

**Learning the magnitude and duration of influence of
infections**

by

Emily Mu

Submitted to the Department of Electrical Engineering and Computer
Science

in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2019

© Massachusetts Institute of Technology 2019. All rights reserved.

Author

Department of Electrical Engineering and Computer Science

May 24, 2019

Certified by

John Guttag

Professor

Thesis Supervisor

Accepted by

Katrina LaCurts

Chair, Master of Engineering Thesis Committee

Learning the magnitude and duration of influence of infections

by

Emily Mu

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 2019, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Clostridioides difficile infections (CDIs) impose a substantial burden on the health-care system leading to poor health outcomes, mortality and costs to the healthcare system estimated at greater than \$5 billion. One of the reasons why CDIs are hard to control is the contribution of individual infections to the risk of transmission is not well understood. In this paper, we propose modeling incident infections using a Hawkes process, which is a self-exciting stochastic process, encoding the intuition that new infections trigger further infections. Using data from a large urban hospital, we demonstrate that our approach reveals different patterns of infection spread across patient care units. These insights can be used to guide unit-specific interventions aimed at interrupting nosocomial transmission.

Thesis Supervisor: John Guttag

Title: Professor

Acknowledgments

Firstly, I would like to thank John Guttag, my thesis advisor, and Maggie Makar for all their work, feedback, and support throughout this project. They were both incredible mentors and I am grateful for their work in helping shape and guide this project.

I would also like to thank the infectious control department at Massachusetts General Hospital and our collaborators at the University of Michigan, for their support and feedback during this work. These individuals include Drs. Erica Shenoy M.D., Ph.D, David Hooper M.D., Lauren West MPH, and Jenna Wiens Ph.D. Their advice and clinical perspective during the progression of this work was invaluable.

Next, I would like to thank all of the members of the Clinical and Applied Machine Learning group at MIT for their continued feedback and input. These include and are not limited to Dr. Adrian Dalca Ph.D., Amy Zhao, Davis Blalock, Jose Javier Gonzalez Ortiz, Katie Lewis, Harini Suresh, Divya Shanmugam, Courtney Guo, and Advaith Anand.

Finally, I would like to thank my family and friends for their love and support throughout this year and through these last four years at this institution.

Contents

1	Introduction	13
1.1	Clinical Contribution	14
1.2	Technical Contribution	15
2	Related Work	17
2.1	Clinical Work	17
2.2	Risk Prediction	18
2.2.1	Early Models	18
2.2.2	Machine Learning Models	19
3	Hawkes Process	21
3.1	Single Hawkes Process	21
3.2	Network Hawkes Process	22
4	Learning the Magnitude and Duration of Influence	25
4.1	Event extraction	25
4.2	Methods	26
4.2.1	Single Hawkes Process Inference	26
4.2.2	Network Hawkes Process Inference	27
4.3	Results	29
4.3.1	Poisson vs. Hawkes Processes	29
4.3.2	Single Hawkes Simulations	31
4.3.3	Learning Single Hawkes Processes	34

4.3.4	Multiple Processes Analysis	36
4.3.5	Learning Network Hawkes Processes	38
5	CDI Risk Prediction	43
5.1	Cohort	43
5.2	Methods	44
5.3	Results	45
5.3.1	Cohort: Entire Hospital	46
5.3.2	Cohort: Top 9 Units	46
5.3.3	Pairwise Analysis	47
6	Discussion	49
A	Tables	51
B	Figures	53

List of Figures

4-1	The number of CDI per day for the entire hospital.	30
4-2	Recovering base rate after varying influence.	32
4-3	Recovering influence after varying influence.	32
4-4	Recovering base rate after varying influence.	33
4-5	Recovering influence after varying influence.	33
4-6	Units plotted by learned influence and decay with error for one day shift with perturbing the data.	35
4-7	Units plotted by learned influence and decay with error for one day shift with bootstrapping.	35
5-1	We plot the fraction of true positives for the number of days predicted in advance with SE unit (left) and HE unit (right).	47
B-1	Units plotted by learned influence and decay with error for zero day shift with perturbing the data.	53
B-2	Units plotted by learned influence and decay with error for zero day shift with bootstrapping.	54
B-3	Units plotted by learned influence and decay with error for two day shift with perturbing the data.	54
B-4	Units plotted by learned influence and decay with error for two day shift with bootstrapping.	55
B-5	Units plotted by learned influence and decay with error for three day shift with perturbing the data.	55

B-6 Units plotted by learned influence and decay with error for three day
shift with bootstrapping. 56

List of Tables

4.1	Anonymized unit acronyms and characteristics.	30
4.2	Test log-likelihood of Hawkes vs Poisson model for different units, 3 day shift. Note that log likelihood values closer to 0 is better.	31
4.3	Number of shifted events per unit	37
4.4	Unit OU A proportion of shifted events	37
4.5	Unit EMD proportion of shifted events	37
4.6	Unit MU proportion of shifted events	37
4.7	Unit GMU A proportion of shifted events	37
4.8	Unit average proportion of shifted events	37
4.9	Unit OU A base rate parameters	39
4.10	Unit OU A decay parameters	39
4.11	Unit OU A influence parameters	39
4.12	Unit EMD base rate parameters	40
4.13	Unit EMD decay parameters	40
4.14	Unit EMD influence parameters	40
4.15	Unit MU base rate parameters	41
4.16	Unit MU decay parameters	41
4.17	Unit MU influence parameters	41
4.18	Unit GMU A base rate parameters	42
4.19	Unit GMU A decay parameters	42
4.20	Unit GMU A influence parameters	42

5.1	Cohort results for the whole hospital. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.	46
5.2	Cohort results for the top 9 units. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.	47
A.1	Test log-likelihood of Hawkes vs Poisson model for different units, 0 day shift	51
A.2	Test log-likelihood of Hawkes vs Poisson model for different units, 2 day shift	52
A.3	Cohort results for the whole hospital. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.	52
A.4	Cohort results for the top 9 units. The model is run with no class balancing and L1-regularization. The AUROC test 95% confidence interval is reported.	52
A.5	Cohort results for the top 9 units with location removed. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.	52

Chapter 1

Introduction

Healthcare-associated infections (HAIs), or infections associated with the delivery of healthcare in hospitals, long-term care facilities and other care facilities, are a substantial cause of morbidity and mortality [18] and excess costs [33]. They have been the focus of several government initiatives (e.g., Healthy People 2020 [28] and CDC's Winnable Battles [9]) but without sufficient success. One of the most common HAIs, and the focus of this study, is *Clostridioides difficile* infection (CDI) (formerly called *Clostridium difficile*). Despite efforts by clinicians and policy makers, CDI continues to impose a significant burden on the healthcare system, resulting in approximately 453,000 infections, 29,000 deaths, and around \$4.8 billion dollars in medical costs every year [17].

The risk of developing CDI is a function of both exposure to CDI spores and patient susceptibility to developing the infection. Exposure to CDI occurs through direct and indirect contact; contamination of the hands of healthcare workers, portable medical equipment, and the hospital environment are all thought to contribute to transmission within the hospital setting. Transmission occurs when spores from the fecal matter of an infected patient are ingested by another patient. Because of the likely routes of transmission, it is reasonable to consider how variability at a unit level (i.e., compliance with hand hygiene, cleaning practices, patient turnover, physical layout/design) could impact the contribution of individual cases to onward transmission. Furthermore, *C. diff* has been shown to be highly resistant to normal methods

of eradication, including heat and regular disinfectants. There is evidence of variation in shedding of *C. diff* spores during the course of illness and recovery. However, the influence a single new CDI case has on increasing the risk of CDI for other patients and the duration of that influence is not known. Knowing that a new infection in a particular unit might have a strong impact in triggering further events could guide contact precaution and isolation measures, allowing focus on prevention efforts in high risk units. Knowing how long that effect lingers could guide infection control practices over the duration of high risk. In this work, we estimate the magnitude and duration of the impact of a new infection by recasting the problem of modeling infection spread using self-exciting temporal point processes. Specifically, we use Hawkes processes to model infections, where each new infection causes a ripple effect triggering further infections.

1.1 Clinical Contribution

Modeling CDI transmission as a Hawkes process addresses unanswered questions. It is known that an infected patient continues to raise the risk of infection in units even after they are discharged from that unit. However, it is unknown how much a new infection elevates the risk of further infections. We refer to this concept as the magnitude of the influence. This uncertainty arises because there are multiple factors that determine the influence of a new infection. These relate to both the susceptibility of the exposed patients and the intensity of the exposure. For example, it is well known that patients who are immunocompromised are more susceptible to developing disease if exposed than are other patient populations. Hence, a reasonable hospital policy might focus resources on ensuring that units occupied by immunocompromised patients have regular environmental disinfection with a sporicidal agent that could inactivate *C. diff*. However, it is also true that these units tend to have private rooms, limiting contact with other patients and reducing *C. diff* exposure. Depending upon the magnitudes of these different effects, interventions might be prioritized by unit to achieve the most cost-effective impact on interrupting transmission. By recasting

the problem as a Hawkes process, we can directly model the effect of a new infection on the instantaneous expected rate of new infections.

Secondly, it is unknown exactly how long a new infection elevates the risk of further infections. We refer to this as the duration of influence. It is known that spores formed by *C. diff* bacteria can survive outside of the human host and in the environment for quite a long time. Estimates of how long the bacterial spores can survive outside the host, range from several days to a year [8, 26, 31, 13, 14], in some cases even surviving the laundry process [27]. In addition to uncertainty regarding the duration of *C. diff* survival in the environment, it is difficult to determine the duration of risk of transmission from residual *C. diff* spores. The amount of environmental contamination, adequacy of environmental cleaning, and adequacy of hand hygiene and contact isolation activities of healthcare provider and visitors could shorten or elongate the duration of risk. Many of these factors are unobserved or unmeasured. Information on the duration of the influence of a patient with *C. diff* could inform approaches to interrupt transmission of *C. diff* to other patients. Our proposed method directly addresses that challenge by learning the decay rate of the effect of a new infection. By modeling each unit separately, we are able to model the differences in decay rates between units even if these differences are not explicitly recorded in the data.

1.2 Technical Contribution

The problem of modeling infection spread has been studied in the epidemiology, biostatistics, and machine learning literatures. Early mathematical models of infectious diseases, including the Susceptible-Infected-Recovered (SIR) and Susceptible-Infected-Susceptible (SIS) models assume that characteristics of immunity, susceptibility, and recovery do not vary by subgroups, e.g., hospital units. While these models provide important estimates for total number of infections over a period of time, they typically do not give detailed insight into the spread of a disease. [16] [7]

In the machine learning literature, the current state-of-the-art approaches for

characterizing the duration and magnitude of influence of CDI rely on manually constructed measures of exposure that assume a specific model for the magnitude of influence (e.g., number of infected patients) and duration (e.g., linear decay), and do not account for variations in these effects across units [29, 23]. In addition, to our knowledge, none of the existing methods characterizing the spread of CDI take into account uncertainty in the actual time of the CDI. We address these issues by suggesting a data-driven approach that learns a different stochastic process for each unit; allowing the different units to exhibit different spread patterns. Our method is able to incorporate the uncertainty in the time of exposure, translating it into uncertainty in the final estimates of the Hawkes process. In addition to guiding cleaning and isolation policies, the parameters learned by our model, specifically the learned expected rates of infection can be used as a meaningful measure of exposure to CDI. We show that combined with information about host susceptibility, our exposure estimates can be used to predict the onset of CDI.

Chapter 2

Related Work

In this section, we present related work from both a clinical and technical perspective. From a clinical perspective, we identify studies that identify relevant factors of exposure. We also present previous work with earlier models, such as the Susceptible-Infected-Recovered (SIR) and Susceptible-Infected-Susceptible (SIS) models and early risk prediction models, in comparison to more recent machine learning risk prediction literature.

2.1 Clinical Work

Exposure to *C. diff* spores is required for transmission, and thus efforts to reduce exposure will reduce incident CDI. Studies have shown that a one percentage point increase in CDI hospital incidence rate had a similar magnitude of impact on risk of *C. diff* as extremes of age and extended length of stay, two risk factors commonly associated with *C. diff* [20]. Incidence rate is often used as a measure of exposure, but this incidence rate does not capture the differences in exposure on a hospital unit level.

Individual medical studies within hospital centers in North America have used genomics and geographical mapping to try to characterize patterns in CDI spread [15] [25]. These studies have confirmed risk factors for CDI, including occupying a room previously held by a CDI patient or sharing a ward with a CDI patient, can

increase patient risk. However, none of these studies compared the differences in exposure on a unit level.

In [15], whole genome sequencing, ward movement, and typing data were used to identify potential CDI donors in six different hospitals. Although colonized patients contributed to transmission, the authors found that CDI cases are more likely to be linked to infected patients than colonized patients, demonstrating the importance of local virulent strains in determining transmission dynamics.

2.2 Risk Prediction

Much technical work in the health-care field has been done in predicting patient risk and identifying patterns of infection spread. In this section, we present and compare early models of infection spread with limited variables and more recent machine learning models with hundreds or even thousands of variables incorporating both patient susceptibility and exposure.

2.2.1 Early Models

The problem of modeling infection spread has been extensively studied in epidemiology. Early mathematical models of infectious diseases, including the Susceptible-Infected-Recovered (SIR) and Susceptible-Infected-Susceptible (SIS) models assume that characteristics of immunity, susceptibility, and recovery do not vary by subgroups, e.g., hospital units [24]. While these models provide important estimates for total number of infections over a period of time, they typically do not give detailed insight into the spread of a disease.

Previous work has relied on manually constructed heuristics that capture patterns in the spread of CDI. These heuristics were developed primarily to be used for estimation of the risk of CDI onset rather than discovery of unit-specific spread patterns. For example, early risk prediction models included hospital CDI incidence level as one of a few dozen variables [5, 6].

In [5, 6], the authors identified symptomatic modified CDI pressure as an independent risk factor for predicting CDI. Patients who were colonized but were not infected were not included in the colonization pressure estimate. The authors found that these estimates were useful independent risk factors. Still, these models did not provide deep insight into transmission dynamics.

2.2.2 Machine Learning Models

In more recent work in the machine learning literature, the authors used a model that uses an estimate of colonization pressure, computed by performing a linear decay of the impact of a new infection over the duration of 14 days. They demonstrated that when combined with thousands of features compiled from electronic health records (EHR) were developed, this exposure estimate is an important risk factor [23, 29, 30].

These authors frame identifying high-risk patients over time as a time-series classification task [29, 30]. They only consider patients whose stays in the hospital last over a couple of days and extract patient risk processes from the EHR in order to model the evolution of patient risk over time. Patient risk processes are constructed by concatenating daily patient risk predictions over time. These daily predictions incorporate hospital and unit colonization pressure estimates based on CDI cases in the hospital and unit previous to that day.

Unlike these efforts, our main aim is not accurate risk estimation. We focus on characterizing patterns in the spread of infection. However, we do show that our method can be used to obtain an estimate of exposure that performs comparable to the state-of-the-art methods.

Chapter 3

Hawkes Process

Temporal point processes are collections of events occurring over time. We say that the process is memory-less if the expected rate of future events is independent of the timing of historical or observed events. The Poisson process is an example of memory-less point processes. By contrast, self-exciting and self-inhibitory temporal point processes model events whose rate depends on the past history of the process. Self-exciting processes describe a set of events where one event encourages the occurrence of future events [12]. Although mathematically, Hawkes processes may be both exciting or inhibitory, we describe these processes as self-exciting for easier reading. Also, we expect infections to incite other infections. Hawkes processes have been used to describe self-exciting phenomena in a variety of fields including finance, geophysical events (i.e., earthquakes), social media, and ecology [32, 1, 11].

3.1 Single Hawkes Process

Formally, a Hawkes process describes a sequence of events in time expressed as :

$$H = \{t_0, t_1, \dots, t_n\},$$

where t_i denotes the time of occurrence of event i . The central equation characterizing the Hawkes process is expressed as

$$\lambda(t) = \lambda_0 + \sum_{t_i \in H_t} g(t, t_i),$$

where λ is the expected infection rate at time t , and λ_0 corresponds to the background rate of CDI caused by exogenous factors, such as the influx of patients who acquire the infection outside the hospital and then get admitted as inpatients on an average day. $g(t, t_i)$ is the time-dependent impact function of the event occurring at time t_i on the expected rate at time t . H_t describes all events $t_i \in H$ where $t_i < t$. It is meant to capture the “self-exciting” component of the process, explicitly modeling the effect of infections on triggering further infections. While the self-exciting component can take on any functional form, in this work we use the most common definition:

$$g(t, t_i) = \nu e^{\frac{-(t-t_i)}{\tau}},$$

where ν is a measure of the influence of one infection happening at time t_i on the expected rate of infection at time t , and τ is a measure of how fast the influence of the event at time t_i decays.

3.2 Network Hawkes Process

In this section, we present a network setting for inferring influence adapted from [10]. In this setting, for each unit, we represent the events occurring in that unit for some number of days before test time as a separate Hawkes process. For example, we let $N_u^0(\cdot)$ represent the counting process of the number of patients with positive test results collected in unit u . We let $N_u^1(\cdot)$ represent the counting process of the number of patients with positive test results who were located in unit u the day before their results were collected. Similarly, we let

$$N_u^d(\cdot)$$

represent the counting process of the number patients with positive test results who were located in unit u d days before their results were collected (d stands for day shift). $N_u^d(\cdot)$ takes as argument an interval $[a, b)$ and returns the number of t day shift patients during that interval.

In our network process, our influence network only contains edges from $N_u^c(\cdot)$ to $N_u^d(\cdot)$ where $c \leq d$. In other words, patients c days before positive test results can only affect patients d days before positive test results if c is less than d . If a patient is infected by another patient, we expect the first patient to be diagnosed first. This may not always be true, but is a reasonable assumption.

We express the stochastic rate for a unit with day shift d as follows

$$\begin{aligned} \lambda^d(t) &= \lambda_0^d(t) + \sum_{c \leq d} \int_0^{t^-} g^{(cd)}(t, t') dN^c(t') \\ &= \lambda_0^d(t) + \sum_{c \leq d} \sum_{t' < t} g^{(cd)}(t, t') \end{aligned}$$

Similarly to the single Hawkes process, $\lambda^d(t)$ represents expected infection rate at time t with day shift d . $\lambda_0^d(t)$ represents the background rate of the Hawkes process with day shift d . For all patients with day shift c , where $c \leq d$, we consider $g^{(cd)}(t, t')$ or the time-dependent impact function of the event occurring at time t' on the expected rate at time t . Analogously to the single Hawkes process, we let

$$g^{(cd)}(t, t') = \nu^{(cd)} e^{\frac{-(t-t')}{\tau^{(d)}}},$$

We keep the rate of decay of the impact of each day shift process the same, specific to the day shift process d . Note that the single Hawkes process can be represented as a single node network Hawkes process. We present the inference algorithm for this network Hawkes model in the next chapter.

Chapter 4

Learning the Magnitude and Duration of Influence

In this section, we discuss our main method for estimation of the magnitude and duration of influence. We present how the data are extracted, details of our method, and main results. We provide both simulated and real results.

4.1 Event extraction

Some of our analyses are conducted using electronic health records from a large urban hospital. We considered all patients who visited the hospital anytime between January 1, 2013 and June 1, 2014. We examined all microbiology lab tests and flagged a CDI event whenever a patient tested positive for CDI. During the relevant timer period, the hospital had a tiered testing system based on two assays. The first test was a combined glutamate dehydrogenase (GDH) and toxin enzyme immunoassay (EIA). If both components were positive, the patient was considered positive and no further testing was done. If both components were negative, the patient was considered negative. In the instance of discordance, a reflex polymerase chain reaction (PCR) test was done as the “tie-breaker.”

We refer to the date of positive CDI sample collection date as the positive test date. However, it is likely that the date of exposure occurs some time before the onset

of symptoms and sample collection. For that reason, we conduct our analyses using *shifted* days. Specifically, we denote the k -day-shifted date as the date k days before the positive test date and run our analysis separately using 0-day shift (meaning taking the exposure date to be the positive test date) to 3-day shift (meaning taking the exposure date to be three days prior to the positive test date). The location of the event, meaning unit where the event happened, is taken to be the unit where the patient was located on the k -day-shifted date. If a patient was located outside the hospital k days before the positive test date, they are not considered for the k -day-shifted date analysis. If a patient was located in multiple units during that day, for example being transferred between rooms, we consider the patient to be located in the unit where they spent the most time. We split the data into training and testing temporally, taking 2013 data to be the training data (number of events = 726) while January to June 2014 is the testing data (number of events = 348).

4.2 Methods

We treat each unit in the hospital and the hospital as whole as separate Hawkes processes, learning independent parameters for each location. We use the training data to compute the Maximum Likelihood Estimates (MLE) of the model parameters.

4.2.1 Single Hawkes Process Inference

For the inference method, we use the inference algorithm described in subsection 4.2.2 of this paper with a single agent.

Like most other time stamps in EHRs, the exact time of the positive CDI test is uncertain. In order to account for that uncertainty, we try two methods: perturbing the data and bootstrapping our results.

For the first method, we perturb the data by a sampling each event from a uniform distribution around the hour at which the sample was collected for each unit. We perturb and compute Hawkes parameters for each unit 100 times. This process gives us a distribution of parameters and lets us determine how robust our model remains

to variation. We select the mean parameters of these 100 runs to be the final MLE parameters of each unit model.

For the second method, we split the unit time series into 100 equally spaced, overlapping intervals of half of the size each. We then compute the Hawkes parameters for each interval and determine the distribution over parameters. We select the mean parameters of these 100 runs to be the final MLE parameters of each unit model.

4.2.2 Network Hawkes Process Inference

In this section, we provide the appropriate priors and an inference algorithm used for our variant of the network Hawkes process described in the previous chapter [10]. For each unit, we have the day shifted times

$$\mathbb{T} = \{\mathbb{T}^{(d)}\}_{d=1}^D$$

but the parameters

$$\Theta = \{\lambda_0^{(d)}, \{\nu^{(cd)}\}_{c \leq d}, \tau^{(d)}\}_{d=1}^D$$

are unobserved. We learn these parameters by estimating their posterior distribution through Bayes' Theorem

$$P(\Theta|\mathbb{T}) \propto P(\mathbb{T}|\Theta)P(\Theta)$$

We compute likelihood as follows

$$P(\mathbb{T}|\Theta) = \prod_{d=1}^D (\exp(-\Lambda^d(T)) \prod_{i=1}^{N^{(d)}(T)} \lambda^d(t_n^d))$$

where

$$\Lambda^d(T) = \int_0^T \lambda^d(t) dt$$

represents the expected number of infections over time \mathbb{T} [4].

We place an improper prior over $\lambda_0^d > 0$, such that the base rate of probability of an infection per day in a unit is greater than 0. An improper prior is a prior whose distribution does not integrate to 1. We use this improper prior to allow for our parameters to make sense. We cannot have a negative expected rate of infection. We also use another prior to ensure that the network Hawkes process is stationary. We use the stationarity condition if M is a $D \times D$ matrix

$$M^{(cd)} = \int_u^\infty |g^{(cd)}(t, u)| dt = \nu^{(cd)} \tau^{(d)}$$

then the spectral radius of M must be less than 1 [2]. Since we can upper-bound the spectral radius by any norm, we use the maximum absolute column sum norm

$$\|M\|_{1 \rightarrow 1} = \max_{\|x\|_1=1} \|Mx\|_1 = \max_{d=1, \dots, D} \tau^{(d)} \sum_{c \leq d} \nu^{(cd)}$$

to construct improper joint priors over $\{\tau^{(d)}\}_{d=1}^D$ and $\{\{\nu^{(cd)}\}_{c \leq d}\}_{d=1}^D$

$$0 < \tau^{(d)} < \frac{1}{\sum_{c \leq d} \nu^{(cd)}}$$

$$0 < \nu^{(cd)} < \frac{1}{\tau^{(d)} - \sum_{r \neq c, r \neq d} \nu^{(rd)}}$$

The resultant distribution $P(\Theta|T)$ is analytically intractable, but we can use the conditional intensity function to draw posterior samples. We use a slice-within-Gibbs algorithm to sequentially sample parameters from its conditional posterior [22]. Like Guo et al., we reduce the computational cost by computing the product over rate functions with the following recurrence

$$\begin{aligned} \lambda^d(t_n^d) &= \lambda_0^d + (\lambda^d(t_{n-1}^d) - \lambda_0^d) \exp\left(-\frac{t_n^d - t_{n-1}^d}{\tau^{(d)}}\right) \\ &+ \sum_{c \leq d} \sum_{m: t_{n-1}^d \leq t_m^c \leq t_n^d} \nu^{(cd)} \exp\left(-\frac{t_n^d - t_m^c}{\tau^{(d)}}\right) \end{aligned}$$

for $n = 2, 3, \dots, N^{(d)}(T)$. The initial term can be computed

$$\lambda^d(t_1^d) = \lambda_0^d + \sum_{c \leq d} \sum_{m: t_m^c \leq t_1^d} \nu^{(cd)} \exp\left(-\frac{t_1^d - t_m^c}{\tau^{(d)}}\right)$$

4.3 Results

In this section, we first present the likelihoods of modeling top CDI incidence units with Poisson and Hawkes processes. We then present simulations of retrieving Hawkes parameters when we vary base rate and influence, and determine the Hawkes parameters when we model each unit with time decay as a single Hawkes process. We then add a constraint in which patients who cause other patients to get infected are diagnosed first, which is not taken into consideration by a single Hawkes process. We propose a network Hawkes process as a better model and analyze learned network Hawkes processes on different types of units.

4.3.1 Poisson vs. Hawkes Processes

Before delving into the Hawkes process results, we first seek to ensure that the data supports our hypothesis; meaning we wish to examine whether or not the events do follow a self-exciting Hawkes process. An alternative model that might explain the data is a simple memory-less Poisson process, which assumes that the incidence of a new CDI case is independent of previous contiguous events, i.e., the inter-arrival rate is independent. To test our hypothesis, we compared the log likelihood (LL) of the test data under each process and the p-value of using the Hawkes process over the Poisson process is computed. For each process, we train a Hawkes model and a Poisson model on the first half of the data in time, roughly January 2013 to September 2013. We then compute the log-likelihood on the remainder of the data. For the Poisson parameters, the rate is invariant and is equivalent to the average number of events expected to occur per day. Table 4.2 shows the unit name, the number of events, the log likelihood under the Hawkes process and under the Poisson process and the p-value of using the Hawkes process over the Poisson process for a 3 day shift. We see that the log likelihood is rarely higher for any of the units under the Poisson

process and that units with higher incidence rates seem to be better explained by a Hawkes process. Results from the two and zero day shift conform with the findings here, and are presented in the appendix. We visualize the events per day over all of our data in the entire hospital in Figure 4-1. We also anonymize all units based upon characteristics described in Table 4.1.

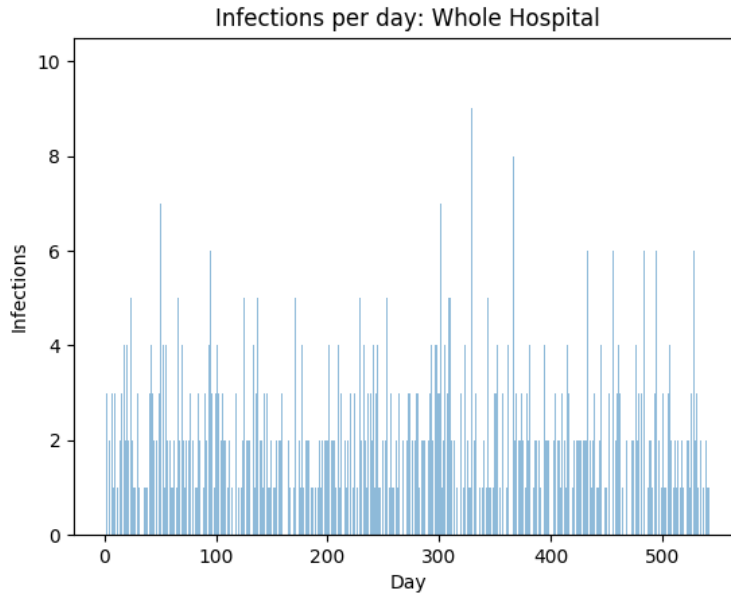


Figure 4-1: The number of CDI per day for the entire hospital.

Table 4.1: Anonymized unit acronyms and characteristics.

Unit Acronym	Unit Description	Unit Characteristics
EMD	Emergency Department	bays and single rooms
GMU	General Medical Unit	mix private and semi private rooms
MICU	Medical Intensive Care Unit	all private
MU	Medical Unit	all private
OU	Oncology Unit	all private

Table 4.2: Test log-likelihood of Hawkes vs Poisson model for different units, 3 day shift. Note that log likelihood values closer to 0 is better.

Unit	Number of Events	Hawkes LL	Poisson LL	p-value (2-tailed)
Whole Hospital	1074 (548 in test)	-165.94	-276.97	< 0.00001
EMD	115 (63)	-156.90	-169.81	< 0.00001
MU	73 (40)	-117.34	-125.96	0.00018
OU A	62 (32)	-100.36	-106.99	0.0013
MICU A	53 (20)	-75.23	-73.74	0.225
GMU C	46 (18)	-68.38	-67.90	0.619
OU B	43 (21)	-75.28	-78.47	0.041
GMU B	42 (25)	-85.50	-91.04	0.0039
GMU A	40 (22)	-78.05	-80.92	0.057
GMU D	40 (22)	-79.78	-82.58	0.0608
MICU B	37 (15)	-59.63	-59.88	0.78

4.3.2 Single Hawkes Simulations

Before working further with the data, we explore simulations in order to test how robust our model is to ranges of parameters.

For a given set of parameters, we simulate 1000 time-steps worth of data. After simulating the data, we use Maximum Likelihood Estimation to recover the most likely parameters from the simulated data. We repeat the simulation and recovery 50 times, and plot the range of recovered parameters as a box-plot.

In our simulations, we test how well we recover parameters when we vary ν and when we vary λ_0 . In our first set of simulations, shown in Figure 4-2 and Figure 4-3, we set τ to be 0.1, the λ_0 to be 0.1 and vary ν across the set [0.0001, 0.001, 0.1, 0.5, 1].

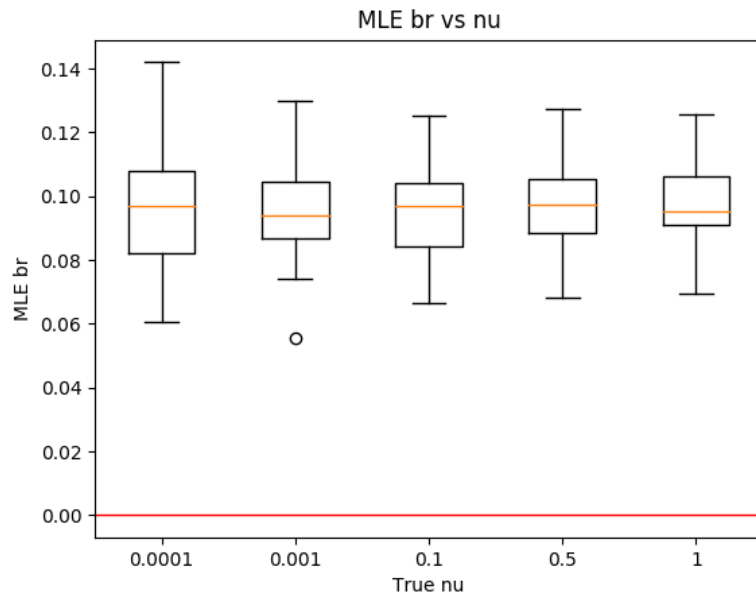


Figure 4-2: Recovering base rate after varying influence.

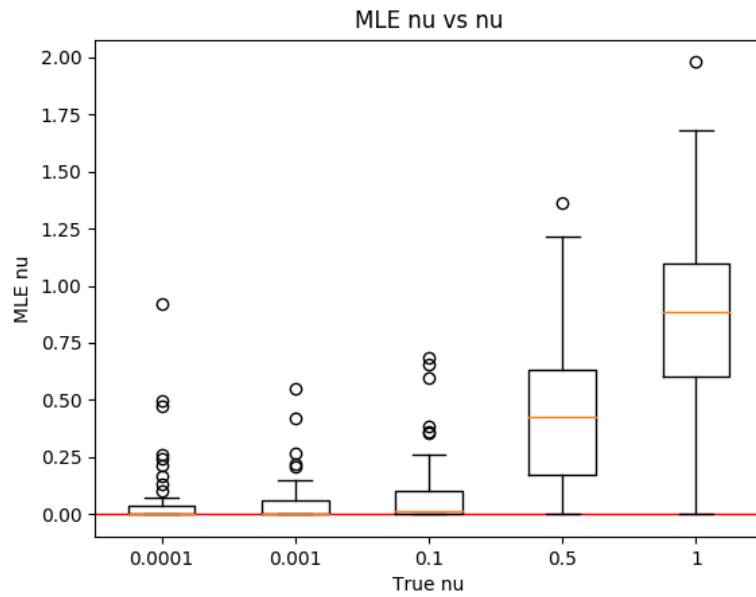


Figure 4-3: Recovering influence after varying influence.

From Figure 4-2 and Figure 4-3, we see that even if events have a strong effect on the rate, we are still able to recover the base rate and the influence parameters with confidence. In the next set of simulations, shown in Figure 4-4 and Figure 4-5, we set τ to be 0.1, the ν to be 0.1 and choose λ_0 from the set $[0.0001, 0.001, 0.01, 0.1, 0.5]$.

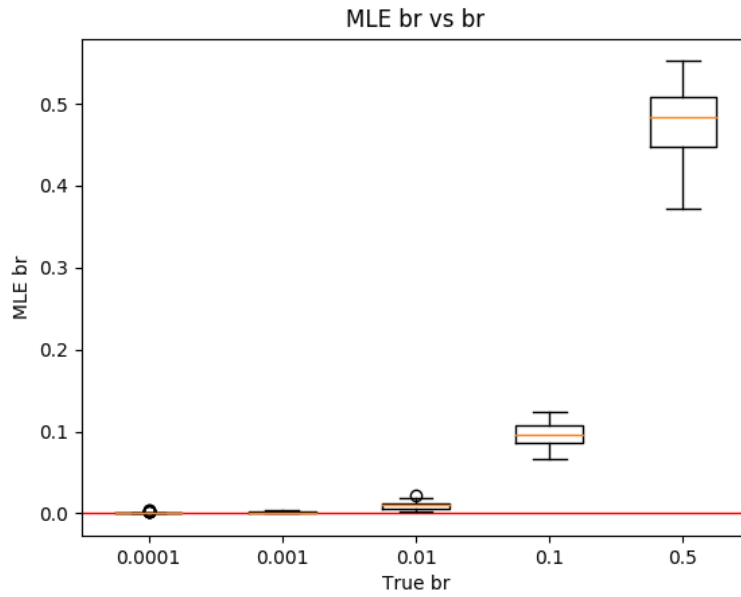


Figure 4-4: Recovering base rate after varying influence.

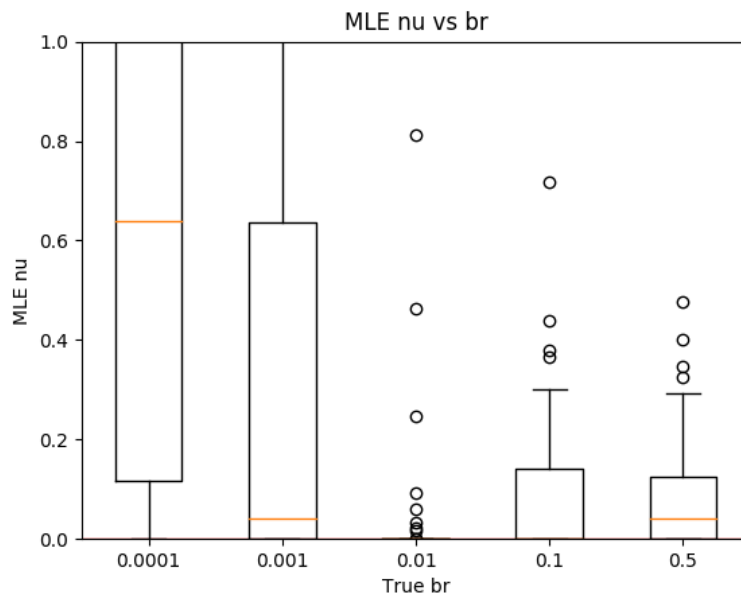


Figure 4-5: Recovering influence after varying influence.

From Figure 4-4 and Figure 4-5, we see that if the base rate is very low, the MLE for the influence parameter is uncertain. This makes sense because if the base rate of the Hawkes process is low, we expect very few events to occur. Because of the lack of data and incidences, it becomes harder to predict the effect of one event on another. Because of this, we decide to look at wards with the most incidences as best candidates for our Hawkes model. For wards with few incidences in the data, we expect our model to be less accurate.

4.3.3 Learning Single Hawkes Processes

Having established that the data seem to support the Hawkes process hypothesis, we next explore the learned magnitude and duration of influence. We present two methods to learn the mean parameters for a single unit day shifted Hawkes process: by perturbing the data or by bootstrapping. We present results from both methods below, plotting units by mean learned influence and decay parameters.

The magnitude is captured by ν , which denotes that instantaneous increase in rate of infections per day when a new infection occurs. Large and positive values of ν imply that a new infection raises the expected rate of infections in the following time step. The duration of influence is captured by the τ , which denotes the decay in the influence of an infection. Higher values of τ suggest that the influence of a new infection decays fast, signaling that a new infection does *not* have a long lasting influence.

Figure 4-6 shows the values of the influence parameter ν on the x -axis and the decay parameter τ on the y -axis for one day shift by perturbing the data. Figure 4-7 shows the same plot for 3 day shift with bootstrapping. Blue represents GMUs, green represents individual units, black represents OUs, and red represents the EMD. Note that both of these plots show similar patterns. Plots for zero, two, and three day shifts are given in the Appendix.

In general, we find that perturbing the data and bootstrapping produce similar patterns of influence and decay across units. In general, we find that oncology units have long decay and low influence, which corresponds to our intuition that these units

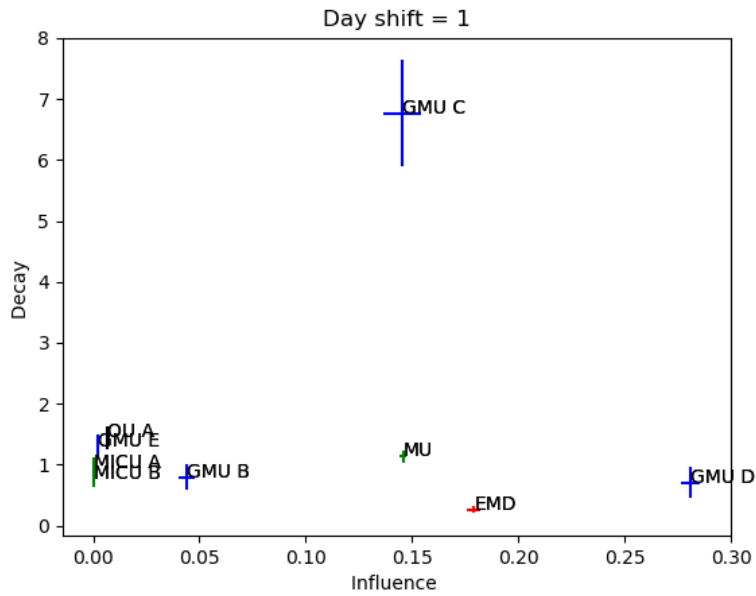


Figure 4-6: Units plotted by learned influence and decay with error for one day shift with perturbing the data.

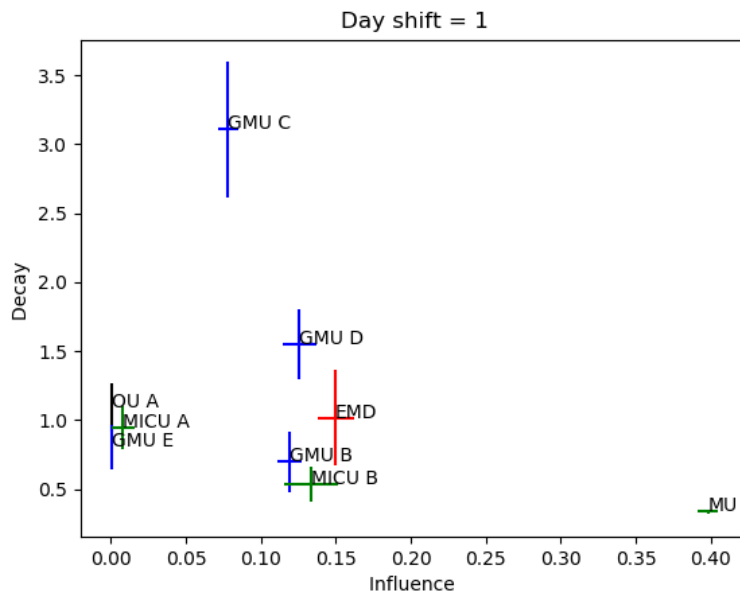


Figure 4-7: Units plotted by learned influence and decay with error for one day shift with bootstrapping.

have good isolation policies but that the effect of an exposure on immunocompromised patients lasts for a longer time. Private room units also generally have low influence being a notable exception. However, the MU unit, where many CDI patients are moved to, is a notable exception.

Interestingly, we find that infections in the emergency room tend to have a high influence and a low decay rate, meaning they trigger further infections and their effect lasts for a long time. This implies that controlling infections in the emergency room should be of the highest priority for purposes of infection control. The fact that the emergency room is a key culprit in the spread of infection does not come as a surprise since it is a location where there is high patient turnover and where cleaning tends to be difficult because there is a constant influx of patients and issues of overcrowding.

4.3.4 Multiple Processes Analysis

One problem that we noticed in our single day shift unit Hawkes processes is that we only consider the situation in which a patient in a unit t days before testing affects another patient in that unit t days before testing. The more likely scenario is that the patient who causes the second patient to get infected is probably going to be diagnosed first.

In order to demonstrate that our data supports this hypothesis, we consider all of the events that happen within a unit, 0, 1, 2, and 3 day shifted. We report the proportion of s day shifted events that have one or more t day shifted events that occur in the three days before. We expect to see a higher proportion of 1 or 2 day shifted events before a 3 day shifted event because we assume that patients more advanced in the disease but not yet diagnosed will likely spread spores to new patients.

In the following tables in entry (t, s) , we report the proportion of s day shifted events with t day shifted events that occur before for an oncology unit, the emergency department, a general medical unit, and a medical unit. We also provide a table of the average of all four units in Table 4.8. For reference, we also provide the number of day shifted events in each unit in Table 4.3.

Note too that in the Table 4.8, the values above the diagonal are larger than the

Table 4.3: Number of shifted events per unit

Days shifted	OU A	EMD	MU	GMU A
0 day shift	66	27	84	52
1 day shift	59	60	50	33
2 day shift	45	12	22	24
3 day shift	40	19	17	18

Table 4.4: Unit OU A proportion of shifted events

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	0.23	0.25	0.29	0.3
1 day shift	0.22	0.25	0.29	0.3
2 day shift	0.17	0.19	0.2	0.2
3 day shift	0.14	0.15	0.16	0.15

Table 4.5: Unit EMD proportion of shifted events

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	0.11	0.12	0.08	0.26
1 day shift	0.37	0.32	0.17	0.37
2 day shift	0.07	0.07	0	0.05
3 day shift	0	0.13	0.08	0.11

Table 4.6: Unit MU proportion of shifted events

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	0.36	0.46	0.41	0.47
1 day shift	0.25	0.34	0.32	0.35
2 day shift	0.15	0.16	0.18	0.18
3 day shift	0.11	0.12	0.18	0.18

Table 4.7: Unit GMU A proportion of shifted events

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	0.33	0.33	0.28	0.28
1 day shift	0.21	0.24	0.25	0.17
2 day shift	0.17	0.19	0.17	0.11
3 day shift	0.13	0.12	0.13	0.11

Table 4.8: Unit average proportion of shifted events

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	0.26	0.29	0.27	0.33
1 day shift	0.26	0.29	0.26	0.30
2 day shift	0.14	0.15	0.14	0.12
3 day shift	0.10	0.13	0.14	0.14

ones below. This signifies that our hypothesis that contagious patients spread the disease to new patients does appear to hold. Patients who are less advanced in the disease have a larger proportion of patients more advanced in the disease preceding them.

4.3.5 Learning Network Hawkes Processes

Given that our assumption appears to hold from the previous section, we then learn network Hawkes processes for different day-shifts. We do this to see if we can determine the influence of different day shifts on each other. For example, we compute how much influence individuals 1 day before their positive test date have on individuals 3 days before their positive test date in a unit.

Since we believe that individuals who are more advanced in the disease are more likely to spread the infection, we enforce the entries below the diagonal in the influence matrix to be 0 during our Gibbs sampling process. For reference, the single Hawkes parameter experiments can be represented as learning different processes for each of the entries along the diagonal in the influence matrix. We report the influence, base rate, and decay parameters for several units in the following sections.

Unit OU A

Table 4.9: Unit OU A base rate parameters

0 day shift	1 day shift	2 day shift	3 day shift
0.12	0.11	0.08	0.07

Table 4.10: Unit OU A decay parameters

0 day shift	1 day shift	2 day shift	3 day shift
0.18	0.15	0.16	0.12

Table 4.11: Unit OU A influence parameters

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	1e-6	1e-6	1e-6	1e-6
1 day shift	-	1e-6	1e-6	1e-6
2 day shift	-	-	1e-6	1e-6
3 day shift	-	-	-	1e-6

We expect oncology units to have low influence parameters. Since patients are well isolated, we expect that patients are unlikely to spread the infection to other patients. In the data, we see that the influence parameters are very small, $1e - 6$.

Unit EMD

Table 4.12: Unit EMD base rate parameters

0 day shift	1 day shift	2 day shift	3 day shift
0.06	0.10	0.02	0.05

Table 4.13: Unit EMD decay parameters

0 day shift	1 day shift	2 day shift	3 day shift
1.0	1.3	0.17	3.9

Table 4.14: Unit EMD influence parameters

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	1e-6	1.8e-2	1.8e-2	1.9e-2
1 day shift	-	2.3e-2	1e-6	1.3e-2
2 day shift	-	-	1e-6	1e-6
3 day shift	-	-	-	6.5e-3

We note that the influence of 0 day shift patients is high on 1, 2, and 3 day shift patients. This means that even on the day of lab testing, patients who are tested in the EMD still have significant influence on later infections. Since many patients pass through the EMD, making it difficult to isolate patients, this influence makes sense. However, in comparison to the other units, the decay parameters are larger. This means that the influence of an infection decays faster. Since patients don't stay long in the EMD, this fast decay makes sense.

Unit MU

Table 4.15: Unit MU base rate parameters

0 day shift	1 day shift	2 day shift	3 day shift
0.15	0.08	0.03	0.02

Table 4.16: Unit MU decay parameters

0 day shift	1 day shift	2 day shift	3 day shift
0.7	0.4	0.7	0.6

Table 4.17: Unit MU influence parameters

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	1e-6	1e-6	1e-6	1e-6
1 day shift	-	1.3e-1	1.7e-2	1.4e-2
2 day shift	-	-	2.7e-2	1e-6
3 day shift	-	-	-	6.0e-2

Although Unit MU has around the same number of expected total infections per day as Unit OU, the influence parameters are very different. Patients on the day of lab testing have low influence of consequent infections. This makes sense since once lab results are collected, patients suspected of infection are isolated. These results also suggest that isolation is well implemented in Unit MU. In comparison to the EMD unit, patients the day before infection have the highest influence on subsequent infections.

Unit GMU A

Table 4.18: Unit GMU A base rate parameters

0 day shift	1 day shift	2 day shift	3 day shift
0.07	0.03	0.04	0.03

Table 4.19: Unit GMU A decay parameters

0 day shift	1 day shift	2 day shift	3 day shift
2.0	2.4	0.17	0.11

Table 4.20: Unit GMU A influence parameters

	0 day shift	1 day shift	2 day shift	3 day shift
0 day shift	0.4e-2	0.4e-2	1e-6	1e-6
1 day shift	-	1e-6	1e-6	1e-6
2 day shift	-	-	1.3e-1	1e-6
3 day shift	-	-	-	1.9e-1

Although the general medical units tend to differ slightly in terms of ward layout, we select Unit GMU A for analysis. Patients 2 and 3 days before lab testing have a larger effect on resulting infections than patients 1 or 0 days before lab testing, but otherwise, influence parameters are low. This difference may be because once patients begin showing symptoms of infections, GMU A does a good job of isolating patients.

Chapter 5

CDI Risk Prediction

In this section, we validate that the Hawkes process is able to capture meaningful patterns in the spread of infection by demonstrating that the learned expected rate of infection is useful as a measure of patient exposure to CDI. To do so, we test the performance of our learned exposure measure when combined with patient characteristics to predict the onset of CDI during their hospital admission.

5.1 Cohort

We replicate the cohort definition and data extraction process used in [23]. Specifically, the study cohort consists of adult inpatients admitted to the hospital between January 1, 2013, and June 1, 2014. We focus on suspected nosocomial, i.e., healthcare-associated cases by excluding patients who tested positive for CDI within the first two calendar days of admission. We extracted variables from the electronic health records that capture patient susceptibility to infection. Briefly, we split variables into 2 main categories: (1) time invariant and (2) time varying. Time-invariant variables are available at the time of admission and do not change over the course of the admission. These variables include patient demographics (e.g., gender), statistics on encounter history (e.g., number of inpatient admissions in last 90 days), and treatment and diagnoses associated with the most recent previous hospitalization. Time-varying variables, extracted daily for each patient, included laboratory results,

procedure codes, medications, and vital signs collected during the hospitalization. We mapped all data to binary values as in [23].

The prediction task is to predict whether or not a patient will acquire CDI during his/her inpatient visit. We extract the labels for each visit according to whether the patient was diagnosed with CDI during that visit. The learning task was to predict in advance of clinical diagnosis which patients would be diagnosed with CDI. We labeled each day from a CDI case as positive, and negative otherwise.

5.2 Methods

We constructed a daily estimate of exposure equal to the daily expected rate of infection of the learned Hawkes process for 0 day shift. To compute this estimate, we learn the parameters using the training data, and compute the expected rate of infection for each unit, and each day t in the testing set using event data up to time $t - 1$. We emphasize that the values of the parameters are not updated using the test data, but the conditional expectation of the event rate is computed using all the events up to time $t - 1$. We refer to our exposure estimate as the Hawkes exposure estimate (HE). We compute the HE estimate for each unit (HE-Unit) and the hospital as a whole (HE-hospital). For the unit-level Hawkes rates, we only compute Hawkes rates for the top 9 wards ranked by infection incidence. For each unit, we transform the computed rates into quintiles and include each quintile as a binary feature variable in the prediction tasks.

As a benchmark, we compute the exposure measure used in [23]. In that work, the authors use a simple heuristic to capture the magnitude and duration of the influence of an infection. They assume that the influence does not change based on unit and define an estimate for colonization pressure, or the average daily proportion of patients colonized. They assume that a linear decay model over a 14-day period; meaning on day 1 of the infection it contributes to the colonization pressure by 1, on day 2 it contributes by $1/2$, on day t it contributes by $\frac{1}{t}$ and the contribution is 0 on day 14. We refer to this model as the state-of-the-art exposure estimate (SE) and use

it as the main benchmark. Similar to the HE estimates, we compute the SE estimate for each unit (SE-Unit) and the hospital as a whole (SE-hospital), transform them into quintiles and include them as variables in the prediction tasks.

For the prediction model, we applied multitask L2-regularized logistic regression with class balancing to produce models for the institution that were used to generate daily estimates of patient risk for the top 9 units ranked by infection incidence and for the entire cohort. We smoothed daily risk scores by averaging over time, a previously validated approach [23, 30]. To learn and evaluate the model, the data were split temporally: using the first year for training data and the last six months as testing data. A temporal split was used because it provides a better estimate of prospective performance than a random split. In addition, from the training data, we excluded data pertaining to the sample collection date and the preceding day for positive cases. This approach prevented the model from using empiric CDI therapy as a factor in predicting CDI.

5.3 Results

We computed 95% empirical bootstrap confidence intervals (CI) for the AUROC using parameters computed from 100 bootstraps of the training set. Using a decision threshold based on the 95th percentile, we classified patients as “predicted CDI positive” or “predicted CDI negative” depending on whether or not they crossed that threshold. Tables in the following sections show the description of the exposure measure used, the number of days in advance of CDI onset that we’re able to predict the infection at the 95th percentile threshold, and the test-set AUROC for various selected testing cohorts. We find that the models incorporating the Hawkes process-based exposure measures have comparable performance to those incorporating the state-of-the-art exposure measure across all cohorts. However, all of the models, including the one without a measure of colonization pressure perform similarly. One reason for this may be that exposure is somehow also being caught by a combination of the thousands of other variables in our model.

5.3.1 Cohort: Entire Hospital

The first cohort we test on are patients in the described cohort from the entire hospital from January 1, 2014 to June 1, 2014. This is the same cohort as described in [23]. Table 5.1 contains results for models run with L2-regularization and class-balancing. We also ran models with L1-regularization and no class-balancing. These results are in the Appendix. The AUROC results are better for models trained with L1-regularization, but the number of days in advance of CDI onset that we’re able to predict the infection at the 95th percentile threshold decreases.

Table 5.1: Cohort results for the whole hospital. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.

Dataset	Early Detection Days	AUROC Test
No Exposure	6.0	0.76 (0.73-0.79)
SE unit	6.0	0.76 (0.74-0.79)
HE unit	6.0	0.76 (0.74-0.79)
SE unit, SE hosp	5.5	0.77 (0.74-0.79)
HE unit, HE hosp	6.0	0.76 (0.74-0.79)
SE unit, HE hosp	6.0	0.76 (0.74-0.79)
HE unit, SE hosp	6.0	0.77 (0.74-0.79)

5.3.2 Cohort: Top 9 Units

Since we only compute the Hawkes rates for the top 9 units sorted by incidence, we also tested reducing our cohort to incidences that occurred in these top 9 units. Table 5.2 contains results for models run with L2-regularization and class-balancing. These results are worse than Table 5.1 because we lose many true negatives when we drop all low-incidence units. Results for models with L1-regularization and no class-balancing are shown in the Appendix.

Table 5.2: Cohort results for the top 9 units. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.

Dataset	Early Detection Days	AUROC Test
No Exposure	4.0	0.63 (0.58-0.68)
SE unit	3.0	0.62 (0.57-0.67)
HE unit	4.0	0.63 (0.58-0.68)
SE hosp	4.0	0.65 (0.61-0.70)
HE hosp	4.0	0.63 (0.59-0.68)
SE unit, SE hosp	4.0	0.65 (0.61-0.70)
HE unit, HE hosp	4.0	0.63 (0.59-0.68)
HE unit, SE hosp	4.0	0.65 (0.61-0.70)

5.3.3 Pairwise Analysis

Finally, we perform pairwise analysis on the days in advance before CDI onset that we're able to predict the infection for the datasets of the top 9 units combined with both SE unit and HE unit. We do this pairwise analysis by the fraction of true positives for the number of days predicted in advance. The results are shown in Figure ?? and Figure ?. From this figure, we see that the distributions are similar, but we are able to prediction a larger fraction of true positives more days in advance for the dataset combined with HE unit.

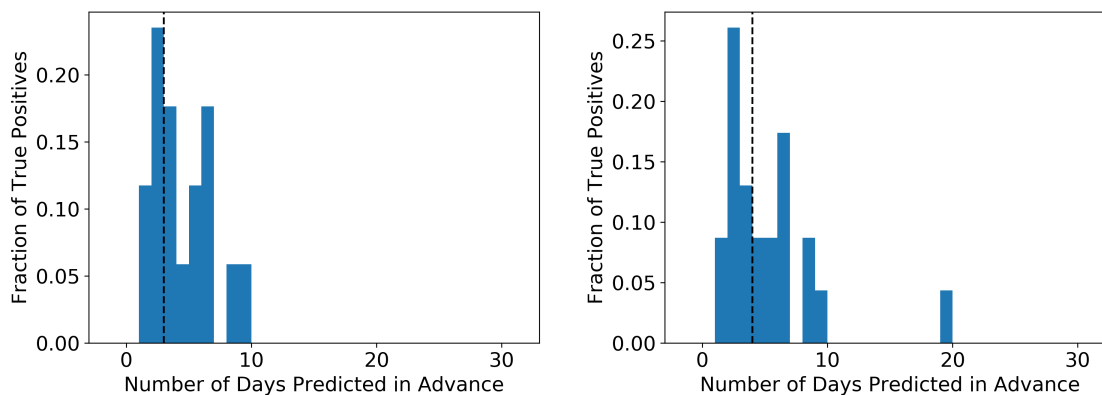


Figure 5-1: We plot the fraction of true positives for the number of days predicted in advance with SE unit (left) and HE unit (right).

Chapter 6

Discussion

We presented a method to estimate the influence of new CDI cases on triggering further infections. Our method separated the magnitude of the influence from the duration of the influence. We showed that different units in the hospital respond differently to new infections, which suggests that unit-specific infection control policies can be tailored to control the spread of infections. We find that the models incorporating the Hawkes process-based exposure measures have comparable performance to those incorporating the state-of-the-art exposure measure. However, all of the models, including the one without a measure of colonization pressure perform similarly. One reason for this lack of improvement may be that exposure is somehow also being caught by a combination of the thousands of other variables in our model.

We modeled each unit and the hospital as a whole as separate network Hawkes processes. Future work should address this task as a shared learning process, allowing different units to have different spread patterns but also accounting for the fact that there are shared patterns across all units. In addition, we made the simplifying assumption that only patients who tested positive for CDI contribute to the spread of infection. However, it is believed that colonized patients who did not test positive for CDI might still spread the infection [19, 21, 3]. These are referred to as asymptomatic carriers. Future work could incorporate the notion that these asymptomatic carriers act as unobserved events.

To our knowledge, this is the first attempt at using self-exciting point processes

to model the spread of CDI. We hope that the work presented here sets the stage for further inquiry into the nature of the spread of infection and optimal policies to curtail the spread.

Appendix A

Tables

Table A.1: Test log-likelihood of Hawkes vs Poisson model for different units, 0 day shift

Unit	Number of Events	Hawkes LL	Poisson LL
Whole Hospital	1074 (548 in test)	-166.02	-277.46
GMU B	50 (29)	-94.33	-100.87
GMU A	52 (27)	-93.15	-94.26
GMU E	42 (18)	-66.35	-67.73
MICU A	52 (23)	-80.05	-82.50
MU	84 (44)	-124.50	-133.31
GMU C	49 (20)	-73.60	-73.65
GMU D	45 (23)	-81.01	-84.42
OU A	66 (34)	-104.57	-111.58
OU B	48 (23)	-80.92	-83.64
GMU F	45 (25)	-86.15	-90.95

Table A.2: Test log-likelihood of Hawkes vs Poisson model for different units, 2 day shift

Unit	Number of Events	Hawkes LL	Poisson LL
Whole Hospital	1074 (548 in test)	-165.94	-276.97
EMD	96 (50 in test)	-135.27	-144.98
GMU A	43 (22 in test)	-77.90	-80.08
GMU E	40 (16 in test)	-61.64	-61.65
MICU A	55 (22 in test)	-79.46	-79.47
MU	75 (40 in test)	-116.67	-125.38
GMU C	45 (18 in test)	-68.03	-68.04
GMU D	41 (23 in test)	-82.72	-85.56
OU A	62 (32 in test)	-100.36	-106.99
OU B	43 (21 in test)	-75.28	-78.47
GMU B	43 (26 in test)	-87.98	-93.9

Table A.3: Cohort results for the whole hospital. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.

Dataset	Early Detection Days	AUROC Test
No Exposure	5.0	0.77 (0.75-0.79)
SE unit, SE hosp	5.0	0.78 (0.75-0.80)
HE unit, HE hosp	5.0	0.77 (0.75-0.80)

Table A.4: Cohort results for the top 9 units. The model is run with no class balancing and L1-regularization. The AUROC test 95% confidence interval is reported.

Dataset	Early Detection Days	AUROC Test
SE unit	5.0	0.66 (0.62-0.71)
HE unit	5.0	0.66 (0.62-0.71))
SE hosp	5.0	0.69 (0.63-0.73)
HE hosp	5.0	0.66 (0.61-0.71)
SE unit, SE hosp	5.0	0.68 (0.63-0.73)
HE unit, HE hosp	5.0	0.66 (0.61-0.71)
HE unit, SE hosp	5.0	0.68 (0.63-0.73)

Table A.5: Cohort results for the top 9 units with location removed. The model is run with class balancing and L2-regularization. The AUROC test 95% confidence interval is reported.

Dataset	Early Detection Days	AUROC Test
No Exposure	4.0	0.63 (0.58-0.68)
SE unit	4.0	0.63 (0.58-0.68)
HE unit	4.0	0.63 (0.58-0.67)

Appendix B

Figures

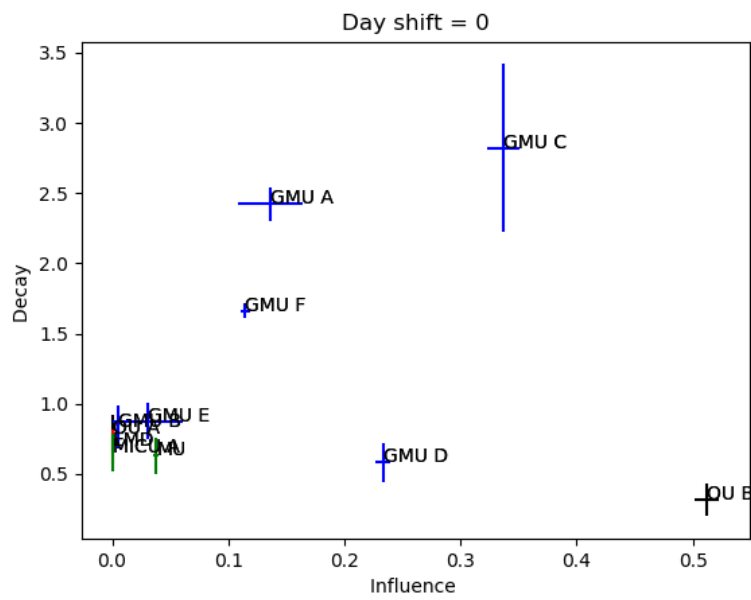


Figure B-1: Units plotted by learned influence and decay with error for zero day shift with perturbing the data.

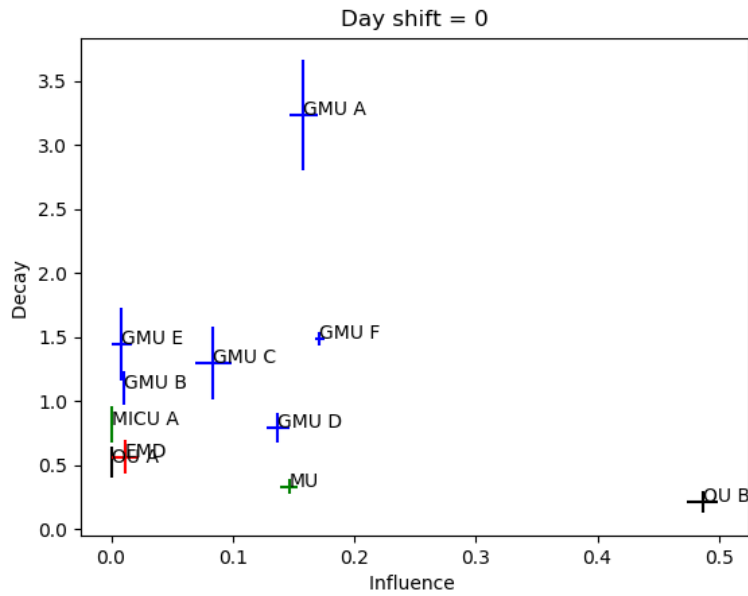


Figure B-2: Units plotted by learned influence and decay with error for zero day shift with bootstrapping.

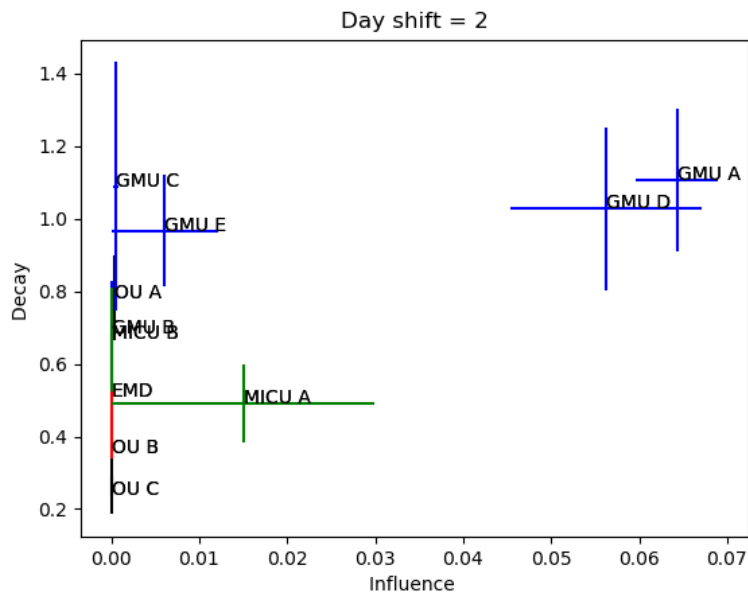


Figure B-3: Units plotted by learned influence and decay with error for two day shift with perturbing the data.

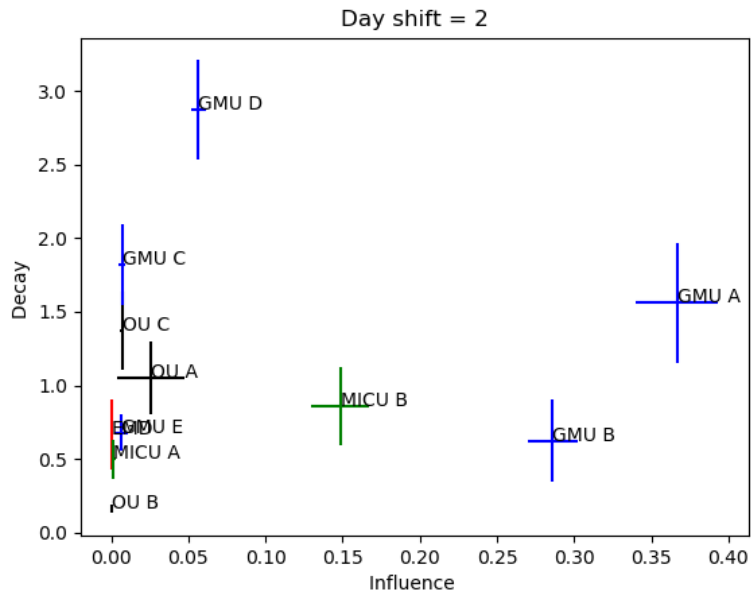


Figure B-4: Units plotted by learned influence and decay with error for two day shift with bootstrapping.

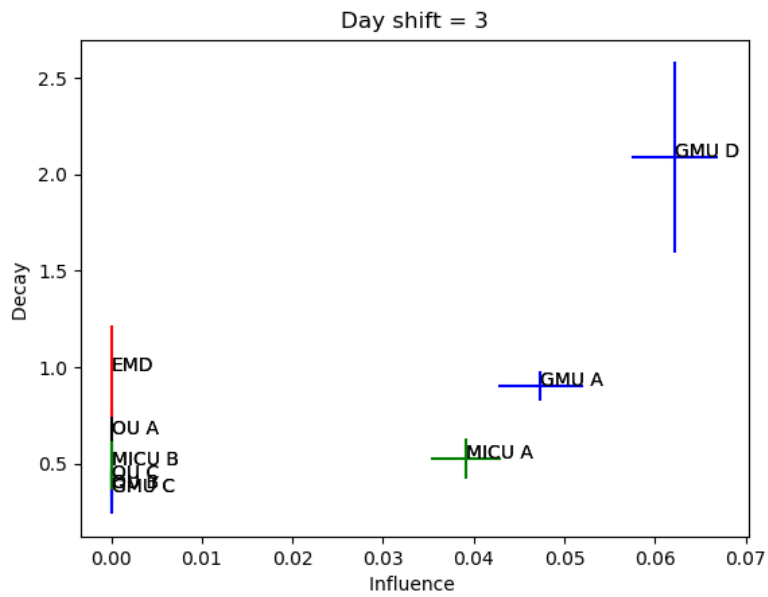


Figure B-5: Units plotted by learned influence and decay with error for three day shift with perturbing the data.

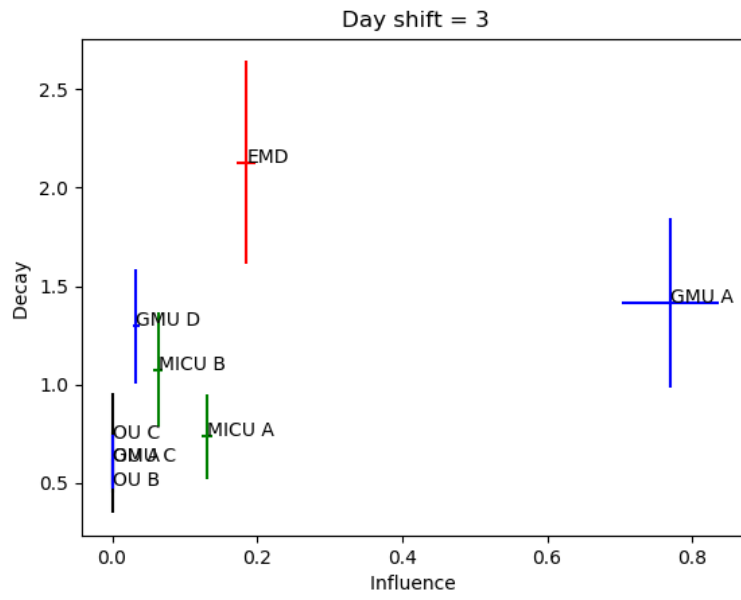


Figure B-6: Units plotted by learned influence and decay with error for three day shift with bootstrapping.

Bibliography

- [1] Emmanuel Bacry, Khalil Dayri, and Jean-François Muzy. Non-parametric kernel estimation for symmetric hawkes processes. application to high frequency financial data. *The European Physical Journal B*, 85(5):157, 2012.
- [2] Pierre Brémaud and Laurent Massoulié. Imbedded construction of stationary sequences and point processes with a random memory. *Queueing systems*, 17(1-2):213–234, 1994.
- [3] Stuart H Cohen, Dale N Gerding, Stuart Johnson, Ciaran P Kelly, Vivian G Loo, L Clifford McDonald, Jacques Pepin, and Mark H Wilcox. Clinical practice guidelines for clostridium difficile infection in adults: 2010 update by the society for healthcare epidemiology of america (shea) and the infectious diseases society of america (idsa). *Infection Control & Hospital Epidemiology*, 31(5):431–455, 2010.
- [4] Daryl J Daley and D Vere-Jones. An introduction to the theory of point processes springer series in statistics, 1988.
- [5] Erik R Dubberke, Kimberly A Reske, Yan Yan, Margaret A Olsen, L Clifford McDonald, and Victoria J Fraser. Clostridium difficile-associated disease in a setting of endemicity: identification of novel risk factors. *Clinical Infectious Diseases*, 45(12):1543–1549, 2007.
- [6] Erik R Dubberke, Yan Yan, Kimberly A Reske, Anne M Butler, Joshua Doherty, Victor Pham, and Victoria J Fraser. Development and validation of a clostridium difficile infection risk prediction model. *Infection Control & Hospital Epidemiology*, 32(4):360–366, 2011.
- [7] Maggie A Dudeck, Lindsay M Weiner, PJ Malpiedi, JR Edwards, KD Peterson, and DM Sievert. Risk adjustment for healthcare facility-onset c. difficile and mrsa bacteremia laboratory-identified event reporting in nhsn. *The Centers for Disease Control and Prevention*, 12, 2013.
- [8] Robert Fekety, Kyung-Hee Kim, Donald Brown, Donald H. Batts, Margaret Cudmore, and Joseph Silva. Epidemiology of antibiotic-associated colitis: Isolation of clostridium difficile from the hospital environment. *The American Journal of Medicine*, 70(4):906 – 908, 1981.

- [9] Thomas R Frieden, Kathleen Ethier, and Anne Schuchat. Improving the health of the united states with a “winnable battles” initiative. *Jama*, 317(9):903–904, 2017.
- [10] Fangjian Guo, Charles Blundell, Hanna Wallach, and Katherine Heller. The bayesian echo chamber: Modeling social influence via linguistic accommodation. In *Artificial Intelligence and Statistics*, pages 315–323, 2015.
- [11] Stephen J Hardiman, Nicolas Bercot, and Jean-Philippe Bouchaud. Critical reflexivity in financial markets: a hawkes process analysis. *The European Physical Journal B*, 86(10):442, 2013.
- [12] Alan G Hawkes. Point spectra of some mutually exciting point processes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 33(3):438–443, 1971.
- [13] Robin LP Jump, Michael J Pultz, and Curtis J Donskey. Vegetative clostridium difficile survives in room air on moist surfaces and in gastric contents with reduced acidity: a potential mechanism to explain the association between proton pump inhibitors and c. difficile-associated diarrhea? *Antimicrobial agents and chemotherapy*, 51(8):2883–2887, 2007.
- [14] K-H Kim, Robert Fekety, Donald H Batts, D Brown, M Cudmore, J Silva Jr, and D Waters. Isolation of clostridium difficile from the environment and contacts of patients with antibiotic-associated colitis. *Journal of infectious diseases*, 143(1):42–50, 1981.
- [15] Ling Yuan Kong, David W Eyre, Jacques Corbeil, Frederic Raymond, A Sarah Walker, Mark H Wilcox, Derrick W Crook, Sophie Michaud, Baldwin Toye, Eric Frost, et al. Clostridium difficile: investigating transmission patterns between infected and colonized patients using whole genome sequencing. *Clinical Infectious Diseases*, 68(2):204–209, 2018.
- [16] Stephen Kralovic, Martin Evans, Loretta Simbartl, and Gary Roselle. Use of a standardized infection ratio (sir) model to monitor a nationwide healthcare-associated clostridium difficile prevention initiative within the us department of veterans affairs (va) healthcare system. In *Open Forum Infectious Diseases*, volume 2. Oxford University Press, 2015.
- [17] Fernanda C. Lessa, Yi Mu, Wendy M. Bamberg, Zintars G. Beldavs, Ghinwa K. Dumyati, John R. Dunn, Monica M. Farley, Stacy M. Holzbauer, James I. Meek, Erin C. Phipps, Lucy E. Wilson, Lisa G. Winston, Jessica A. Cohen, Brandi M. Limbago, Scott K. Fridkin, Dale N. Gerding, and L. Clifford McDonald. Burden of clostridium difficile infection in the united states. *New England Journal of Medicine*, 372(9):825–834, 2015.
- [18] Shelley S Magill, Jonathan R Edwards, Wendy Bamberg, Zintars G Beldavs, Ghinwa Dumyati, Marion A Kainer, Ruth Lynfield, Meghan Maloney,

- Laura McAllister-Hollod, Joelle Nadle, et al. Multistate point-prevalence survey of health care-associated infections. *New England Journal of Medicine*, 370(13):1198–1208, 2014.
- [19] Maggie Makar, John Guttag, and Jenna Wiens. Learning the probability of activation in the presence of latent spreaders. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.
- [20] Aaron C Miller, Linnea A Polgreen, Joseph E Cavanaugh, and Philip M Polgreen. Hospital clostridium difficile infection (cdi) incidence as a risk factor for hospital-associated cdi. *American journal of infection control*, 44(7):825–829, 2016.
- [21] Carlene A Muto. Asymptomatic clostridium difficile colonization: is this the tip of another iceberg?, 2007.
- [22] Radford M Neal et al. Slice sampling. *The annals of statistics*, 31(3):705–767, 2003.
- [23] Jeeheh Oh, Maggie Makar, Christopher Fusco, Robert McCaffrey, Krishna Rao, Erin E Ryan, Laraine Washer, Lauren R West, Vincent B Young, John Guttag, et al. A generalizable, data-driven approach to predict daily risk of clostridium difficile infection at two large academic health centers. *infection control & hospital epidemiology*, 39(4):425–433, 2018.
- [24] Eroboghene H Otete, Anand S Ahankari, Helen Jones, Kirsty J Bolton, Caroline W Jordan, Tim C Boswell, Mark H Wilcox, Neil M Ferguson, Charles R Beck, and Richard L Puleston. Parameters for the mathematical modelling of clostridium difficile acquisition and transmission: a systematic review. *PLoS One*, 8(12):e84224, 2013.
- [25] Amy Priddy, Linell Santella, and Barbara Moran. Geographical mapping of clostridium difficile case locations: A system for understanding transmission patterns within the hospital. *American Journal of Infection Control*, 41(6):S55, 2013.
- [26] Ardeshir Rineh, Michael J Kelso, Fatma Vatansever, George P Tegos, and Michael R Hamblin. Clostridium difficile infection: molecular pathogenesis and novel therapeutics. *Expert review of anti-infective therapy*, 12(1):131–150, 2014.
- [27] Joanna Tarrant, Richard O. Jenkins, and Katie T. Laird. From ward to washer: The survival of clostridium difficile spores on hospital bed sheets through a commercial uk nhs healthcare laundry process. *Infection Control Hospital Epidemiology*, 39(12):1406–1411, 2018.
- [28] US Department of Health and Human Services. Healthy people 2020, 2000.
- [29] Jenna Wiens, Wayne N Campbell, Ella S Franklin, John V Guttag, and Eric Horvitz. Learning data-driven patient risk stratification models for clostridium

difficile. In *Open forum infectious diseases*, volume 1. Oxford University Press, 2014.

- [30] Jenna Wiens, Eric Horvitz, and John V. Guttag. Patient risk stratification for hospital-associated c. diff as a time-series classification task. In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 25*, pages 467–475. Curran Associates, Inc., 2012.
- [31] Mark H Wilcox. Clostridium difficile infection and pseudomembranous colitis. *Best practice & research clinical gastroenterology*, 17(3):475–493, 2003.
- [32] Ke Zhou, Hongyuan Zha, and Le Song. Learning social infectivity in sparse low-rank networks using multi-dimensional Hawkes processes. In *Artificial Intelligence and Statistics*, pages 641–649, 2013.
- [33] Eyal Zimlichman, Daniel Henderson, Orly Tamir, Calvin Franz, Peter Song, Cyrus K Yamin, Carol Keohane, Charles R Denham, and David W Bates. Health care-associated infections: a meta-analysis of costs and financial impact on the us health care system. *JAMA internal medicine*, 173(22):2039–2046, 2013.