RT PCR result does not necessarily imply that the student is not sick with a virus, as there may have been a poor sample (in which markers of the virus were not present), a non-tested (within the RT PCR library) virus may have been responsible for illness, and the RT PCR test is itself imperfect. In the context of this study, across the three years, 897 viral etiology results based on RT PCR were constituted.

In addition to the aforementioned RT PCR tests, we also used the available blood samples to perform gene-expression analysis. Let $\boldsymbol{x}_q \in \mathbb{R}^G$ represent the expression data for subject q, for G genes. We performed sparse Bayesian factor analysis on the set of data $\mathbf{X} \in \mathbb{R}^{G \times Q}$, where column q of $\mathbf{X}$ corresponds to $\boldsymbol{x}_q$. Details on the factor analysis method may be found in [21], [22], [23]. In [24], [22], [23] it was demonstrated that one of the factors in such an analysis may be linked to the host response to virus, and in multiple experiments this signature has been found invariant to the particular type of virus studied. Therefore, for the Q samples defining $\mathbf{X}$, we used this factor analysis to define which of these subjects appear to be infected by a virus (by the presence of an elevated form of a specific factor [24], [22], [23]). We emphasize that this test is also imperfect, but it provides more generality than the RT PCR tests, which are constrained to specific types of viruses. In these experiments $Q = 34$ and $G = 22277$.

Based upon the RT PCR and gene-expression tests outlined above, we can (imperfectly) label each of the subjects for whom samples were collected as being sick with a virus or not, at a given point in time; these tests are used to assess the quality of the model developed in Sections III and IV to detect sickness based on the symptom scores alone (with state of health defined by the inferred latent state, S or I). It is important to emphasize that the labels we will use to assess performance are imperfect, in the sense that the RT PCR and gene-expression tests are only testing for the presence of a virus (and even these tests are not perfect). It is possible that an individual may be sick for another reason ($e.g.$, due to allergies or bacteria); in this case the virus-driven labels may indicate that the student is not sick, while the symptoms indicate otherwise (our symptom-based declaration of health or sick may indicate state I, while the virus-driven labels may indicate S). These issues will be revisited when presenting results.

C. *Impact of form of data on the developed model*

This paper is principally directed toward analyzing the self-reported symptom-score data, with a focus on community-to-person transmission of pathogens. In each student dorm (living facility), roughly 10% to 20% of the students participated in the study, and therefore the close contacts are very sparsely sampled. Further, many of the students spend most of their time outside the dorm, and interact infrequently with many members of the same dorm. It was therefore deemed inappropriate to try to model person-to-person pathogen transfer. We also considered developing dorm-dependent models for pathogen transfer, but the dynamics across the different dorms ($e.g.$, fraction of students sick at any given time) did not vary substantially, and therefore it was deemed most appropriate to develop a single community-to-person model for all students who participated in the study.

However, as detailed below, the close contacts constitute a separate set of data (with associated "truth" for the presence/absence of virus), within the context of the imperfections of the RT PCR and gene-expression data. We therefore use these data for model testing.

D. *Questions to be examined in this study*

- We have access to data over three academic years, with one year corresponding to the presence of novel H1N1 virus. During that year there was heightened awareness on campus about non-pharmacological ways to reduce virus transmission, with many highly visible reminders ($e.g.$, students were prominently reminded about hand washing, use of disinfectants, not touching eyes and nose, etc.). Disinfectant soap was widely accessible throughout the campus, at locations in which people congregate. We wish to examine how this heightened awareness affected the time-dependent hazard of community-to-person transmission of pathogens, relative to the other two years of the study, in which pathogen transmission was far less of a focus.
- The students who participated in this study primarily lived on a separate campus dedicated for first-year students. Therefore, most of the students were Freshman, and at the beginning of the academic year most of these students were coming together, and living in close proximity (in dorms), for the first time. We wish to examine the impact of this new mixing of people on the time-dependent hazard of community-to-person transmission of pathogens.
- The Duke University campus resides within the surrounding city of Durham, NC. We wish to examine how the time-dependent hazard of community-to-person transmission of pathogens of Duke students relates to such metrics as Google Flu Trends. Specifically, we wish to examine the extent to which

Google Flu Trends for Durham, NC predicts the hazard of pathogen transmission on the Duke campus.

- We have access to which dorm room each student lived in. Based upon the symptom scores, the model predicts whether each student is infected at a given time. While we do not explicitly model person-to-person transmission within the model (for reasons stated above), we may use model predictions on the state of health to examine whether someone getting infected at time t within a given dorm raises (or lowers) the incidence of infection of other students in the study who lived in the same dorm (and more specifically, on the same dorm floor). For example, an infected neighbor in a dorm may *heighten* awareness of the danger of pathogen transfer, yielding phenomenon like that associated with the exposure to novel H1N1 (heightened awareness, and hence precautions). We examine this issue in detail, as a function of the type of virus associated with each index case (with virus type imperfectly determined via RT PCR).
- We examine and analyze real-world characteristics of daily self-reported symptom data. This includes day-of-the-week dependent phenomenon in the data (weekly, semi-periodic effects), and connections to data missingness.
- We examine the utility of using symptoms alone for classification of the latent state S/I, with comparisons to RT PCR. This is of clinical relevance, as clinicians typically make a diagnosis based directly on symptoms. We also examine the presence of non-symptomatic individuals who are shedding the virus. Further, we examine cases for which symptoms are clear, but extensive RT PCR testing is negative.

III. BASIC MODELING SETUP

A. *Observed symptoms and the latent state of health*

Assume access to self-reported data from N individuals, provided daily over multiple months. The data correspond to the strength of various infectious-disease-related symptoms, reported separately by each of the N students. Eight symptoms are recorded: nasal discharge, nasal congestion, sneezing, cough, malaise, throat discomfort, fever and headache. Each of the eight symptoms is reported on an ordinal scale, from 0 to 4, with 0 being no symptoms, and 4 "maximum" symptoms. Before the study each of the students is instructed on how to connect perceived symptoms to this scale. Nevertheless, there is clearly subjectivity to the mapping from perceived symptoms to ordinal data, and this subjectivity should be accounted for in the statistical analysis. We note that such subjectivity is always present when

individuals report symptom severity to a doctor or nurse.

Let $\boldsymbol{y}_{nt} \in \{0, \ldots, M\}^J$ represent the J symptom scores reported by individual $n \in \{1, \ldots, N\}$ on day t, where for our study $J = 8$ and $M = 4$; we use generalized notation because the basic modeling strategy may be applied to other types of related data. It is assumed that, at a given time, individual n is either in an infective state I or in susceptible state S. When in state S, the student is not currently sick from a virus, and therefore does not display ILI symptoms; however, the student is assumed susceptible to virus infection. When in state S, different individuals may have distinct levels of susceptibility to virus-borne illness, and this is accounted for in the model. We have tied state I to symptoms, as is common [1], [2], [3], [4], [5], [6]. However, there are asymptomatic individuals who shed virus [11] and hence are in the infective state I; these individuals are identified and discussed when presenting results.

We employ an ordinal probit model to link the J observed symptoms to the latent state. Specifically, consider $\boldsymbol{r}_{it} \in \mathbb{R}^J$, drawn conditioned on the latent state as

$$\boldsymbol{r}_{nt} | z_{nt} \sim \mathcal{N}(\boldsymbol{\mu}_{z_{nt}}, \boldsymbol{\Sigma}_{z_{nt}}^{-1}) \tag{1}$$

where $z_{nt} = S$ or $z_{nt} = I$. Let r_{njt} represent the jth component (symptom) of $\boldsymbol{r}_{nt}$, and y_{njt} similarly represent the jth component of $\boldsymbol{y}_{nt}$. The mapping from real r_{njt} to ordinal y_{njt} is manifested via a traditional probit model as

$$y_{njt} = m \quad \text{if} \quad \tau_{j,m-1} < r_{njt} \leq \tau_{j,m} \tag{2}$$

where each $\tau_{j,m} \in \mathbb{R}$, $\tau_{j,m-1} < \tau_{j,m}$, $\tau_{j,-1} = -\infty$, $\tau_{j,0} = 0$, and $\tau_{j,M} = \infty$. We wish to infer $\{\tau_{j,1}, \ldots, \tau_{j,M-1}\}$, with this performed by considering an improper uniform prior on $\tau_{j,1} < \cdots < \tau_{j,M-1}$ [25]. Uniform improper priors result in proper posterior distributions under mild conditions, as detailed in [25], yielding practically useful sufficient conditions that are met in our study. As discussed in the Appendix A, for identifiability purposes the covariance matrices $\boldsymbol{\Sigma}_S^{-1}$ and $\boldsymbol{\Sigma}_I^{-1}$ are restricted to correspond to correlation matrices [26], with diagonal elements all equal to one.

Note that we assume that the statistics of the symptoms for the infective individuals, characterized by $\mathcal{N}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I^{-1})$, are independent of the length of time in which the individual has been in state I. This is a modeling simplification, and one may also link the symptom statistics to the length of time the subject has been in the infective state. The variability in the symptom scores within a given state, characterized

by $\boldsymbol{\mu}_I$ and $\boldsymbol{\Sigma}_I$, account for variability in how a given individual maps perceived symptom strength to ordinal values. Further, $\boldsymbol{\Sigma}_I$ accounts for variability in symptom strength across an extended period of infection (typically from weak, to strong, and back to weak symptoms over the period of infection).

B. Semi-Markov latent-state dynamics

The probability of individual n transiting from a state of susceptibility at time $t-1$, $z_{n,t-1} = S$, to a state of illness/infection at time t, $z_{n,t} = I$, is modeled as

$$p(z_{n,t} = I | z_{n,t-1} = S) = \Phi(\gamma_{nt}) \tag{3}$$

where $\Phi(x) = \int_{-\infty}^{x} d\eta \mathcal{N}(\eta; 0, 1)$ is a cumulative distribution function, with $\mathcal{N}(\eta; 0, 1)$ a normal distribution function for variable η, characterized by zero mean and unit variance (probit transition statistics). The model of the time-evolving variable $\gamma_{nt} \in \mathbb{R}$ is discussed in Section III-C. Related forms of time and covariate dependent probabilities of community-to-person transfer have been considered in [6]; however, in that work, and much of the literature, the state S or I was assumed observed, where here the state is latent and is to be inferred upon the observed symptoms.

If individual n transits from $z_{n,t-1} = S$ to $z_{n,t} = I$, then it is assumed that $z_{n,t+d} = I$ for $0 \leq d \leq D_{nt}$, where D_{nt} is a random variable defining the number of days of infection. We employ the model

$$D_{nt} = c + \hat{D}_{nt} \ , \quad \hat{D}_{nt} \sim \text{Pois}(\lambda_n) \tag{4}$$

where $c > 0$ is a minimum number of days infected, and the rate parameter λ_n is assumed drawn from a gamma distribution. We discuss setting c when presenting experimental results; the imposition of a lower bound c on the number of days of being infected ($i.e.$, in the state I) helps distinguish isolated days when one may not feel well, for various reasons, from actual extended periods of infection.

In [1] the length of time D_{nt} in state I was a real random variable, and was drawn from a gamma distribution. Here we observe discrete temporal data (days), and employ Poisson random variables for the length of time in state S; the lower bound c assures that we are not undermined by draws from Pois(λ_n) that could be equal to zero.

C. *Modeling the time-dependent probability of becoming infected*

The time-evolving parameter γ_{nt}, in concert with a probit link function, defines the probability with which one transits from a susceptible to infective state. We model this time-evolving parameter via four terms:

$$\gamma_{nt} = a_n + \sum_{i=1}^{3} \gamma_t^{(i)} \tag{5}$$

with $\gamma_t^{(1)}$ modeling the general trend within the population to become infected, $\gamma_t^{(2)}$ is associated with periodic (weekly) effects characterizing unique aspects of the day of the week, and $\gamma_t^{(3)}$ is a regression term. Concerning $\gamma_t^{(3)}$, we specifically perform regression to the Google Flu Trends data. Note that $\{\gamma_t^{(i)}\}_{i=1,3}$ are independent of the individual index n, and are therefore shared across the population. The term $a_n \in \mathbb{R}$ is an individual-dependent tendency to get infected, which we place a normal prior on (when a_n is large and positive the nth individual has a heightened susceptibility toward illness, with the opposite true when a_n is negative with large magnitude).

The model in [6] also imposed covariate-dependent state-transition statistics. However, the length of the data considered in that study, and in most of the literature, precluded the need to consider semi-periodic terms. Further, most such models are not performed at the level of symptoms, and therefore they do not have to address semi-periodic missing data phenomenon, and other characteristics of the symptoms.

1) ***General-trend term****:* An autoregressive model is employed for $\gamma_t^{(1)}$:

$$\gamma_t^{(1)} \sim \mathcal{N}(\omega \gamma_{t-1}^{(1)}, \beta^{-1}) \tag{6}$$

where a gamma prior is placed on β and $\omega \in (0, 1)$ is drawn from a truncated normal distribution, $\omega \sim \mathcal{N}_{(0,1)}(\mu_\omega, \sigma_\omega)$. This imposes that the time dependence of the general trend toward illness varies smoothly.

2) ***Weekly or periodic term****:* The observed data are characterized by clear dependencies on the day of the week on which symptoms are reported. The day of the week may impact general feelings of well being (Monday vs. Friday), and certain portions of the week may be characterized by heightened student workload, stress, and lack of sleep/exercise. A seven-day semi-periodic term is therefore employed to model $\gamma_t^{(2)}$. This term builds upon modeling strategies discussed in [16] (Chapter 8.6); for completeness, we here provide some details.

Using notation from [16], the unique terms of a periodic function may be represented in terms of φ_j, for $j = 0, \ldots, p-1$, where p is the period of the weekly/repeating term (for our problem $p = 7$, for the days of the week). Basic Fourier analysis dictates that

$$\varphi_j = a_0 + \sum_{r=1}^{h} [a_r \cos(\alpha r j) + b_r \sin(\alpha r j)] = a_0 + \sum_{r=1}^{h} S_r(j) \tag{7}$$

where h is the largest integer not exceeding $p/2$, $\alpha = 2\pi/p$, and $a_r \in \mathbb{R}$ and $b_r \in \mathbb{R}$ are Fourier components. One may readily demonstrate that

$$S_r(j) = \boldsymbol{e}^T \boldsymbol{\theta}_r(j) \ , \quad \boldsymbol{\theta}_r(j) = \mathbf{J}(\alpha r) \boldsymbol{\theta}_r(j-1) \tag{8}$$

where $\boldsymbol{\theta}_r(0) = (a_r \ b_r)^T$, $\boldsymbol{e} = (1 \ 0)^T$ and $\mathbf{J}(w) = \begin{pmatrix} \cos(w) & \sin(w) \\ -\sin(w) & \cos(w) \end{pmatrix}$.

Generalizing (8) to the stochastic case, and motivated by a dynamic linear model (DLM) [16], the time dependence of the rth Fourier component within a particular time period is modeled as

$$S_r(j) = \boldsymbol{e}^T \boldsymbol{\theta}_r(j) + \nu_r(j) \ , \quad \boldsymbol{\theta}_r(j) = \mathbf{J}(\alpha r) \boldsymbol{\theta}_r(j-1) + \boldsymbol{\epsilon}_r(j) \tag{9}$$

where the components of the two-dimensional vector $\boldsymbol{\epsilon}_r(j)$ are drawn $\boldsymbol{\epsilon}_r(j) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_{\theta r}^{-1})$, and $\nu_r(j) \sim \mathcal{N}(0, \zeta_{Sr}^{-1})$, with a gamma prior placed on ζ_{Sr} and a Wishart prior on $\boldsymbol{\Sigma}_{\theta r}$. The term $\boldsymbol{\epsilon}_r(j)$ models noise in the Fourier components over a given time period, and $\nu_r(j)$ represents measurement noise.

A prior is placed on $\boldsymbol{\theta}_r(0)$, corresponding to the Fourier components over the first week of data, and then (9) is repeated cyclically over the multiple weeks, through sequential draws of $\{\nu_r(j)\}$ and $\{\boldsymbol{\epsilon}_r(j)\}$. Note that with the zero mean priors on $\{\nu_r(j)\}$ and $\{\boldsymbol{\epsilon}_r(j)\}$, conditioned on $\boldsymbol{\theta}_r(0)$, the expectation of (9) corresponds to (8). A zero-mean normal prior is placed on $\boldsymbol{\theta}_r(0)$, for each r. With $S_r(j)$ so drawn, one may superpose the Fourier components to constitute $\gamma_t^{(2)}$; for the weekly data under consideration, there are $h = 3$ Fourier components, in addition to the mean a_0.

3) ***Regression term****:* Assume that we have access to a time-dependent covariate f_t, which in our problem corresponds to the Google Flu Trends [17] data for the region in which the individuals under

study reside. The regression term is modeled as

$$\gamma_t^{(3)} \sim \mathcal{N}(\xi f_t, \alpha_f^{-1}) \tag{10}$$

where a zero-mean normal prior is placed on ξ and a gamma prior is placed on α_f.

IV. ADDITIONAL MODEL CONSIDERATIONS

In the previous section it was assumed that the parameters λ_n and a_n were drawn i.i.d., with the former controlling the length of time individual n tends to be in an infective state, and the latter controlling the tendency of individual n to get infected. The parameter a_n has the impact of controlling the degree to which one is susceptible to virus, and hence to transition from state S to I (large a_n implies higher susceptibility).

It is anticipated that individuals may cluster in terms of their ($e.g.$, genetic or behavioral) tendency to get infected, and in the length with which they stay infected. It is desirable to account for this in the model (it allows sharing of statistical strength between individuals). Additionally, for the dataset that motivates this paper, we have access to the residence location of each student, and therefore it is possible to use this spatial information as a covariate. For example, one may consider the spatial location of each student when modeling the time-dependent tendency to get infected, via including spatial information in $\gamma_t^{(1)}$, for example. Other modeling issues discussed below include consideration of missing data, and the joint modeling of data from multiple years.

A. *Clustering tendency toward infection, and length of infection*

A natural means of clustering λ_n and a_n is to employ a Dirichlet process, with which the number of clusters may be inferred nonparametrically. Specifically, we draw

$$\lambda_n \sim G_\lambda \ , \quad G_\lambda \sim \text{DP}(\alpha_{0\lambda} G_{0\lambda}) \tag{11}$$

$$a_n \sim G_a \ , \quad G_a \sim \text{DP}(\alpha_{0a} G_{0a}) \tag{12}$$

where the base measures $G_{0\lambda}$ and G_{0a} correspond, respectively, to gamma and normal distributions. Gamma priors are placed on the DP parameters $\alpha_{0\lambda}$ and α_{0a}.

B. Spatial covariates

As indicated above, for the motivating data, we have knowledge of the residence location (dorm room) of each individual (student), and therefore it is possible to exploit spatial information when modeling the general trend toward being infected, reflected in $\gamma_t^{(1)}$. One could also consider utilizing spatial information when modeling the weekly (semi-periodic) term $\gamma_t^{(2)}$ and the regression term $\gamma_t^{(3)}$, but spatial dependencies for these terms are less well motivated.

In our numerical experiments, we considered assigning a separate $\gamma_t^{(1)}$ for each floor of a dorm. In this case all students on a given floor shared the same floor-dependent variant of $\gamma_t^{(1)}$ ($i.e.$, rather than sharing a single $\gamma_t^{(1)}$ across all students, a separate such term was employed for each door floor). We also considered assigning a separate term of the form $\gamma_t^{(1)}$ to each dorm ($i.e.$, all residents in a given dorm, independent of floor, shared the same $\gamma_t^{(1)}$). In our experiments, we found that such added modeling complexity did not improve the predictive performance of the model, and in some cases reduced performance (since the students were spatially segregated in these tests, fewer students were associated with a particular floor/dorm-dependent $\gamma_t^{(1)}$, and therefore statistical strength was diffused). There did not appear to be clear situations for which a given dorm or specific dorm floor had a greater tendency toward health (state S) or sickness (state I) than the general population. A potential reason for this is that students spend a significant portion of their time away from their dorm, in classes and other activities, mixing with the general population.

For these reasons, for the results below we do not explicitly leverage spatial covariates for student dorm rooms within the model. However, when presenting results we will examine some of the inferred parameters in the context of student residency location.

C. Missing data

There is a substantial quantity of missing data in self-reported studies, and it is anticipated that the missingness is *not* at random. It is likely that individuals are less likely to pay attention to reporting symptoms when they are feeling well, with greater attention paid during the time of actual illness. If data are missing from individual n on day t, the "observations" are denoted $\boldsymbol{y}_{nt} = \emptyset$. The probability of the null observation in states S is defined as $\eta \in (0, 1)$, and the probability of a null observation in state I is $\rho \in (0, 1)$. We now consider the case of missing data a null observation, and the observation

probability for symptom j, individual n and day t is generalized to [27]

$$y_{njt}|S \sim [\eta\delta_\emptyset + (1-\eta)\sum_{k=0}^{M} p(y_{njt}=k|S)\delta_k] \tag{13}$$

$$y_{njt}|I \sim [\rho\delta_\emptyset + (1-\rho)\sum_{k=0}^{M} p(y_{njt}=k|I)\delta_k] \tag{14}$$

where $p(y_{njt}=k|S)$ and $p(y_{njt}=k|I)$ are the observation probabilities from Section III-A (assuming symptoms are not missing); the symbol δ_k is a unit measure concentrated at the point k. It is assumed that η and ρ are drawn from uniform priors over [0,1].

D. Modeling multiple years of data

The experiments detailed in Sections II and V correspond to (ideally) daily student recording of symptom scores, over an entire academic year; imperfections in this process naturally manifest missing data. Data of this type were collected over three academic years. It is desirable to analyze all of these data jointly, to achieve maximal statistical strength in the results. However, because of the influx of new (freshman) students, and the exit/graduation of others (seniors), the explicit set of students considered on consecutive years is largely distinct. Additionally, each year is characterized (for example) by a distinct respiratory viral illness season, and this must be accounted for when deciding which components of the model to share between multiple years. For example, one of the years during which we collected data corresponded to the presence of an unusual (and potentially dangerous) novel H1N1 flu, which had characteristics ($e.g.$, time of arrival) distinct from typical flu seasons.

So motivated, in the experiments that follow, the explicit $\{\gamma_t^{(1)}, \gamma_t^{(2)}, \gamma_t^{(3)})\}$ are modeled as being distinct among the three years of data. However, the priors on parameters with which these time-dependent functions are constituted are shared across years. Specifically, for the AR(1) model of $\gamma_t^{(1)}$, the priors for ω and β are shared across the multiple years of data. For $\gamma_t^{(2)}$, the parameters $\boldsymbol{\Sigma}_{\theta r}$ and ζ_{Sr} are shared across years, as is the prior on $\boldsymbol{\theta}_r(0)$. Finally, for $\gamma_t^{(3)}$, the priors for ξ and α_f are shared across the multiple years.

Concerning λ_n and a_n, the DP-drawn priors G_λ and G_a are shared across the multiple years, and therefore the clustering of types of people (by susceptibility toward illness, and length of illness) is performed jointly across the multiple years. Finally, concerning the observed symptoms, the parameters $\{\boldsymbol{\mu}_S, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_S, \boldsymbol{\Sigma}_I\}$ are shared across the multiple years, as are the ordinal probit cut points $\{\tau_{j,1}, \ldots, \tau_{j,M-1}\}_{j=1,J}$, and η and ρ (for missing data).

V. RESULTS

The modeling software was implemented in MATLAB™. On a laptop with a 2.7 GHz dual core CPU, each Gibbs iteration takes about 30 seconds, to process all three years of data. We considered 7000 MCMC samples, with the first 2000 discarded as burn-in.

A. *Symptom correlation*

The inferred correlation matrix $\mathbf{\Sigma}_I^{-1}$ for the infective state is shown in Figure 1, where here we present the maximum *a posteriori* MCMC collection sample. As expected, all symptoms are relatively highly correlated within the infective state, with minimum correlation between any two symptoms in excess of 0.65 . Note that nasal discharge and nasal congestion are particularly highly correlated, as are throat discomfort, malaise and cough.

We apply the following approach to help the model distinguish between states S and I, imposing a strong identifiability condition. Based upon the model construction above, the only way states S and I are distinguished is via a requirement that an individual remain in the I state for a minimum of c days. Based on expertise of the infectious disease medical doctors who are co-authors on this study, we set $c = 3$ days, consistent with the minimum length of time one would be anticipated to manifest symptoms due to infectious disease of the type associated with common viruses. In addition, recall from Section II-B that a subset of the subjects within the study were confirmed via RT PCR and/or gene-expression analysis to be ill due to a virus. A small subset of these infective individuals had their data removed from the subsequent analysis, and the correlation between the symptoms of this subset of confirmed cases were used to set the hyperparameters in the prior for $\mathbf{\Sigma}_I$. This setting of the model parameters significantly distinguished the S and I states, yielding interpretable results.

B. *Example student trajectories and inferred diagnoses*

For each individual in the study, as a function of time (day), we infer the probability that the student is in state I ($i.e.$, that they are sick). To demonstrate this, and to give a sense of the reported symptom scores, in Figure 2 self-reported data and the inferred probability of being in state I are depicted for four example individuals; the order of the symptoms (1-8, top to bottom) in Figure 2 is consistent with the order of the symptoms in the correlation matrix of Figure 1. In Figure 2, based upon averaging across all MCMC collection samples, we plot the probability the individual is in the infective state I, for each day.

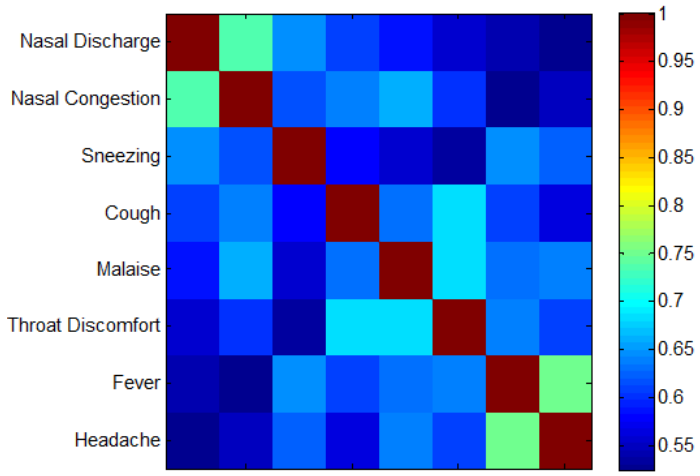

Fig. 1. Inferred correlation matrix for infective state I, $\mathbf{\Sigma}_I^{-1}$, with the approximate MAP solution depicted, corresponding to the maximum *a posterior* collection sample.

The results in Figure 2 are based upon a joint analysis of all self-reported symptom-score data, across all three years.

The inference of the state of health of the subjects in Figure 2 is illustrative of model prediction over all time, the results of which provide *interpretable* values for analysis of infectious disease. However, in a clinical setting one would like to make a prediction about the health of an individual based on all symptoms up to the current point in time (not based on all data, even into the future). We utilize the model for this practical purpose in Section V-F.

C. *Characteristics of missing data*

TABLE I
SUMMARY ON PROPERTIES OF STUDENT REPORTING FREQUENCY AND ASSOCIATED REPORTED SYMPTOM SCORES. THE AVERAGE SYMPTOM SCORE REPORTED IS THE AVERAGE OF THE *sum* OF THE SCORES FOR EIGHT SYMPTOMS.

Missingness	0-20%	20-40%	40-60%	60-80%	80-100%
# Students	22	158	303	178	205
Avg. Symp. Score	2.67	2.22	2.64	3.27	4.33
Avg. Sick Prob.	0.12	0.11	0.15	0.2	0.32

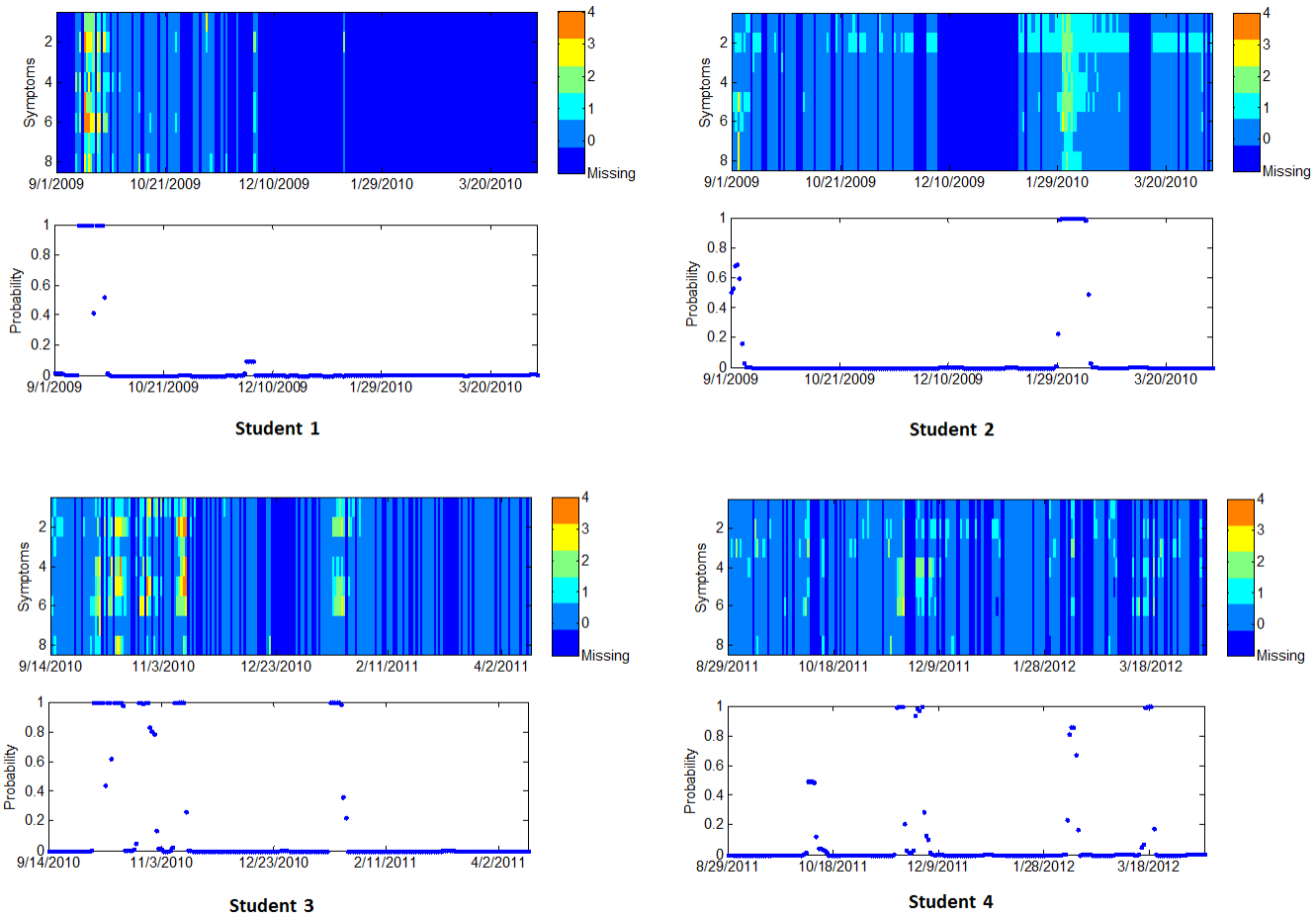

Fig. 2. State of health of four students. For each student, self-reported symptom scores are shown in the top figure. Different colors denote different scores (missing, 0, 1, 2, 3, 4). The probability that a student is in an infective state I at a given time is presented in the bottom subfigure for each of the four students.

In Section IV-C we proposed a model for the missing data. Specifically, it was assume that if a student is in the susceptible state, S, they do not report symptoms (which are likely negligible) with probability η, thereby manifesting missing data. By contrast, when in the infective state I, it is anticipated that one may be more likely to report symptom scores (which are non-negligible, by definition); the probability of not reporting when in state I is represented by ρ (see Section IV-C). Within the analysis, we inferred a mean $\eta = 0.65$, with standard deviation of η equal to 0.01 (reflecting the uncertainty in this parameter from the approximate posterior); the inferred mean for ρ was 0.28, with standard deviation 0.03. Hence, the model infers that when a student is in state S (healthy), a student doesn't report any symptoms 65% of the time, while when in state I (sick) the students don't report symptoms 28% of the time.

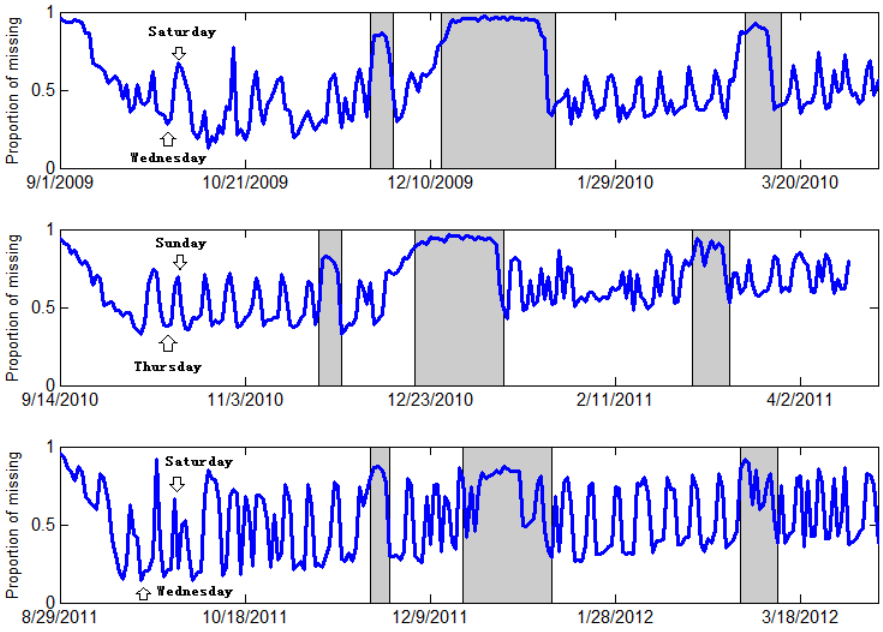

Fig. 3. Fraction of missing data over days. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

In Table I we show data on the characteristics of student reporting and associated symptom scores. In this table is depicted the percentage of days students didn't report data, and the number of such students in each class of missingness. Note that the largest group of students, with 303 members, did not report symptoms on 40-60% of the days. For each class of missingness, we also report the average reported symptom score, recalling that the values were 0 to 4, with 4 the largest/strongest symptom (eight different symptoms are considered). Note that the average symptom score is particularly large for those students who report data infrequently, and the probability that students are in state I when reporting is heightened for the group that rarely reports. The data in Table I motivates the model in Section IV-C, in which the degree of missingness is assumed to be linked to the latent state of health.

In Figure 3 we show the fraction of students who do not report symptoms (fraction of missing data), as a function of day for each of the three years of the study. There is clearly a weekly semi-periodic effect, which has motivated the term $\gamma_t^{(2)}$ in the model. This is discussed further in Section V-G below.

D. Virus infection probability over time

We examine the probability of being in the infective state for students living in proximity to infective individuals. As mentioned in Section II-B, RT PCR test results are available for 897 samples (each from an individual student, and infection case), for the set of viruses discussed in Section II-B. If a positive RT PCR-based virus detection occurred for a given student, that student was deemed to be in an infective state (note that, with RT PCR, most sources of error occur with false negatives, so a positive RT PCR test does have a high chance of actually corresponding to someone infected with a virus – this is discussed further below).

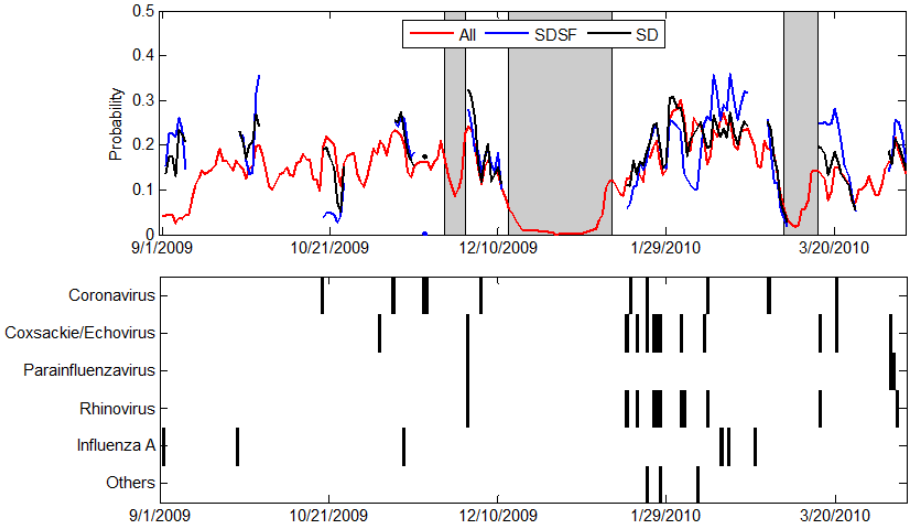

Fig. 4. Top figure: Probability of being in the infective state I on a given day, for academic year 2009-2010. "All" refers to the average across all the students. "SDSF" refers to the average of students living in the same dorm and same floor with infective individuals. "SD" refers to the average of students living in the same dorm with infective individuals. The vertical gray bars represent, from left-to-right, Thanksgiving break, inter-semester (Winter/Christmas) break, and Spring break. Bottom figure: RT PCR test results, black line denotes at that time certain type of virus was detected.

We wish to examine the probability of whether a student is in state I, relative to that student's living conditions with respect to another student who had a positive RT PCR test. Specifically, assume that a given student has a positive RT PCR test. Over a period of a week after that positive RT PCR test, we examine the probability of being in an infective state for all students who shared a dorm with the student confirmed by RT PCR as being in state I. We also examined the probability of being in state I for all students on the same dorm floor (not just the same dorm) of a student confirmed by RT PCR as being infected, again for a week after RT PCR confirmation. When multiple instances overlap in time (multiple

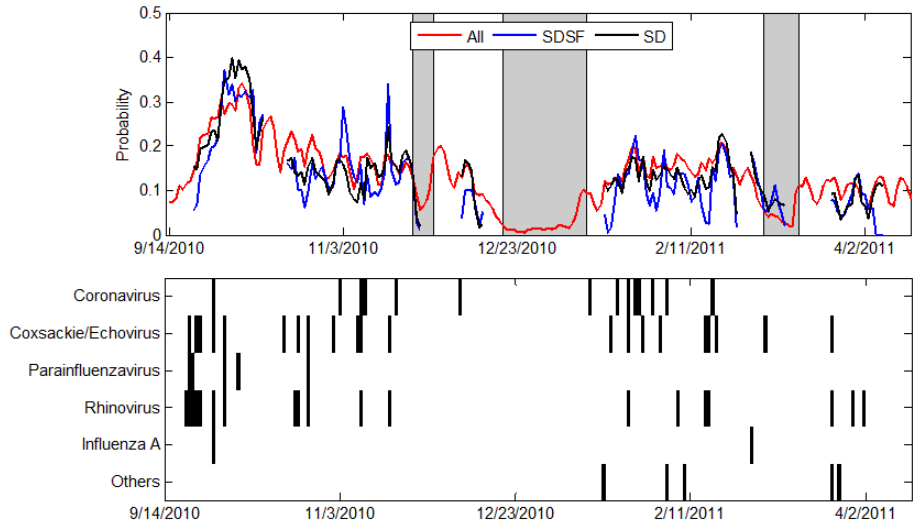

Fig. 5. As in Figure 4, for academic year 2010-2011.

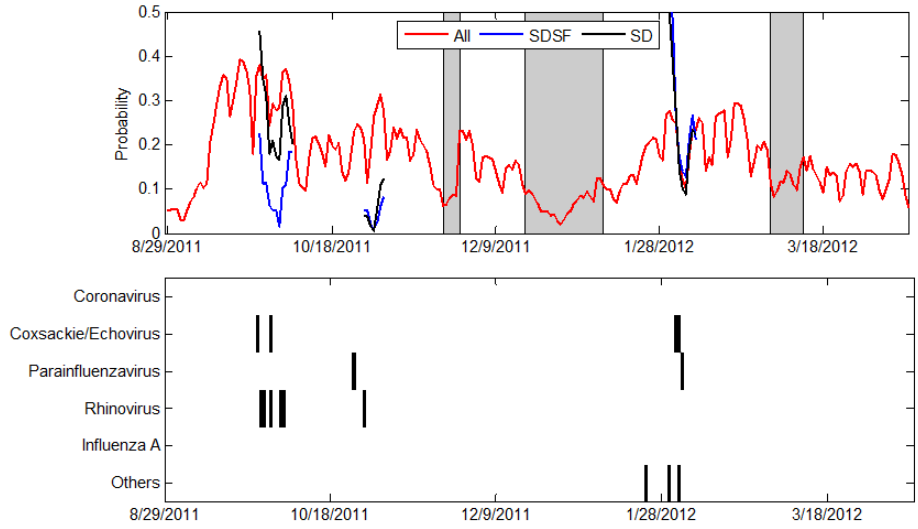

Fig. 6. As in Figure 4, for academic year 2011-2012.

positive RT PCR tests), average results are presented across those multiple instances.

To be precise, let $\mathcal{X}$ represent a particular set of students ($e.g.$, a set of students in the same dorm of a RT PCR-confirmed infective student, or a set of students on the same dorm floor of a RT PCR-confirmed infective student). Let $|\mathcal{X}|$ represent the number of individuals in this set. Then we are interested in computing $S_{\mathcal{X}} = \frac{1}{|\mathcal{X}|} \sum_{n \in \mathcal{X}} p(z_{nt} = I|\mathbf{y}_{nt})$, where $p(z_{nt} = I|\mathbf{y}_{nt})$ is computed from our model. This

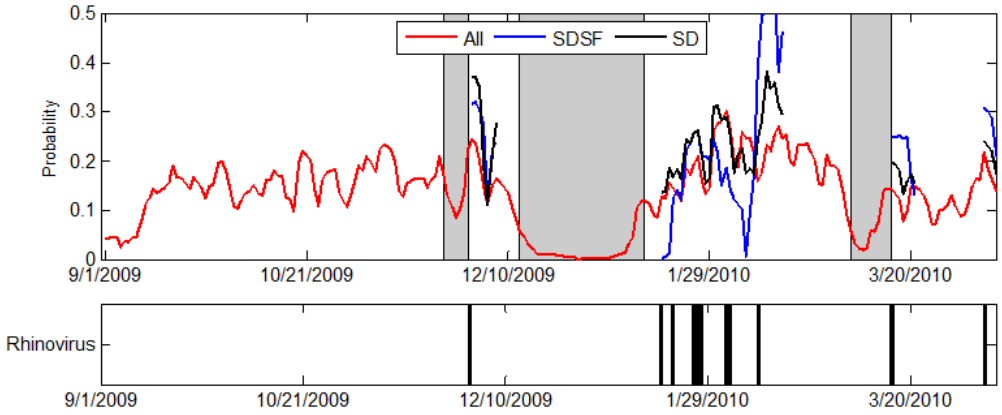

Fig. 7. Top figure: Probability of being in the infective state I on a given day, for academic year 2009-2010. "All" refers to the average across all the students. "SDSF" refers to the average of students living in the same dorm same floor with an infective individual. "SD" refers to the average of students living in the same dorm with an infective individual. The SD and SDSF cases are only for confirmed cases of Rhinovirus. The vertical gray bars represent, from left-to-right, Thanksgiving break, inter-semester (Winter/Christmas) break, and Spring break. Bottom figure: RT PCR test results, black line denotes at that time Rhinovirus was detected.

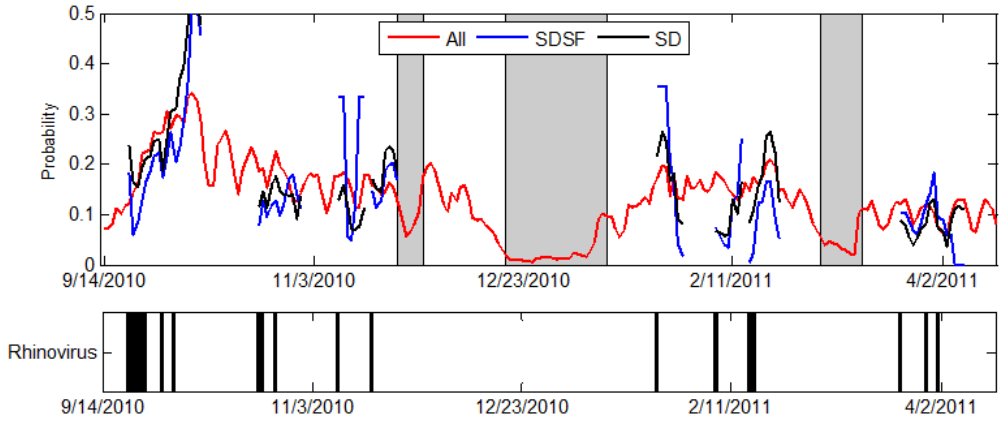

Fig. 8. As in Figure 7, for academic year 2010-2011.

provides a means of examining the inferred degree of enhanced probability of becoming ill with virus, given a nearby confirmed case (recognizing the imperfections in our $p(z_{nt} = I|\mathbf{y}_{nt})$, most notably that one may become sick for other reasons than virus transfer). Such that we have enough individuals in a given set to make this investigation meaningful, we only consider cases for which $|\mathcal{X}| \geq 5$; $e.g.$, when examining propagation of infectious disease on a dorm floor, we only consider cases for which 5 or more students within the study live on the same floor.

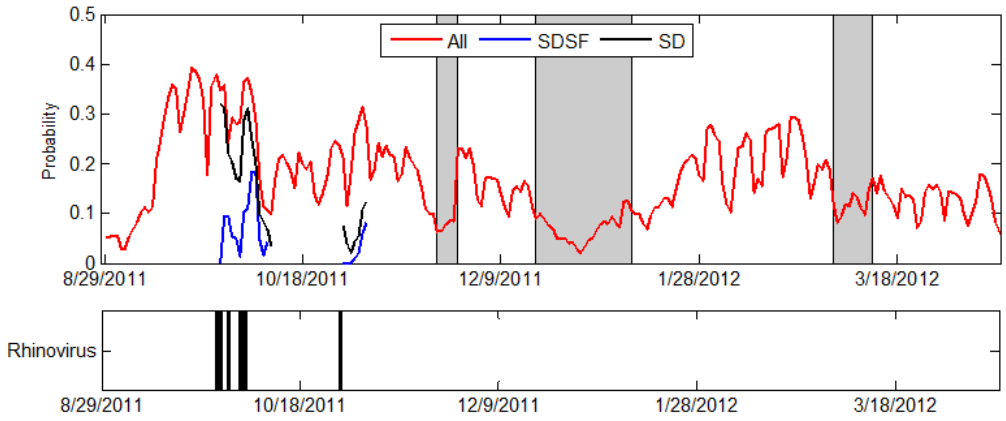

Fig. 9. As in Figure 7, for academic year 2011-2012.

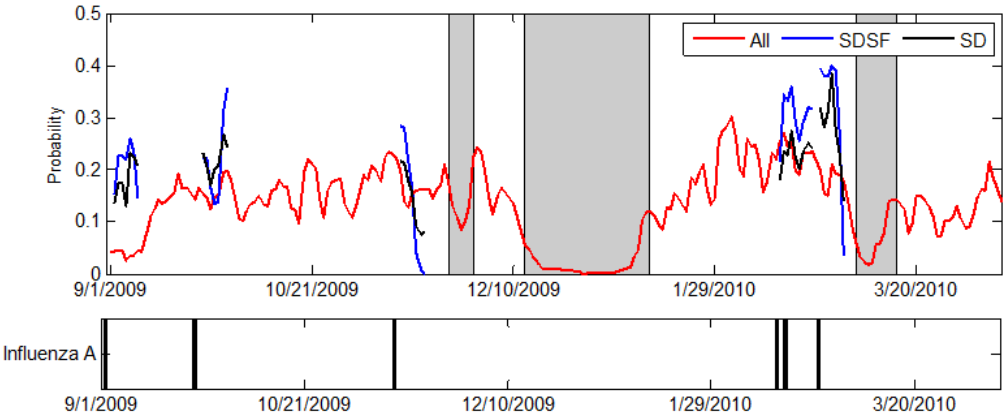

Fig. 10. Top figure: Probability of being in the infective state on a given day, for academic year 2009-2010. "All" refers to the average across all the students. "SDSF" refers to the average of students living in the same dorm and same floor with an infective individual. "SD" refers to the average of students living in the same dorm with an infective individual. The SD and SDSF cases are only for confirmed cases of Influenza A. The vertical gray bars represent, from left-to-right, Thanksgiving break, inter-semester (Winter/Christmas) break, and Spring break. Bottom figure: RT PCR test results, black line denotes at that time Influenza A was detected.

To summarize the form of the results, in Figure 4 are shown results for the 2009-2010 academic year. On the bottom of Figure 4, a black bar represents the presence of a RT PCR-confirmed virus of noted type. At the top in Figure 4 is shown the average probability of being in the infective state, under three circumstances. In red are shown results for all students and all times, and therefore in this case $\mathcal{X}$ denotes the set of all students. The blue curve corresponds to the case for which $\mathcal{X}$ corresponds to the set of students from the same dorm and same floor (SDSF) of a RT PCR-confirmed case. Finally, for the black curve, $\mathcal{X}$ corresponds to the set of students in the same dorm (SD) of a RT PCR-conformed case. Unlike

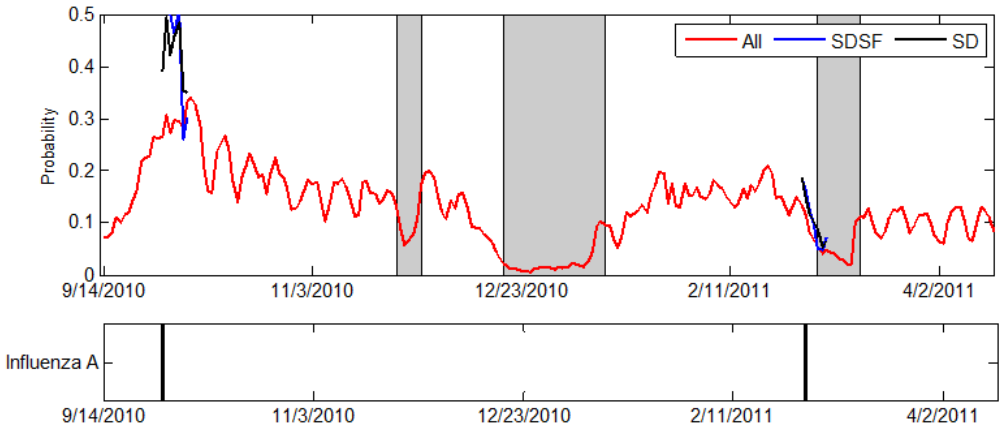

Fig. 11. As in Figure 10, for academic year 2010-2011.

the case of all students (red), for the cases of SDSF and SD the curves are not shown at all times, because we only look within a window of seven days after a RT PCR-based detection, and in some cases there are no data ($e.g.$, a given RT PCR-confirmed student doesn't have a sufficient number of students in the study on the same floor, or there were no RT PCR-confirmed detections in the last seven days).

From Figure 4, we typically see the following trend. If a student gets sick (is in state I) within a given dorm, from one of the specified viruses, then over the proceeding seven days the average probability of students within the same dorm (SD) will have a heightened probability of being in the infective state as compared to the general student population. Moreover, if a student gets infected, within a week students on the same dorm floor (SDSF) typically have, on average, an even higher probability of being in the infective state. This is not always true, but it seems to be a fairly common situation.

In Figures 5 and 6 results are shown in this same format as in Figure 4, for the 2010-2011 and 2011-2012 academic years, respectively. Given the relatively small number of RT PCR-confirmed cases relative to the total population size, it is difficult to make strong conclusions from Figures 4-6. There are periods in which being on the same dorm as an infective student clearly manifests increased probability of being in the infective state, over a subsequent 7 day period, but during other times this trend is not evident ($e.g.$, during the 2011-2012 academic year). One interpretation is that the presence of dorm colleagues who are sick may heighten attention to protecting oneself, through hand washing, etc. Therefore, from this perspective, the presence of a sick student may actually encourage more-healthful behavior in others.

To examine this issue from a finer perspective, we now examine these same types of curves, but for two specific viruses: Rhinovirus and Influenza A. Rhinovirus is associated with the "common cold," and therefore it is a virus that all students will come in contact with, in and outside their dorm. Therefore, in the case of Rhinovirus, the connection to the dorm, and who is infected there at a given time, may be more tenuous (students will come in contact with Rhinovirus and associated infective students in their classes, and other activities outside their dorm). Influenza A occurs more rarely, and therefore if someone is confirmed as infected with Influenza A, it is anticipated that students within the same dorm (SD) and same dorm floor (SDSF) may be at higher risk of infection.

In Figures 7-9 we show results like discussed above, but now for the SD and SDSF cases we only consider situations in which there was PCR-confirmed Rhinovirus-induced illness. For the case of Rhinovirus, we generally observe that if in the same dorm (SD) or on the same dorm floor (SDSF), when a given student is infected his/her dorm neighbors have a heightened probability of being in the infective state over the next week. However, there are cases for which this is not the case, which indicates that for Rhinovirus transmission activities outside the dorm may be as or more important that the degree of infection within the dorm.

In Figures 10-11 similar results are shown as above, but now only Influenza A is considered for the SD and SDSF cases. There are fewer Influenza A cases than Rhinovirus, so conclusions must be drawn with care. Nevertheless, for the case of Influenza A, the SD probability of being infected within a week of a confirmed Influenza A case is heightened relative to the general population, and the SDSF is generally further heightened. Note that in many cases, after roughly 5 days from a confirmed Influenza A case, the SD/SDSF probability of being infected is *less than* that of the general population; right after the confirmed Influenza A case the SD/SDSF probability of being infected increases, but then it diminishes relative to the general student population ($e.g.$, see the case in November 2009, in Figure 10). This phenomenon may be attributed to acquired immunity, after being infected.

An interesting phenomenon is observed in Figures 7-9, when considering the probability of being in an infective state for all of the students (red curve). Note that when the students come together at the beginning of the school year, and after the long Winter break, a general increase in the probability of being in an infective state is observed. Note that at the beginning of the school year, this is particularly

evident in the 2010-2011 year (Figure 8), and in 2011-2012 (Figure 9). Therefore, in Figure 8 and 9 the students are coming together for the first time at the beginning of the school year, from all over the United States, and from many other countries across the world. This phenomenon of increased probability of infection as students come together for the first time, or after extended break, may be associated with the general spread of infectious disease caused by a new mix of people, as been observed previously in the literature [12], [13].

Note that Figure 7 for 2009-2010 has temporal dependence (red curve) that is distinct from 2010-2011 and 2011-2012 (respectively Figures 7-8). This may be attributed to the fact that the 2009-2010 academic year was the year of the novel H1N1 virus, and significantly heightened on-campus attention to protecting oneself from virus transfer. These results seem to indicate that the heightened attention to viral transfer manifested by the novel H1N1 virus had a significant impact in reducing the probability of students transiting from the S to I state (not just from H1N1 virus, but from all viruses), when compared to two years in which such attention to viruses was far more muted on campus.

E. Classification performance based on symptoms

In the above results, we considered the average probability that students were in state I at a particular point in time. We wish to now examine the accuracy of the prediction of infection, relative to an objective "truth." To do this, we considered all 897 individuals for whom RT PCR-based virus-identification was performed (for a subset of these, for which the RT PCR test was negative, confirmation gene-expression analysis was also performed).

In our study there were two reasons a given student could have a RT PCR test performed: (i) based upon their self-reported data, a doctor in (near) real-time determined that they were infected, and therefore they were contacted for acquisition of a sample; (ii) a given individual was a close contact of a person who was sampled in the case of (i). Therefore, for the close contacts, the student may not be in an infective state at the time of sampling, either because there was no disease transfer, or because the onset of illness was manifested at a later time.

Based upon the symptoms, and the doctor-based diagnosis, all individuals in case (i) above are defined as being in the infective state, essentially by definition (these students were only contacted because their symptoms were deemed above a threshold of illness). For the close contacts, all individuals for whom the

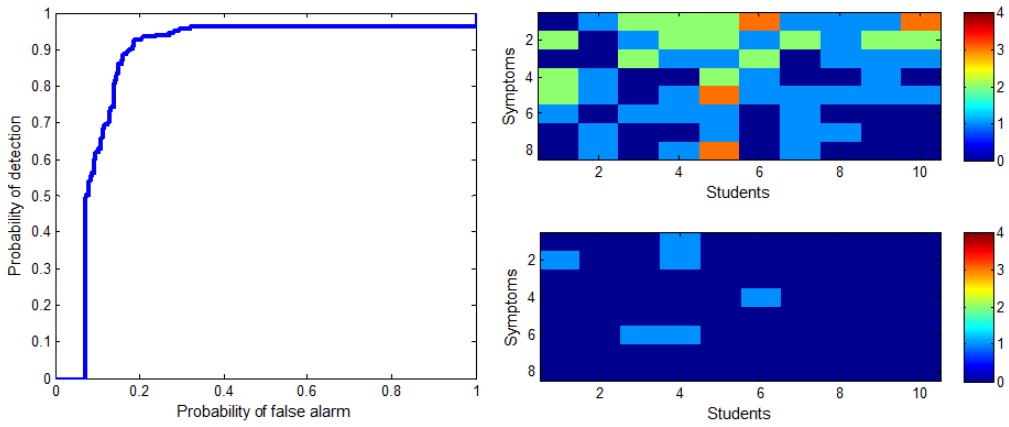

Fig. 12. Left column: ROC curve. Right column: Top figure is the symptom scores of students who are healthy (in state S) but labeled infective (in state I) with high probability by the model. The bottom figure shows the symptom scores of students who are infective (in state I) but labeled healthy (in state S) by the model. The order of the symptoms (1-8, top to bottom) is consistent with the order of the symptoms in the correlation matrix of Figure 1.

RT PCR test was negative will be deemed as healthy (in susceptible, S) state, and all others were deemed to be in the infective state (the RT PCR test may miss some infected people, which the gene-expression analysis can pick up, and this issue is discussed when presenting results).

The receiver operating characteristic (ROC) curve is manifested by thresholding $p(z_{nt} = I|-)$, and is shown in the left column of Figure 12. Note that the model achieves a 90% detection rate at a false-alarm rate of 15%. However, the quality of the ROC is undermined by imperfections in the definition of "truth." People who are sick as a result of illnesses other than virus will deemed as healthy in the truth (negative RT PCR test), but in reality they are sick. Another source of errors are manifested by positive RT PCR tests for the presence of virus, but the individual shows no symptoms – these are termed "shedders" in the medical community [28]. These individuals are carrying the virus, and shedding the virus, but they do not show any symptoms. The RT PCR test will deem these individuals as being in the infective state I, but from the standpoint of symptoms, which is what our analysis considers, these people are not infected (there are no symptoms present that would allow one to declare they are infected, based on symptoms alone).

In Figure 12, left, note (a) the presence of many false alarms before any detections are achieved (left-most part of the ROC), and (b) after a probability of false alarms of about 0.35, the detection probability

is stuck at around 0.95, until the very rightmost part of the ROC. Concerning (a), on the top-right of Figure 12, we show the symptom scores for the ten students who characterize the individuals detected as being infected, but RT PCR deems as being healthy, or in state S (the false alarms at the beginning of the ROC). Based upon the symptoms (right in Figure 12), these students are almost certainly sick due to some cause other than virus, or because of limitations of the RT PCR test ($e.g.$, poor samples, or because the illness was caused by a virus other than that tested by the RT PCR).

At bottom-right in Figure 12 is shown the symptom scores of students who were deemed infective via RT PCR, but our model deemed as healthy, based upon the symptoms. We see that the symptoms of these students are indeed very mild, or absent. These individuals were likely carrying a virus that was tested, and that was detected via RT PCR. However, these individuals were likely recently infected with the virus, and still carrying it, but no longer infected. Alternatively, these individuals may have been asymptomatic shedders.

Gene expression data were available for 6 of the individuals considered at right in Figure 12. In all of these cases, the gene-expression analysis was able to confirm the labels inferred by our algorithm based on symptoms.

F. Online prediction of health

The results in Figure 12 on predicting the state of health were based on *all* of the self-reported data, at all times for which data were reported. Of course, in a clinical setting a clinician must predict the state of health only based upon symptoms up to the point at which a diagnosis is made. It is desirable to predictive probability that a particular student is in state I on day $t+1$, based on symptom scores up to day t.

Let $\mathbf{y}_1^t = \{\mathbf{y}_1, ..., \mathbf{y}_t\}$ represent the symptom scores up to time t for the student in question. The probability that the student is in state I on day $t+1$ may be expressed as

$$p(z_{t+1} = I|\mathbf{y}_1^t, \Omega) = \frac{\sum_{d_{t+1}=1}^{D_{max}} p(z_{t+1} = I, d_{t+1}, \mathbf{y}_1^t|\Omega)}{\sum_{d_{t+1}=1}^{D_{max}} p(z_{t+1} = I, d_{t+1}, \mathbf{y}_1^t, |\Omega) + p(z_{t+1} = S, \mathbf{y}_1^t|\Omega)} \tag{15}$$

where $p(z_{t+1} = I, d_{t+1}, \mathbf{y}_1^t|\Omega)$ represents the joint probability of data $\mathbf{y}_1^t$, that the student is in state I on day $t+1$, and that they are in day d_{t+1} of being infected; $d_{t+1} \in \{1, .., D_{max}\}$ is the number of days left

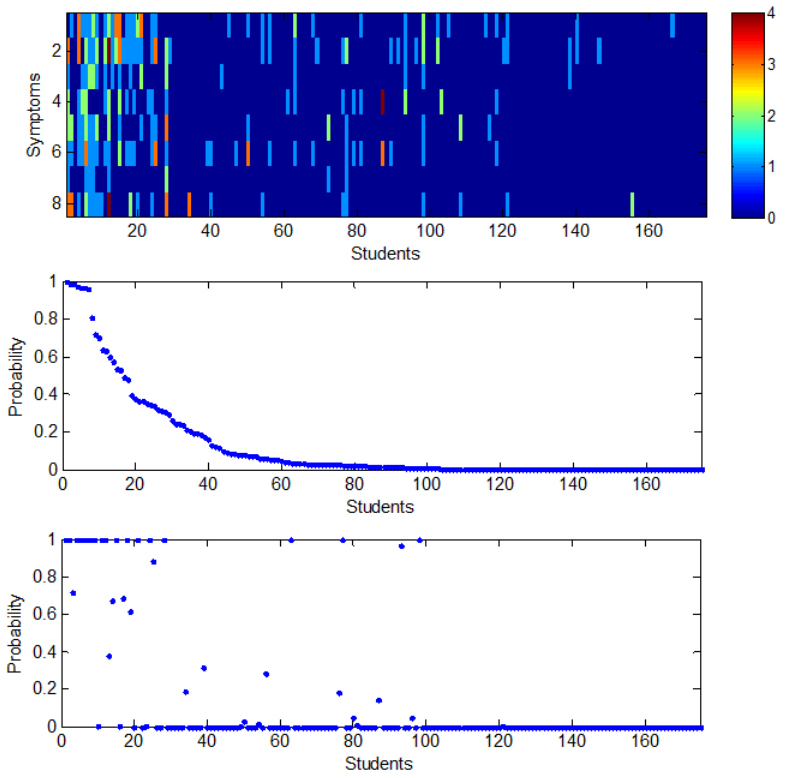

Fig. 13. The top figure is the symptoms scores for students at time $t+1$ (can be considered as "truth"). The middle figure is $p(z_{nt+1} = I|\mathbf{y}_{n1}^t, -)$, the predictive probability that a given student is in the infective state at $t+1$. The bottom figure is the probability that students stay in infected at $t+1$ given all the data. The order of the symptoms (1-8, top to bottom) is consistent with the order of the symptoms in the correlation matrix of Figure 1.

in infective state at time $t+1$. $p(z_{t+1} = S, \mathbf{y}_1^t|\Omega)$ represents the joint probability of the data and being in the susceptible state S. In both cases, Ω represents model parameters learned from data up to day t. The details for calculating $p(z_{t+1} = I, d_{t+1}|\mathbf{y}_1^t, \Omega)$ and $p(z_{t+1} = S, \mathbf{y}_1^t|\Omega)$ are provided in Appendix B.

In this experiment, the first two years of data, and the third year of data up to day $t = 140$ are employed to learn Ω (these are typical results for many values of t selected in Year 3). In Figure 13 are shown the model predictions for all students in Year 3 (2011-2012), where in Figure 13 the students are ordered from left to right from the most to least probable of being in state I on day $t+1$. At the top in Figure 13 are shown the symptoms reported on day $t+1$ (for those for whom scores were provided), and it is evident that the individuals who are deemed most likely to be in state I on day $t+1$ (based on data up to day t) tend to have the strongest symptoms on that day. In the middle in Figure 13 is shown the probability of being in an infective state on day $t+1$, based on data up to day t. Finally, the

bottom part of Figure 13 shows the probability of being in an infective state on day $t+1$ based on all of the data. Note that there is generally good agreement (middle and bottom figures) on which students are most likely to be in the infective state on day $t+1$.

G. Breaking out model components

In Figure 14 are plotted the posterior mean of the general trend terms $\gamma_t^{(1)}$ for academic years $2009-2010$, $2010-2011$ and $2011-2012$; the error bars reflect one standard deviation (estimated from the Gibbs collection samples). The weekly parameter $\gamma_t^{(2)}$ is displayed in Figure 15. In this figure the weeks are identified, with the beginning of a week defined here as Monday. We observe the that the weekly pattern (impacting the probability of transiting from healthy to infective state) is typically peaked at either Wednesday or Thursday, and tends to be smaller around the weekend. This is possibly reflective of the fact that students are more likely to report symptoms during the school week than they are on the weekend, when they may be distracted by funner activities. Of course, another interpretation is that the probability that the students will *feel* infected/sick is diminished during the weekend, relative to the middle of the week, when they may be under greater stress.

Recall Figure 3 from above, which depicts the degree of missingness on average as a function of days. By construction, heightened missingness is deemed associated with health, and weekends tend to be periods of high missingness. Whatever the cause of the weekly effects (student laziness/distraction or actual health), model interpretation may be improved by removing this effect. We consider this below.

In Figure 16 we show $\gamma_t^{(3)}$ associated with the Google Flu Trend data. In this plot we show the mean and one standard deviation, again from posterior collection samples. The posterior distribution in this term is manifested by the posterior distribution on ξ, as the total term is ξf_t, and $f(t)$ represents the deterministic/observed Google FluTrends (for the city of Durham, NC). Note that the contribution of the Google Flu Trend term is relatively small (large mass concentrated around zero, particularly for the first two years), which implies that the spread of infectious disease among students on the Duke University campus is a relatively isolated ecosystem, distinct from the city and community of Durham used here for f_t.

In Figure 17 we depict the inferred probability of transiting from state S to state I, as a function of day, for each of the three years of the study. The data were analyzed using all components of the

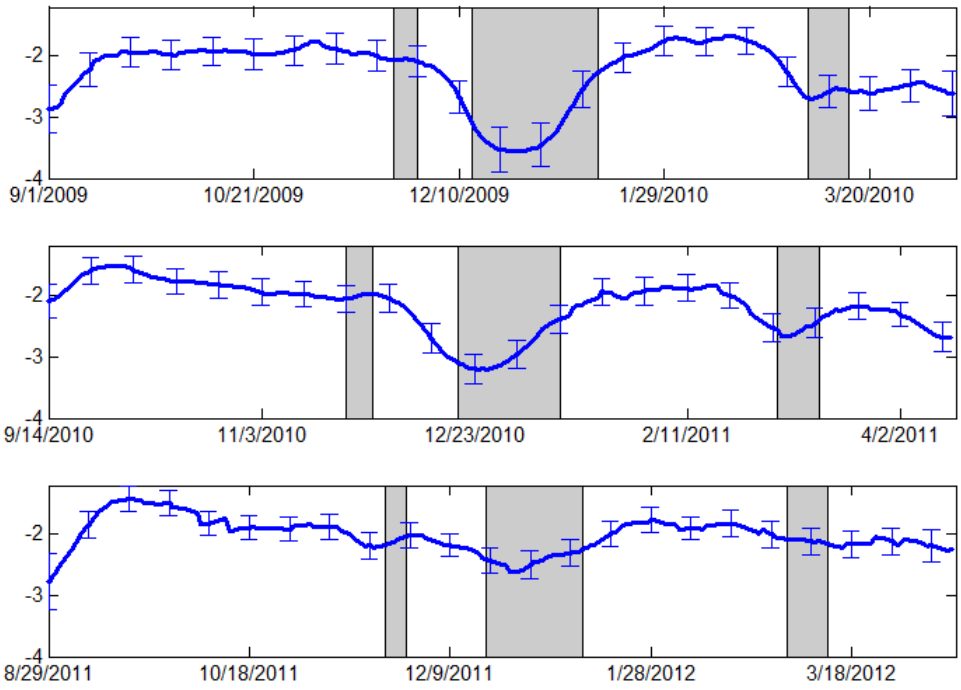

Fig. 14. General trend term $\gamma_t^{(1)}$. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The error bars reflect one standard deviation. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

model. However, after this analysis, to remove the effects of the weekly term, we show the model-inferred probability of transiting from $S \rightarrow I$, and with the weekly term removed. It is interesting to examine the red curve in Figure 17, in which the weekly effects are removed. Recall that the 2010-2011 and 2011-2012 academic years were distinct from 2009-2010, as the latter was associated with the novel H1N1 virus. Note that at the beginning of the academic year in 2010-2011 and 2011-2012, there is a clear increased probability of getting infected within the first month or so the students are together, presumably a mixing effect [12], [13] caused by interactions of many people who have never met before, coming from all over the United States, and also from outside the United States. It appears that the heightened attention to viruses (from the alarm associated with novel H1N1) dampened this phenomenon in 2009-2010. During the first semester of 2009-2010, when there was so much attention to viruses on campus, there is a noticeable decrease in the probability of transiting from state S to I, after the weekly effects are removed (relative to 2010-2011 and 2011-2012).

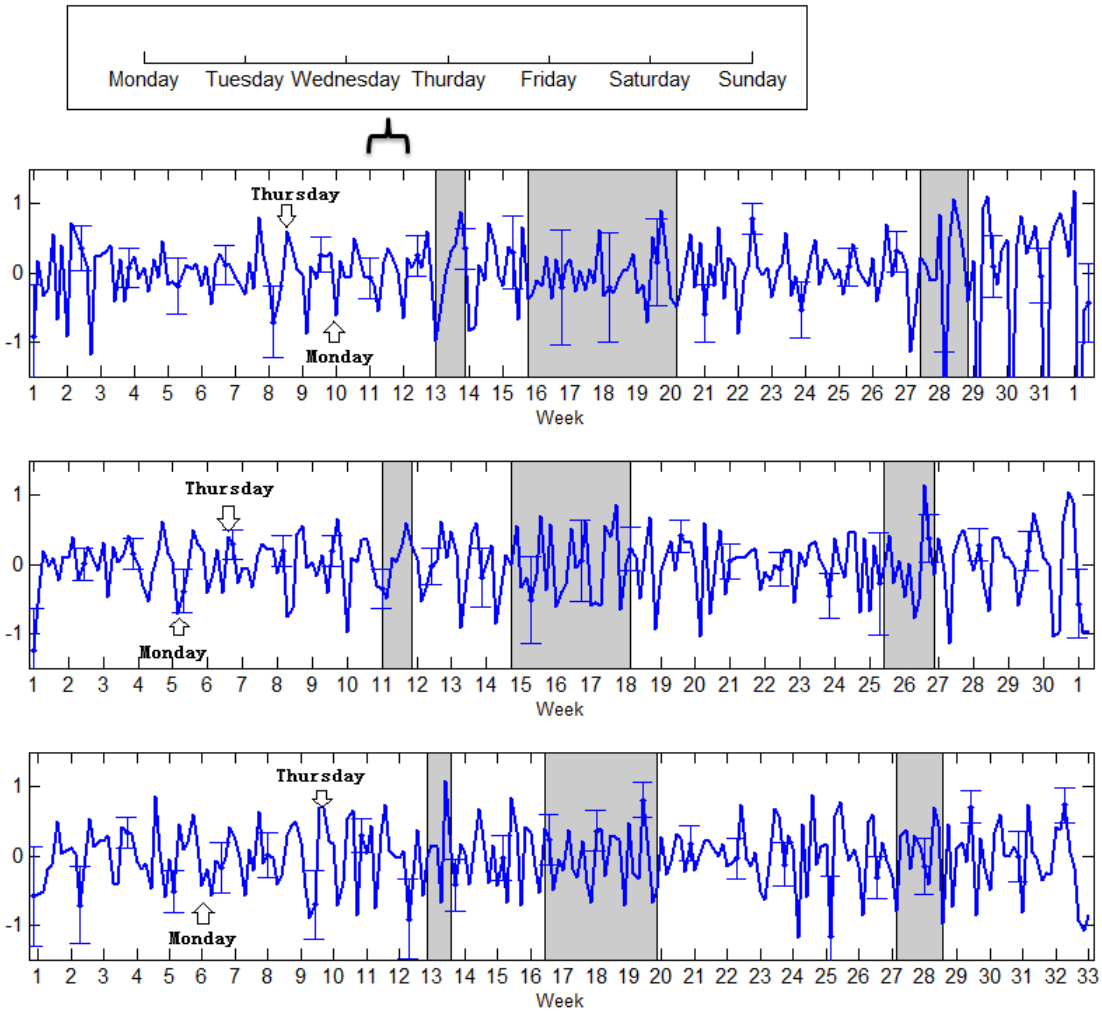

Fig. 15. Weekly or semi-periodic term $\gamma_t^{(2)}$. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

VI. Conclusions

A statistical model has been developed for analysis of the time-dependent symptom scores provided by a large group of undergraduate college students. Unlike almost all studies of data related to infection transfer, the model has operated directly on the observed symptoms, and the state of the students were assumed to be latent. The community-to-person mechanism for pathogen transfer has been modeled in terms of a SIS analysis, and computations have been performed using Bayesian (MCMC) methods.

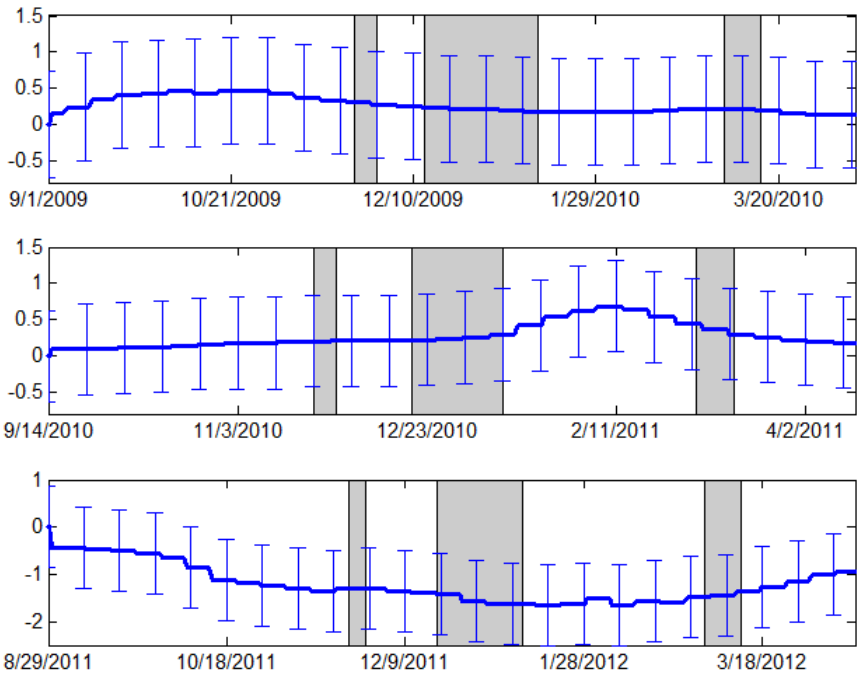

Fig. 16. Google Flu Trends (for Durham, NC, USA) regression term $\gamma_t^{(3)}$. From top to bottom are the results for academic year 2009-2010, 2010-2011 and 2011-2012. The gray bars reflect, from left-to-right, Thanksgiving break, Winter/Christmas break, and Spring break.

A detailed characterization of the data and the scientific questions that have motivated this study are discussed in Section II; a comprehensive answering of these questions with the available data has been provided in Section V. For brevity, we do not repeat these details here. We note that these data are presented here for the first time, and were collected by the authors; all data will be made available to the research community.

There are further questions that may be examined with the collected data, and that are worthy of future study. The identification of the virus responsible for each illness has been (imperfectly) constituted via RT PCR, for a large set of common viruses considered. We have access to the dorm in which each individual resided. A more detailed analysis of pathogen transfer as a function of virus type can be examined. In this paper we have presented results in this direction, but more explicit modeling could be performed (not necessarily at the symptom level, but after the responsible virus has been identified by RT PCR).

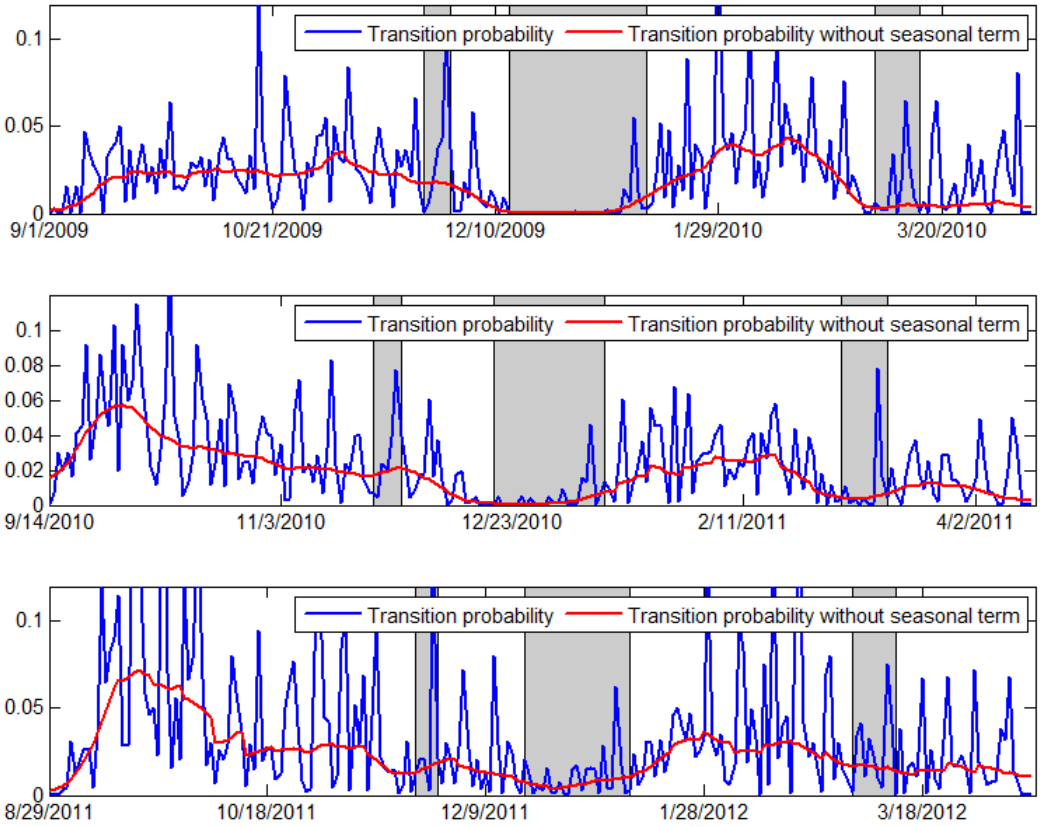

Fig. 17. Probability of transiting from state S to state I. The blue curve represents the total probability, and the red curve represents the probability with the weekly term $\gamma_t^{(2)}$ removed.

The gene expression data from this study have only been employed here in a limited manner, as the focus has been on self-reported symptom scores. However, for the close contacts, we have daily gene expression data for a week. For close contacts who transited from state S to I, we have the opportunity to analyze the time trajectory of the gene expression data as the host responds to the (known) virus. We have performed work of this type for people enrolled in challenge studies (controlled experiments) [23]; the data from this study offers the potential for similar studies on data from individuals who became ill in natural settings. We have preliminary results in this direction on these data, which are encouraging and will be presented elsewhere.

REFERENCES

[1] P. Birrella, G. Ketsetzisb, N. Gayc, B. Cooperd, A. Presanisa, R. Harrisb, A. Charlettb, X.-S. Zhangb, P. Whiteb, R. Pebody, and D. D. Angelisa, "Bayesian modeling to unmask and predict influenza A/H1N1pdm dynamics in London," *Proc. Nat. Acad. Sci.*, vol. 108, Nov. 2011.

[2] S. Cauchemez, F. Carrat, C. Viboud, A. Valleron, and P. Boelle, "A bayesian MCMC approach to study transmission of influenza: application to household longitudinal data," *Statist. Med.*, vol. 23:34693487.

[3] D. Clancy and P. O'Neill, "Exact Bayesian inference and model selection for stochastic models of epidemics among a community of households," *Scandinavian J. Stat.*, vol. 34, 2007.

[4] G. Jones, W. O. Johnson, W. D. Vink, and N. French, "A framework for the joint modeling of longitudinal diagnostic outcome data and latent infection status: Application to investigating the temporal relationship between infection and disease," *Biometrics*, vol. 68, June 2012.

[5] P. O'Neill, D. Balding, N. Becker, M. Eerola, and D. Mollison, "Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods," *Appl. Statist.*, vol. 49, 2000.

[6] Y. Yang, M. Halloran, M. Daniels, I. L. Jr., D. Burke, and D. Cummings, "Modeling competing infectious pathogens from a bayesian perspective: Application to influenza studies with incomplete laboratory results," *J. Am. Stat. Assoc.*, vol. 105, Dec. 2010.

[7] H. Hethcote, "The mathematics of infectious diseases," *SIAM Rev.*, vol. 42, 2000.

[8] W. Dong, A. Pentland, and K. Heller, "Graph-coupled hmms for modeling the spread of infection," *Uncert. Artificial Intell. (UAI)*, vol. 4, 2012.

[9] S. Stebbins, D. Cummings, J. Stark, C. Vukotich, K. Mitruka, W. Thompson, C. Rinalso, L. Roth, M. Wagner, S. Wisniewski, V. Dato, H. Eng, and D. Burke, "Reduction in the incidence of Influenza A but not Influenza B associated with use of hand sanitizer and cough hygiene in schools," *Pediatric Infectious Disease J.*, vol. 30, Nov. 2011.

[10] F. Carrat, C. Sahler, M. Leruez, S. Roger, F. Freymuth, C. L. Gales, M. Bungener, B. Housset, M. Nicolas, and S. Rouzioux, "Influenza burden of illness: estimates from a national prospective survey of household contacts in France," *Archives of Internal Medicine*, 2002.

[11] E. Patrozou and L. Mermel, "Does influenza transmission occur from asymptomatic infection or prior to symptom onset?" *Public Health Rep.*, MarApr 2009.

[12] Y. Sun, Z. Wang, Y. Zhang, and J. Sundell, "In China, students in crowded dormitories with a low ventillation rate have more common colds: evidence for airborne transmission," *PLoS One*, Nov. 2011.

[13] V. Kak, "Infections in confined spaces: cruise ships, military barracks, and college dormitories," *Infect. Dis. Clin. North Am.*, pp. 773–784, 2007.

[14] M. Jahrer and A. Toscher, "Collaborative filtering ensemble," *ACM International Conference on Knowledge Discovery and Data Mining (KDD), KDD Cup Workshop*, 2011.

[15] J. Silva and L. Carin, "Active learning for online bayesian matrix factorization," *Proc. ACM SIGKDD Conf. Knowledge Discovery and Data Mining*, 2012.

[16] M. West and J. Harrison, *Bayesian Forecasting and Dynamic Models*. Springer, 1989.

[17] J. Ginsberg, M. Mohebbi, R. Patel, L. Brammer, M.S.Smolinski, and L. Brilliant, "Detecting influenza epidemics using search engine query data," *Nature*, vol. 457, Feb. 2009.

[18] V. Dukic, H. Lopes, and N. Polson, "Tracking epidemics with google flu trends data and a state-space SEIR model," *J. Am. Stat. Ass.*, Dec. 2012.

[19] E. Fox and D. Dunson, "Bayesian nonparametric covariance regression," *arXiv:1101.2017v2*, Feb. 2011.

[20] V. Trifonov, H. Khiabanian, and R. Rabadan, "Geographic dependence, surveillance, and origins of the 2009 Influenza A (H1N1) virus," *New England J. Med.*, vol. 361, pp. 115–119, 2009.

[21] C. Carvalho, J. Chang, J. Lucas, J. Nevins, Q. Wang, and M. West, "High-dimensional sparse factor modelling: Applications in gene expression genomics," *Journal of the American Statistical Association*, vol. 103, pp. 1438–1456, 2008.

[22] B. Chen, M. Chen, J. Paisley, A. Zaas, C. Woods, G. Ginsburg, A. H. III, J. Lucas, D. Dunson, and L. Carin, "Nonparametric Bayesian factor analysis: Application to time-evolving viral gene-expression data," *BMC Bioinformatics*, vol. , 11:552, 2010.

[23] M. Chen, A. Zaas, C. Woods, G. Ginsburg, J. Lucas, D. Dunson, and L. Carin, "Predicting viral infection from high-dimensional biomarker trajectories," *J. Am. Stat. Ass.*, vol. 106:496, pp. 1259–1279, 2011.

[24] A. Zaas, M. Chen, J. Lucas, T. Veldman, A. Hero, J. Varkey, R. Turner, C. Oien, S. Kingsmore, L. Carin, C. Woods, and G. Ginsburg, "Peripheral blood gene expression signatures characterize symptomatic respiratory viral infection," *Cell Host & Microbe*, vol. 6, pp. 207–217, 2009.

[25] S. O'Brien and D. Dunson, "Bayesian multivariate logistic regression," *Biometrics*, vol. 60, pp. 739–746, 2004.

[26] M. Chen and D. Dey, "Bayesian analysis for correlated ordinal data model," *Generlized Linear Models: A Bayesian Prospective*, 2001.

[27] S. Yu and H. Kobayashi, "A hidden semi-markov model with missing data and multiple observation sequences for mobility tracking," *Signal Processing*, 2003.

[28] N. Dimmock, A. Easton, and K. Leppard, *Introduction to modern virology*. Blackwell, 2007.

[29] K. Murphy, "Hidden semi-markov models," *UBC Technical Report*, 2002.

[30] S. Yu and H. Kobayashi, "An efficient forward-backward algorithm for an explicit-duration hidden markov model," *IEEE Signal Processing Letters*, 2003.

[31] M. Johnson and A. Willsky, "Bayesian nonparametric hidden semi-markov models," *Journal of Machine Learning Research*, 2013.

[32] S.Yu, "Hidden semi-markov model," *Artificial Intelligence*, pp. 215–243, 2010.

[33] J. Gael, Y. Saatci, Y. Teh, and Z. Ghahramani, "Beam sampling for the infinite hidden markov model," *International Conference on Machine Learning*, 2008.

[34] M. Dewar, C. Wiggins, and F. Wood, "Inference in hidden markov models with explicit state duration distributions," *IEEE Signal Processing Letters*, 2012.

[35] X. Zhang, W. Boscardin, and T. Belin, "Sampling correlation matrices in bayesian models with correlated latent variables," *Journal of Computational Graphics Statistics*, 2006.

[36] C. Carter and R. Kohn, "On gibbs sampling for state space models," *Biometrika*, 1994.

[37] S. Fruhwirth-Schnatter, "Data augmentation and dynamic linear models," *Journal of Time Series Analysis*, 1994.

[38] J. Albert and S. Chib, "Bayesian analysis of binary and polychotomous response data," *Journal of American Statistical Association*, 1993.

APPENDIX A: MCMC UPDATE EQUATIONS

The full posterior distribution can be approximate via Gibbs sampler, with Metropolis-Hastings updates for a subset of parameters. We briefly describe how to sample some of the most interesting parameters based on their conditional posterior distribution.

Sampling from the latent states

Sampling from the latent states z_{nt} is achieved by forward and backward sampling method [29], [30], [31], [32]. Define the forward equation

$$\alpha_{nt}(m,d) = p(\mathbf{y}_{n1}^{t}, z_{nt} = m, d_{nt} = d), m \in \{S, I\}, d = 1, .., D_{max}$$

where $\mathbf{y}_{n1}^{t} = [\mathbf{y}_{n1}, ..., \mathbf{y}_{nt}]$ and d_{nt} is the number of days students n left in infective state I after day t. For the susceptible state S, d_{nt} is not necessary and omit for brevity. Let denotes $E_{nt} = \{z_{nt}, d_{nt}\}$, the forward equation can be calculated from the following induction function.

For the Markovian states $E_{nt} = \{S\}$, transition into states $\{S\}$ at time t takes place either from $\{I, 1\}$ or $\{S\}$ at time $t-1$

$$\alpha_{nt}(S) = (\alpha_{nt-1}(I,1) + \alpha_{nt-1}(S)p(z_{nt} = S|z_{nt-1} = S))p(y_{nt}|z_{nt} = S)$$

For the semi-Markovian state $E_{nt} = \{I, d\}$, transition into states $\{I, d\}$ at time t takes place either from $\{I, d+1\}$ or $\{S\}$ at time $t-1$

$$\alpha_{nt}(I,d) = (\alpha_{nt-1}(I,d+1) + \alpha_{nt-1}(S)p(z_{nt} = I|z_{nt-1} = S)p(d_{nt} = d))p(y_{nt}|z_{nt} = I)$$

Then we can sample E_{nt} (the state z_{nt} and the duration d_{nt}) from the backward sampling step. For $t = T$, sample E_{nt}

$$p(z_{nT} = I, d_{nT} = d|y_{n1}^{T}) = \frac{\alpha_{nT}(I,d)}{\alpha_{nT}(S) + \sum_{d=1}^{D_{max}} \alpha_{nT}(I,d)}$$

$$p(z_{nT} = S|y_{n1}^{T}) = \frac{\alpha_{nT}(S)}{\alpha_{nT}(S) + \sum_{d=1}^{D_{max}} \alpha_{nT}(I,d)}$$

For $t \in T-1, ..., 1$, sample E_{nt}

$$p(z_{nt} = I, d_{nt} = d | y_{n1}^t, E_{nt+1}) = \frac{\alpha_{nt}(I,d) p(z_{nt} = I, d_{nt} = d | E_{nt+1})}{\alpha_{nt}(S) p(z_{nt} = S | E_{nt+1}) + \sum_{d=1}^{D_{max}} \alpha_{nt}(I,d) p(z_{nt} = I, d_{nt} = d | E_{nt+1})}$$

$$p(z_{nt} = S | y_{n1}^t, E_{nt+1}) = \frac{\alpha_{nt}(S) p(z_{nt} = S | E_{nt+1})}{\alpha_{nt}(S) p(z_{nt} = S | E_{nt+1}) + \sum_{d=1}^{D_{max}} \alpha_{nt}(I,d) p(z_{nt} = I, d_{nt} = d | E_{nt+1})}$$

The above method need to specify the maximum number of duration D_{max} in order to avoid infinite number of states E_{nt}. We may employ the beam sampling idea developed in [33], [34] to avoid setting D_{max}. The main idea of beam sampling is introducing auxiliary random variables $\mathbf{u}_{n1}^t$ for slice sampling. The forward equation is modified as

$$\begin{aligned} \alpha_{int}^*(E_{nt}) &= p(E_{nt}, \mathbf{y}_{n1}^t, \mathbf{u}_{n1}^t) \\ &= \sum_{E_{nt-1}} I(0 < u_{nt} < p(E_{nt} | E_{nt-1})) \alpha_{nt-1}^*(E_{nt-1}) p(y_{nt} | E_{nt}) \end{aligned}$$

The backward sampling part is

$$p(E_{nt-1} | E_{nt}, \mathbf{y}_{n1}^T, u) \propto I(0 < u_{nt} < p(E_{nt} | E_{nt-1})) \alpha_{nt-1}^*(E_{nt-1})$$

where $I(.)$ is the indicator function. $I(g(u_{nt})) = 1$ if $g(u_n t)$ is true and $I(g(u_{nt})) = 0$ otherwise.

Sampling the correlation matrix

The parameter extension method introduced in [35] is employed to sample the correlation matrix $\mathbf{\Sigma}_I$. Consider an unrestricted covariance matrix $\mathbf{\Sigma_1} \sim Wishart(m_1, \mathbf{V}_1)$, which can be decomposed as $\mathbf{\Sigma_1} = \mathbf{L}^{1/2} \mathbf{\Sigma}_I \mathbf{L}^{1/2}$, where $\boldsymbol{L}$ is the diagonal of the matrix with diagonal elements equivalent to the diagonal of $\mathbf{\Sigma_1}$. The prior for correlation matrix $\mathbf{\Sigma}_I$ is as following,

$$P(\mathbf{\Sigma}_I, \mathbf{L}) = Jacobian_{\mathbf{\Sigma}_1 \to (\mathbf{\Sigma}_I, \mathbf{L})} P(\mathbf{\Sigma}_1)$$

where $Jacobian_{\mathbf{\Sigma}_1 \to (\mathbf{\Sigma}_I, \mathbf{L})} = \prod_{i=1}^{J} q_i^{\frac{J-1}{2}}$ is the Jacobian transformation from $\mathbf{\Sigma_1}$ to $(\mathbf{L}, \mathbf{\Sigma}_I)$. Then the MH algorithm for sampling posterior distribution of $\mathbf{\Sigma}_I$ is as follows: at iteration t, generate the candidate values $\mathbf{\Sigma}_I^*$ from $\mathbf{\Sigma_1^*} = \mathbf{L}^{*1/2} \mathbf{\Sigma}_I^* \mathbf{L}^{*1/2} \sim Wish(m_1, \mathbf{V}_1)$, accept the candidate value with probability $\alpha = min\{1, \frac{p(\mathbf{L}^*, \mathbf{\Sigma_I^*} | -)}{p(\mathbf{L}^t, \mathbf{\Sigma_I^t} | -)} \frac{q(\mathbf{\Sigma}_I^t | \mathbf{\Sigma}_I^*)}{q(\mathbf{\Sigma}_I^* | \mathbf{\Sigma}_I^t)}\}$, where $q(.|\mathbf{\Sigma}_I^t)$ is the proposal distribution given by product the jacobian term and Wishart density $Wishart(m_0, \mathbf{\Sigma}_I^t)$. Sampling $\mathbf{\Sigma_S}$ is performed using a similar procedure.

Sample γ_{nt}

Define $b_{nt} = 1$ if $z_{nt-1} = S$ and $z_{nt} = I$, $b_{nt} = 0$ if $z_{nt-1} = S$ and $z_{nt} = S$. b_{nt} is treat as missing data, if $z_{nt-1} = I$ and $z_{nt} = I$. We use q_{nt} to denote the missing data, with $q_{nt} = 0$ refers to missing and $q_{nt} = 1$ otherwise. We can sample γ_{nt} as following,

$$\gamma_{nt} \sim \mathcal{N}_{(0,+\infty)}(\textstyle\sum_{i=1}^{3} \gamma_t^{(i)} + a_n, 1), \text{ if } b_{nt} = 1$$
$$\gamma_{nt} \sim \mathcal{N}_{(-\infty,0)}(\textstyle\sum_{i=1}^{3} \gamma_t^{(i)} + a_n, 1), \text{ if } b_{nt} = 0$$

$\mathcal{N}_{(0,+\infty)}$ and $\mathcal{N}_{(-\infty,0)}$ are the truncated normal distributions with truncation level $(0, +\infty)$ and $(-\infty, 0)$.

Sample $\gamma_t^{(1)}$ and $\gamma_t^{(2)}$

Sampling $\gamma_t^{(1)}$ and $\gamma_t^{(2)}$ are achieved via forward filtering and backward sampling method [36], [37]. Here we detail the update equations for sampling $\gamma_t^{(2)}$ and sampling $\gamma_t^{(1)}$ is performed in the similar way.

Define $\mathbf{F} = [1, 0, 1, 0, 1, 0]^T$, $\mathbf{G} = \begin{pmatrix} \mathbf{J}(w) & 0 & 0 \\ 0 & \mathbf{J}(2w) & 0 \\ 0 & 0 & \mathbf{J}(3w) \end{pmatrix}$, then $\gamma_t^{(2)} = \mathbf{F}^T \boldsymbol{\theta}_t$ and $\boldsymbol{\theta}_t = \mathbf{G}\boldsymbol{\theta}_{t-1} + \boldsymbol{\epsilon}_t$.

In the forward filtering step, assume $\boldsymbol{\theta}_0 \sim \mathcal{N}(\mathbf{m}_0^{(2)}, \mathbf{C}_0^{(2)})$, it can be shown the posterior at time t is

$$\boldsymbol{\theta}_t \sim \mathcal{N}(\mathbf{m}_t^{(2)}, \mathbf{C}_t^{(2)})$$

where $\mathbf{m}_t^{(2)} = \mathbf{C}_t^{(2)}(\sum_{n=1}^{N} \mathbf{F}\hat{\gamma}_{nt}^{(2)} q_{nt} + \mathbf{R}_t^{(2)-1}\mathbf{a}_t^{(2)})$, $\mathbf{C}_t^{(2)} = \mathbf{R}_t^{(2)} - \mathbf{A}_t^{(2)} Q_t^{(2)} \mathbf{A}_t^{(2)T}$, $\mathbf{A}_t^{(2)} = \sqrt{N_t}\mathbf{R}_t^{(2)}\mathbf{F}Q_t^{(2)-1}$, $Q_t^{(2)} = 1 + N_t \mathbf{F}^T R_t^{(2)} \mathbf{F}$, $\mathbf{a}_t^{(2)} = \mathbf{G}\mathbf{m}_{t-1}^{(2)}$, $\mathbf{R}_t^{(2)} = \mathbf{G}\mathbf{C}_{t-1}^{(2)}\mathbf{G}^T + \mathbf{W}_t$ and $\hat{\gamma}_{nt}^{(2)} = \gamma_{nt} - a_n - \gamma_t^{(1)} - \gamma_t^{(3)}$.

In the backward sampling step, first sample $\boldsymbol{\theta}_T \sim \mathcal{N}(\mathbf{m}_T^{(2)}, \mathbf{C}_T^{(2)})$ and then for day $T-1$ to day 1, sample

$$\boldsymbol{\theta}_t \sim \mathcal{N}(\hat{\mathbf{m}}_t^{(2)}, \hat{\mathbf{C}}_t^{(2)})$$

where $\hat{\mathbf{m}}_t^{(2)} = \mathbf{m}_t^{(2)} + \hat{\mathbf{B}}_t^{(2)}(\boldsymbol{\theta}_{t+1} - \mathbf{a}_{t+1}^{(2)})$, $\hat{\mathbf{C}}_t^{(2)} = \mathbf{C}_t^{(2)} - \hat{\mathbf{B}}_t^{(2)}\mathbf{R}_{t+1}^{(2)}\hat{\mathbf{B}}_t^{(2)T}$, $\hat{\mathbf{B}}_t^{(2)} = \mathbf{C}_t^{(2)}\mathbf{G}^T\mathbf{R}_{t+1}^{(2)-1}$. $\mathbf{W}_t$ is a block diagonal covariance matrix with each block equals to $\boldsymbol{\Sigma}_{\theta_r}$.

Sample r_{nt} and τ

Following the Gibbs sampling algorithm in [38], sample $\mathbf{r}_{nt} \sim \mathcal{N}_{(\boldsymbol{\tau}_{\mathbf{y}_{nt}-\mathbf{1}}, \boldsymbol{\tau}_{\mathbf{y}_{nt}})}(\boldsymbol{\mu}_{z_{nt}}, \boldsymbol{\Sigma}_{z_{nt}})$ where $\boldsymbol{\tau}_{\mathbf{y}_{nt}-\mathbf{1}}$ and $\boldsymbol{\tau}_{\mathbf{y}_{nt}}$ are the truncation level of the multivariate normal. For $m = 1, ..., M-1$, sample τ_{jm} from uniform

distribution with interval $[max(max(r_{ntj} : y_{ntj} = m-1), \tau_{jm-1}), min(min(r_{ntj} : y_{ntj} = m, \tau_{jm+1}))]$.

APPENDIX B: PREDICTION

Denote the current data as $\mathbf{y}_{n1}^t = \{\mathbf{y}_{n1}, ..., \mathbf{y}_{nt}\}$, we may derive the one step predictive probability for student n as following,

$$p(z_{nt+1} = I|\mathbf{y}_{n1}^t, \Omega) = \frac{\sum_{d_{nt+1}=1}^{D_{max}} p(z_{nt+1} = I, d_{nt+1}, \mathbf{y}_{n1}^t|\Omega)}{\sum_{d_{nt+1}=1}^{D_{max}} p(z_{nt+1} = I, d_{nt+1}, \mathbf{y}_{n1}^t, |\Omega) + p(z_{nt+1} = S, \mathbf{y}_{n1}^t|\Omega)} \tag{16}$$

where Ω is the model parameter learned from current data $\mathbf{y}_{n1}^t$ and $d = 1, .., D_{max}$ is the number of days left in infective states at time $t+1$. If we define $\hat{\alpha}_{nt+1}(I, d) = p(z_{nt+1} = I, d_{nt+1} = d, \mathbf{y}_{n1}^t|\Omega)$ and $\hat{\alpha_{nt+1}}(S) = p(z_{nt+1} = S, \mathbf{y}_{n1}^t|\Omega)$, the induction equation for $\hat{\alpha_{nt+1}}$ can be derived.

Similar with deriving the forward induction function for α_{nt} in Appendix A, for the Markov states $\{S\}$, transition into $\{S\}$ at day $t+1$ can only take place from $\{S\}$ and $\{I, 1\}$ at time t.

$$\begin{aligned}
\hat{\alpha}_{nt+1}(S) &= p(z_{nt+1} = S, \mathbf{y}_{n1}^t|\Omega) \\
&= \alpha_{nt}(I, 1) + P(z_{nt+1} = S|z_{nt} = S)\alpha_{nt}(S)
\end{aligned}$$

For the semi-Markov states $\{I, d\}$, transition into $\{I, d\}$ takes place from $\{I, d+1\}$ and $\{S\}$.

$$\begin{aligned}
\hat{\alpha}_{nt+1}(I, d) &= p(z_{nt+1} = I, d_{nt+1} = d, \mathbf{y}_{n1}^t|\Omega) \\
&= \alpha_{nt}(I, d+1) + \alpha_{nt}(S)p(z_{nt+1} = I|z_{nt} = S)p(d_{nt+1} = d)
\end{aligned}$$

$\alpha_{nt}(I, 1)$, $\alpha_{nt}(I, d+1)$ and $\alpha_{nt}(S)$ is obtained from training the model with current data. The one step forward prediction of transition probability $P(z_{nt+1} = I|z_{nt} = S) = \Phi(\hat{\gamma}_{nt+1})$ is obtained based on the properties of AR model.

$$\hat{\gamma}_{nt+1} \sim N(\mu_{\hat{\gamma}_{nt+1}}, \sigma_{\hat{\gamma}_{nt+1}})$$

where $\mu_{\hat{\gamma}_{nt+1}} = \mathbf{F}^T\mathbf{G}\mathbf{m}_t^{(2)} + \mu_{a_n} + \omega m_t^{(1)}$ and $\sigma_{\hat{\gamma}_{nt+1}} = \mathbf{F}^T(\mathbf{G}\mathbf{C}_t^{(2)}\mathbf{G}^T + \mathbf{W}_t)\mathbf{F} + \omega^2 C_t^{(1)} + \beta^{-1} + 1 + \sigma_{a_n}$, where $\mu_t^{(1)}$ and $C_t^{(1)}$ are the mean and variance obtained in the forward filtering step when sample $\gamma_t^{(1)}$. μ_{a_n} and σ_{a_n} are the posterior mean and variance of a_n. Notice for prediction, we do not take into account Google Flu Trend data.