Mumps at Harvard: Modeling the Spread of Disease on College Campuses

Contents

1 INTRODUCTION

College campuses provide ideal breeding grounds for infectious disease. Students live in close quarters, pack into dry lecture halls, share food and drinks in the dining areas, and engage in intimate contact. The dense social environment is coupled with the daily stress and lack of sleep of the average college student, weakening the immune system and enhancing disease transmission. Thus, infectious disease outbreaks at universities pose a unique public health challenge because of the intense rapidity with which they spread. Indeed, in March 2014, meningitis quickly infiltrated Princeton University, eventually claiming the life of one student. The Centers for Disease Control and Prevention (CDC) later reported the attack rate of the disease on Princeton's campus to be 134 per 100,000 students – 1400 times greater than the national average [1].

The most recent string of outbreaks on college campuses involves mumps, once a common childhood viral disease. For instance, in 2016, colleges in Iowa, Indiana, Ohio, and Wisconsin all experienced a spike in this disease. Meanwhile, in the Greater Boston area, 210 confirmed mumps cases were identified between January 1 and August 31, 2016, with most occurring at Harvard University [2]. As a highly contagious disease, mumps has the potential to travel quickly and pervasively on a crowded college campus. But, whereas mumps spread rapidly at Ohio State University in 2014 and the University of Iowa in 2006 and 2016, Harvard employed careful precautions and interventions that mitigated excessive spread of the disease and contained it over just a few months [3]. The CDC is currently investigating the techniques with which Harvard so effectively contained mumps on its campus, with the hope that Harvard's approach can be generalized to future university outbreaks.

Thus, to aid the CDC in this task, this paper constructs a mathematical model to simulate the dynamics of mumps on a college campus and quantify the impact of various interventions.

Most epidemiological models have at least one of three flaws: (i) inability to handle small populations, (ii) inability to handle missing or unobserved data, or (iii) inability to evaluate the effects of interventions. The modified stochastic susceptible-exposed-infectious-recovered (SEIR) model presented in this paper addresses these three issues. We fit this model on case data for Harvard's 2016 mumps outbreak provided by the Massachusetts Department of Public Health (MDPH), and find that the three primary interventions implemented by the university were crucial in reducing the size and duration of the outbreak. In particular, Harvard's policies drastically increased the reporting rate of infection and shortened the time a person remains infectious in a susceptible population, relative to the baseline. As a result, one mumps case at Harvard infected an average of one susceptible individual, compared to a case at a school like Ohio State University, which infected an average of three susceptible individuals. Universities that adopt similar strategies can better contain and abate future infectious disease outbreaks.

We divide our analysis in this paper into four stages. We begin by exploring and understanding the Harvard outbreak to determine which interventions were most important. The three interventions that Harvard University Health Services (HUHS) invested the most time and resources in were (i) an email awareness campaign, (ii) more aggressive diagnoses, and (iii) formal isolation of infectious cases.

Because basic epidemiological models are incompatible with characteristics of most campus outbreaks, we next develop a model that accounts for a time-varying infection rate, random fluctuations in a small population like Harvard's, and the possibility of unobserved or overlooked cases. This is accomplished by fitting a modified stochastic SEIR model, that allows for control interventions, within the framework of a Partially Observed Markov Process (POMP) model. While the SEIR model represents the unseen true process that calculates the changing numbers of cases each day, a measurement model that overlays the SEIR model determines the number of cases actually reported publicly.

We then apply the Harvard case data to the model and estimate its parameters using Monte Carlo techniques, and discover that the vigilance on campus was unmatched. Approximately 99% of cases were reported to Harvard's health services, compared to the 4% of cases reported in the overall US population. Additionally, despite a highly contagious disease, HUHS measures suppressed both the average number of secondary infections that a single case could cause and the length of time an infectious person interacted with susceptible persons. Lastly, the impact of new interventions immensely lowered the mumps transmission rate, which soon led to the end of the outbreak.

Although the combination of the three interventions produced these impressive results, in the final stage of this paper, we perform two types of comparative analyses to speculate the individual effects of each intervention. First, we strive to understand the impact of the email awareness campaign and change in diagnostic procedures, which occurred at a specific time point during the outbreak. The outbreak size is three times larger without these two interventions, according to the model. The second comparative analysis contrasts the model parameters of Harvard's outbreak, which had formal isolation policies, with the model parameters of Ohio State University's outbreak, which did not. We find the time an infectious person spreads their symptoms to be substantially lower at Harvard than at Ohio State, leading to a shorter outbreak.

The conclusions from this paper are relevant in guiding future responses to infectious disease outbreaks on college campuses. Without effective measures in place, diseases like mumps and meningitis penetrate these congested environments at much faster rates than in the overall population and can lead to serious health complications. Simple interventions that ensure most cases are detected, treated, and separated from susceptibles make a significant difference, as Harvard’s outbreak and response prove.

2. THE HARVARD OUTBREAK

2.1 Biological Background of Mumps

Mumps is an infectious viral disease, spread through respiratory tract secretions. Standing within three to six feet of an infected person when she coughs or sneezes or having direct contact with infected secretions facilitates contraction of mumps. Once infected, the incubation period can range from 12 to 25 days (with an average of 17 days); generally, one does not show symptoms during this period and cannot infect others [4]. Then, approximately two days before symptoms arise, a person becomes contagious. The most common symptom that occurs in a mumps patient is parotitis, i.e. swelling of the salivary glands. It is recommended that those suspected of mumps be isolated until five days after the onset of parotid swelling, at which point they are no longer considered infectious. Non-specific symptoms of mumps that may precede parotitis include low-grade fever, headache, anorexia, and malaise. Because these symptoms are generic, a correct diagnosis can be delayed, allowing an infectious person to continue transmitting the disease, especially in crowded settings like dormitories [2].

In the pre-vaccine era, over 90% of US-born children had experienced mumps by age 20. Incidence significantly declined with the licensure of a live attenuated vaccine, known as the Jeryl Lynn strain, in 1967. In 1977, it became routine to include the Jeryl Lynn strain in the measles-mumps-rubella (MMR) vaccine administered to infants. After a series of outbreaks in children in middle school and high school in the late 1980s, children were recommended to receive a second MMR dose between ages 4 and 6. After the introduction of the two-dose vaccination program in 1989, the count of mumps cases in the US plummeted further, reducing disease rates by 99% by 2005. The CDC reports that two doses of the vaccine are 88% effective at protecting against mumps while one dose is 78% effective [3].

Nevertheless, despite a rising vaccinated population, there has been a recent resurgence of mumps, particularly on college campuses, with a steep jump from 229 cases in 2012 to 5833 cases in 2016 [2]. These statistics are troubling for two reasons. Firstly, although a typically mild disease in children, up to 10% of mumps infections acquired after puberty can cause severe complications, including orchitis, meningitis, and deafness. Of those who develop orchitis, which is inflammation of one or both testicles, 13% may later suffer impaired fertility [5]. In the mumps outbreak in Boston, 8.5% of infected males developed orchitis, as seen in **Figure 1**. Secondly, a majority of recent mumps cases have occurred in young adults who had received the recommended two vaccine doses. This suggests that vaccine-derived immunity wanes over time, unlike natural immunity – protection acquired from contracting the disease – which is permanent. Indeed, Lewnard and Grad (2018) estimate that 33.8% of young adults (ages 20 to 24) were susceptible to mumps in 1990, in contrast to the 52.8% susceptible in 2006, as vaccinations have replaced contraction as the source of immunity [6].

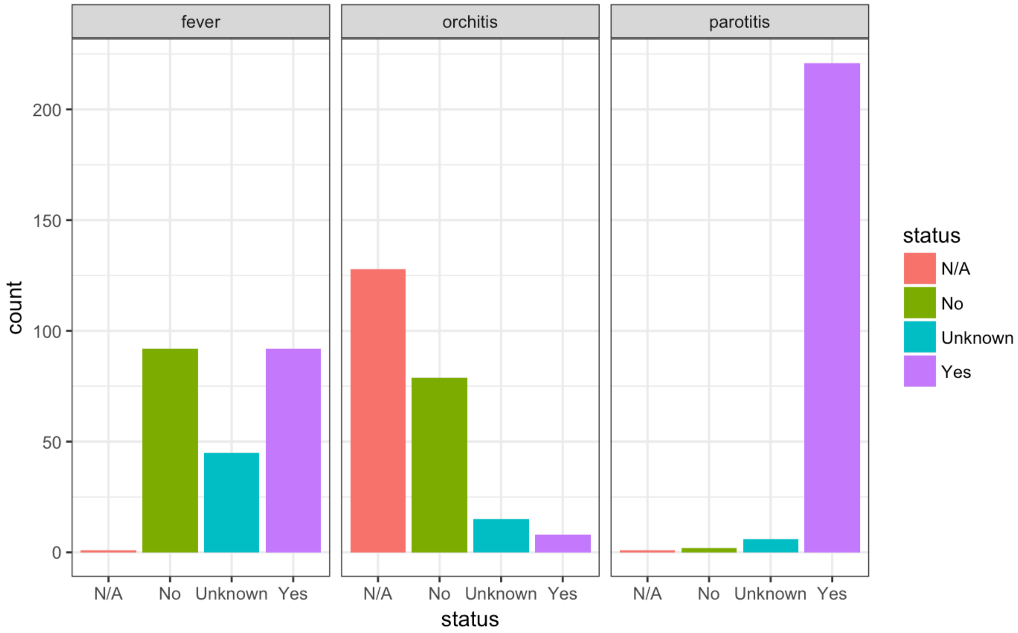

Figure 1: The frequency of fever, orchitis, and parotitis in mumps patients across Boston in 2016. While orchitis was rare in most cases, fever was seen in nearly half the patients, and parotitis was identified in the majority of individuals.

The realization that vaccine-derived immunity wanes has propelled a new body of research (see *Section 3*) aimed at understanding whether administration of a third MMR dose could prevent future mumps outbreaks [6, 7]. This paper, however, shifts the focus from prevention to intervention. Even if the use of a booster vaccination is proven theoretically, it is unlikely that universities with limited resources will proactively invest in a third dose. A rough cost analysis conducted by HUHS showed that, while the total mumps care expenses for Harvard was approximately $75,000, the cost of providing a third MMR dose to every member of the Harvard community (at $83 per vaccination) was $1.7 million [8]. Therefore, at least in the short term, a third MMR dose cannot be the only answer to handling mumps outbreaks; we must explore more immediate solutions and interventions.

2.2 Outbreak Summary

2.2.1 Number of Cases

The mumps outbreak at Harvard officially began in February 2016, when six students reported onset of parotitis to HUHS. For the next three months, the number of cases continued to rise, until finally plateauing in late May and early June, when summer break began; thus, this paper focuses its analysis of the outbreak between mid-February and late June. The number of cases each day during this time period is shown in **Figure 2**.

From **Figure 2**, we also see there are two waves of the outbreak – one occurring in the month of March and a larger one occurring in mid-April – totaling 189 cases. While a majority of these cases were undergraduate students, some involved employees and members of Harvard Law School, Harvard Business School, or the Graduate School of Arts and Sciences [9].

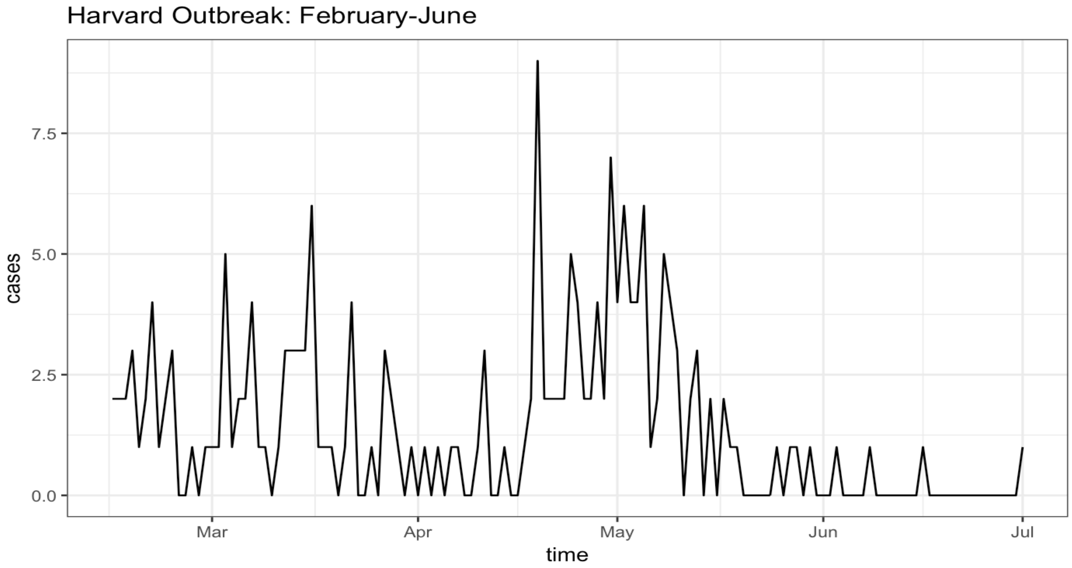

Figure 2: The daily number of new mumps cases (probable or confirmed) at Harvard between February and June 2016. Both probable and confirmed cases display clinical symptoms of mumps, but only confirmed cases have a positive PCR result.

At Harvard, 99.4% of undergraduate students and 98% of graduate students and employees have been vaccinated. Unlike many other universities, which verify mumps immunizations through a questionnaire, Harvard requires documentation of one's vaccine status [10]. Yet, despite these extensive precautions, Harvard could not prevent disease transmission, for close contact on a college campus is inevitable. In fact, most cases seen in **Figure 2** had received the recommended number of MMR doses.

2.2.2 Interventions

Although Harvard's prevention efforts through requiring vaccination failed, their intervention efforts were effective in mitigating the spread of mumps. The university employed three tactics that we center our analysis on: (i) an email awareness campaign, (ii) more aggressive diagnoses, and (iii) formal isolation of infectious persons.

First, the Harvard community was kept well-informed of the spread of mumps. Between February and May 2016, Paul J. Barreira, Director of HUHS, sent six different emails to Harvard students, employees, and colleagues with information on the gravity of the outbreak, recommendations on how to prevent transmission, and instructions on how to identify mumps. These emails served as consistent reminders to stay alert on campus and receive care at the first sign of symptoms. Particularly at the peak of the outbreak, roommates, resident deans, and athletic coaches all played essential roles in reporting potential cases of mumps, so that few cases likely went undetected and untreated by HUHS [11, 12].

Second, Harvard acted aggressively to treat and isolate anyone suspected of mumps. At the beginning of the outbreak, HUHS personnel struggled to diagnose mumps because its symptoms can be non-specific and its manifestation is less extreme in vaccinated people. Thus, they used PCR tests to determine if one had mumps virus. Later, upon recommendations from the MDPH, HUHS stopped automatically ruling out those with negative PCR results, given that false negatives were quite frequent in vaccinated individuals and individuals who reported their infection to the clinic belatedly (see **Appendix** for details on negative PCR tests) [13]. Policies were improved so that anyone who entered HUHS displaying clinical symptoms of mumps was now deemed infected and infectious. Therefore, **Figure 2** includes both confirmed and probable cases of mumps. Confirmed cases are those with a positive laboratory test for mumps virus. Probable cases are those who either tested positive for the anti-mumps IgM antibody or had an epidemiologic linkage to another probable or confirmed case [11, 12].

Third and perhaps most notably, Harvard isolated most confirmed or probable cases of mumps. While many universities simply suggest self-isolation in one’s room or dormitory (which leaves roommates and friends highly susceptible to the disease), Harvard removed anyone with clinical symptoms from the population. Of the 230 total cases at Harvard between February 2016 and November 2017, 96 were isolated in alternate housing on campus while 110 were isolated offsite. Although a person remains infectious with mumps for five days, Harvard isolated patients for six days out of caution [11].

For the purposes of this paper, we study these three interventions, but Harvard used a variety of smaller techniques to contain the disease. For instance, water fountains with a weak upward flow were repaired in late March when it became apparent that students were directly touching the fountain with their water bottles or mouths [12].

Figure 3 shows a timeline of the interventions used by HUHS as well as periods when the population was fluctuating (such as during spring and summer break). The orange lines represent when HUHS emails were sent out to the community to raise awareness, the purple lines delineate vacation times, and the blue line marks when the MDPH recommended that HUHS more carefully assess those with negative PCR results. Because isolation of infectious cases occurred continuously throughout the entire outbreak, this control intervention is not included in **Figure 3**.

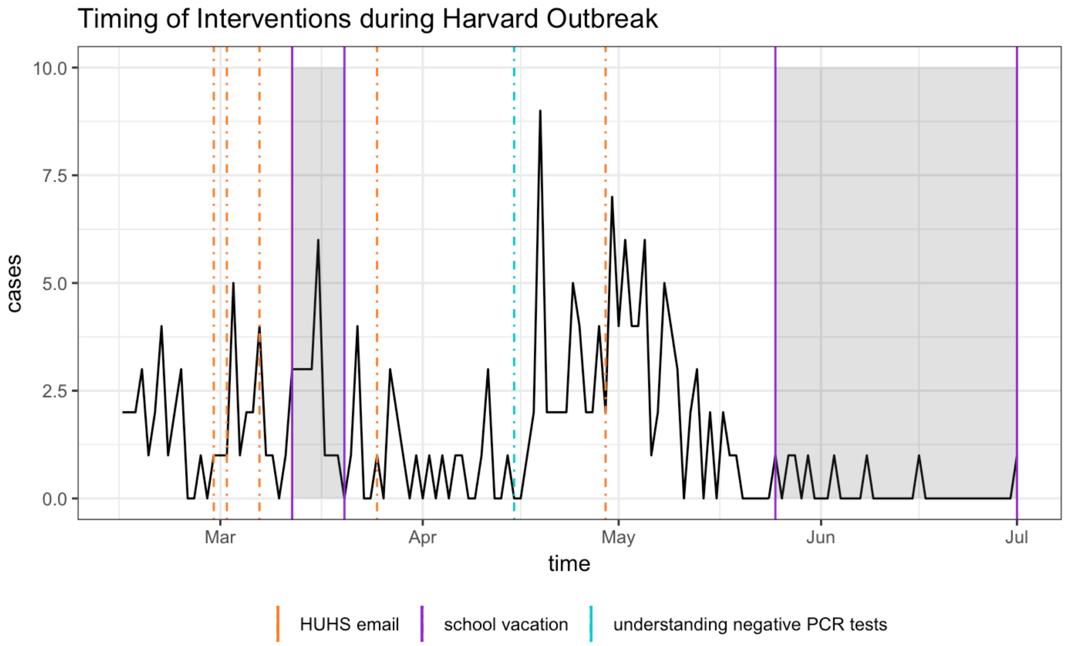

Figure 3: The timeline of school vacations and control interventions employed by HUHS (apart from its isolation policy). HUHS sent multiple emails over the course of the outbreak, raising awareness about the spread of mumps. Additionally, in mid-April, HUHS began more carefully diagnosing mumps, rather than automatically ruling out those with negative PCR tests.

Visually, **Figure 3** shows a spike in cases a few weeks after spring break, which is consistent with the fact that the mumps incubation period lasts approximately three weeks. Furthermore, the interventions in April seem to have had a lasting effect. Shortly after HUHS improved its criteria for diagnosis and sent an urgent email regarding the recent increase in cases (in late April), there was a steep decline in the number of new cases. We quantify the effects of these actions further in the modeling section of this paper.

2.3 Harvard Data

2.3.1 Data Description

For the remainder of this paper, we use data provided by the Massachusetts Department of Public Health, which documented every mumps case between 2015 and 2017 at schools across Massachusetts [14]. Though our analysis is primarily centered on Harvard's outbreak, an understanding of outbreaks at other schools can provide further insight into Harvard's unique response. Each row of data consists of the following:

- Background information about the patient, including their gender, age, county, and institution
- Details about their symptoms and vaccination status
- The date they reported their symptoms and the date of symptom onset
- Lag time, a column we have constructed to track the number of days between the date of symptom onset and admission to a medical clinic

2.3.2 Cluster Analysis

Before designing a formal model, we first test our hypothesis that Harvard's response to the mumps outbreak truly was distinct compared to other schools through cluster analysis.

Cluster analysis involves grouping a set of points in such a way that the objects in the same group (or cluster) are more similar to each other than objects in other clusters. If our hypothesis is correct, we should expect to see a few different clusters, with one cluster solely composed of Harvard data points. Alternatively, to determine if there are any unique qualities about Harvard's outbreak, we could also compare Harvard cases to cases at every other institution. Because there are 23 different institutions in the MDPH data, however, this strategy is too laborious.

We perform k-medoids clustering using the PAM algorithm to group the data based on four features – age of patient, date of symptom onset, institution affiliation, and lag time. PAM stands for "partition around medoids" and is a more robust version of the K-means algorithm. The goal of PAM is to search for K representative observations that are to be the medoids of the K clusters. These observations are chosen in such a way that they minimize the dissimilarities or distances of the remaining data points to their closest representative observation. In this scenario, we minimize Gower distances (as opposed to the more commonly-calculated Manhattan or Euclidean distances), which is a useful approach when trying to calculate distances between observations with both quantitative features (like age) and categorical features (like institution affiliation). See **Appendix** for a more detailed explanation of Gower distances.

To find the optimal number of clusters K, we use the elbow method. This procedure involves calculating $W(C_k)$, the sum of squared Gower distances between all pairs of points in cluster k, and then totaling $W(C_k)$ over all the different clusters in a particular clustering:

$$T_K = \sum_{k=1}^{K} W(C_k)$$

We can then let K vary over a range of values, such as between 1 and 10, and compute the total intra-cluster variation, T_K, for each K. See **Appendix** for a figure plotting T_K against K. To balance goodness-of-fit with dimensionality, we look for a sharp change in the gradient of T_K in the plot, known as the bend or "elbow." This elbow occurs at $K = 3$, thus giving us the optimal number of clusters.

After running PAM on our dataset and defining K as 3, we see that our hypothesis was indeed confirmed. While two of the clusters consist of observations affiliated with numerous different institutions, Cluster 1 is solely made up of Harvard observations. Other institutions are in fact more similar to each other than they are to Harvard. Refer to the **Appendix** for a visualization and summary statistics of the different clusters.

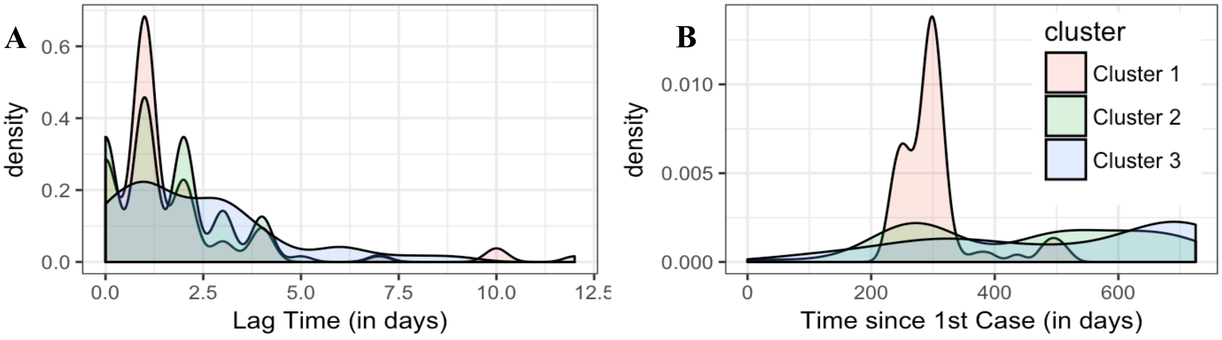

Figure 4: Density plots that compare the different features of each cluster. (a) Distribution of lag times in the three clusters. Cluster 1 (with Harvard data points) has the most right-skewed distribution. (b) Distribution of the days that cases occurred. The outbreak for Cluster 1 occurs in the most concentrated manner.

Figure 4 gives some insight into the distinctive characteristics of Harvard's outbreak, such as how it occurred in a concentrated manner (**4B**). Importantly, the mean lag time for Cluster 1 is 1.568 days, compared to 1.677 days for Cluster 2 and 2.357 days for Cluster 3 (**4A**). Since we define lag time as the time between the onset and the reporting of symptoms, having a lower lag time means that one is infectious in the population for a shorter amount of time and diagnoses their symptoms earlier. It is likely no coincidence that Harvard's lag time is slightly shorter than that of other institutions, given HUHS's email awareness campaign. Therefore, the cluster analysis does indeed justify our treatment of Harvard as different from other schools.

3. LITERATURE REVIEW

With the resurgence of mumps outbreaks across the United States, numerous papers have been released in the past year on the prevention of mumps through additional vaccination. A new study (2018) published by researchers at the Harvard School of Public Health found that vaccine-derived immunity lasts, on average, 27 years after the last dose. To sustain protection in adulthood, they recommend a third dose at age 18 or booster shoots [6]. Similarly, Shah et al. (2018) conduct a vaccine campaign during a mumps outbreak at the University of Iowa in 2016 and find statistically significant results supporting the efficacy of an additional MMR dose. While 25% of the cases occurred five months after this vaccine intervention, 75% of the cases occurred five months before the intervention [7].

However, literature addressing and analyzing the importance of alternative, more immediate interventions during a mumps outbreak is sparse. A recent paper by Li et al. (2017) models mumps transmission dynamics and the impact of control interventions in mainland China [5]. It concludes that, apart from increasing vaccine coverage, the most effective measures for the control and prevention of mumps are (i) cutting off transmission routes of the disease by increasing awareness and promoting good personal hygiene habits and (ii) reducing the length of the infectious period through earlier treatment and isolation of contagious persons. Although these findings are relevant to our research, the model by Li et al. does not translate well to college campuses because it was built to handle large populations like mainland China.

Models that are more applicable to our data are found in papers by He et al. (2009) and Lekone and Finkenstadt (2006) [15, 16]. He et al. successfully simulate the dynamics of measles in both large and small populations by relying on a stochastic model that can handle random fluctuations in the population. Moreover, unlike most models, which assume that all cases are reported, He et al. account for partially observed data. One important element missing from their model, however, is a parameter that can measure the effect of control interventions. The epidemic model designed by Lekone and Finkenstadt to model the spread of Ebola addresses this issue. Rather than fixing the transmission rate to a constant, they introduce a transmission rate function that varies temporally in response to new interventions. We leverage the strengths of the two models introduced in these papers to construct our own model. This is described further in the following section.

4. MODEL FOUNDATIONS

In this section, we develop the foundations for our model by integrating components of compartmental models with Partially Observed Markov Process (POMP) models. Compartmental models simplify the mathematical modeling of infectious disease by splitting the population into non-intersecting classes or "compartments" that reflect characteristics of the disease. However, these models often fail to address the problem of missing or unobserved data, and so we rely on POMP models to address such limitations.

4.1 Compartmental Models

To develop a compartmental model that aptly describes the spread of disease on a college campus, we use a basic epidemiological model as a baseline and subsequently address its assumptions that are not compatible with the Harvard outbreak.

4.1.1 SIR Model

Proposed in 1927, the Kermack-McKendrick model was one of the first epidemic models in the field. This model is known as SIR because the population is divided into three compartments:

- The first class of individuals, known as **susceptibles**, are those that are healthy but may contract the disease. The size of this compartment is denoted by *S*.
- The second class of individuals, known as the **infectious**, are those who have contracted the disease and can spread it to those that are susceptible. The size of this compartment is denoted by *I*.
- The third class of individuals, known as the **removed**, are those who have been removed from the population or recovered from the disease and cannot contract the

disease again; in other words, they are now immune and cannot re-enter the susceptible population. The size of this compartment is denoted by R.

The Kermack-McKendrick model describes the rate of movement between compartments as the derivative of the sizes of the classes with respect to time. The parameters β, the transmission rate, and α, the removal rate, partially control how quickly people change compartments. We can mathematically express the model as a system of ordinary differential equations:

$$\frac{dS}{dt} = -\beta IS$$

$$\frac{dI}{dt} = \beta IS - \alpha I$$

$$\frac{dR}{dt} = \alpha I$$

$$N = S + I + R \tag{1}$$

In Equation 1, members of the susceptible population become infected at a rate proportional to the number of infectious people, βI. Meanwhile, the size of the infectious population changes as susceptibles become infected at a rate of βI and removals from the population occur at a rate of α. Finally, the size of the removed population increases as infectious people recover (or are removed) at a rate of α. Because the population is closed, the sum of the susceptible, infectious, and removed populations should equal N, the total size of the population, at all times [17].

4.1.2 Assumptions of SIR Model

The SIR model simplifies the world in which disease spreads and thus relies on many assumptions that we should consider. Below we determine which assumptions are consistent with the dynamics of mumps on a college campus, and suggest ways in which our own model will either emulate or deviate from the baseline model:

1. *The population is fixed.* This assumption does not necessarily fit with mumps transmission on a college campus, as it is not guaranteed that the population on a college campus will remain fixed throughout the year. Interviewing season, spring break, and summer vacation are all examples in which the population is in flux. We can account for these fluctuations by including demographic stochasticity in our model. Demographic stochasticity refers to chance independent events that cause random fluctuations in the population.
2. *The population is completely homogeneous – there is no inherent age, spatial, or social structure and thus every individual in the population interacts the same as another.* This is reasonable because the majority of the population is in the same age range. Furthermore, all students interact similarly, in that they attend classes, live in packed dormitories, and eat in common areas.
3. *All persons who are not either infected or recovered are equally susceptible to the disease.* Given that approximately 99% of Harvard's campus has received the recommended number of doses, we can assume all persons are equally susceptible. On the other hand, in large, heterogeneous populations, this assumption would not translate well. Some individuals have received no vaccinations, some have received one MMR dose, and some have received two doses, leading to varying amounts of susceptibility.
4. *The incubation period of the infectious agent is instantaneous. In other words, a person immediately becomes infectious once exposed to the disease, and so there is no intermediate compartment between S and I.* This assumption is not consistent with the course of mumps. The incubation period for mumps ranges from 12 to 25 days. Hence, in our model we must also include a compartment between S and I that accounts for this period in which a person has been exposed to the virus but cannot yet transmit it. In related

literature, a model that includes this compartment is known as a **SEIR model**, because in addition to *S, I,* and *R*, there is a class for **exposed** (*E*) individuals.

5. *The disease confers immunity, preventing reinfection and reentry into the susceptible class; instead, a recovered person moves into a new compartment R.* This assumption is realistic because contracting mumps confers permanent immunity. An infected person can never again be susceptible to mumps.
6. *The rates of transfer between compartments is constant*. Because this paper aims to determine the effect of interventions on the length and size of the outbreak, this assumption will not be viable in our own model. After an intervention is employed, we expect the rates of transfer between certain compartments to fall instantaneously, and thus will allow for rate changes in our model.
7. *The epidemic process is deterministic.* A model based on this assumption is reasonable in large populations, but in small populations, there can be significant fluctuations in incidence and prevalence of infection that occur merely by chance. Stochastic models, a complement to deterministic models, are particularly useful when compartment sizes are small, as they are in this Harvard example. These models possess inherent randomness and thus the same set of parameter values and initial conditions leads to different outputs each time. Nevertheless, it is important to keep in mind that although stochastic models are a better representation of the natural world, they are also more complicated than deterministic models.

We conclude this analysis of assumptions by settling on using a SEIR model that accounts for demographic stochasticity and changing rates of transfer between compartments. Further, we consider the merits of stochastic processes over deterministic processes, given the size of our population. These alterations should make our model more accurate and applicable to the transmission of mumps on college campuses.

4.2 POMP Models

Although we have addressed some issues with the basic compartmental model, it is important to note that these models assume access to fully observed disease data. In reality, not all mumps cases are reported and latent mumps carriers exhibit no symptoms at all. The exact number of cases per day is more difficult to estimate than compartmental models make it seem.

Thus, throughout this paper we will assume that we only have partially observed mumps data by building a **Partially Observed Markov Process** model [18]. POMP models combine the interpretable elements of the compartmental model approach to modeling disease transmission with a probabilistic model for the unobserved data.

POMP models (see **Figure 5**) represent data $y_1^*, \ldots, y_N^*$ collected at times $t_1 < \cdots < t_N$ as noisy, incomplete observations or measurements of an unobserved Markov process $\{X(t), t \geq t_0\}$. $\{X(t)\}$ is Markovian if only its current value, and not its history, inform the future of the process. Disease transmission, represented by compartmental models as we saw above, is indeed a Markov process because the number of infectious people at time t is solely determined by the number of infectious people at time $t - \delta$.

A POMP model is characterized by the transition density and measurement density of its stochastic processes. The one-step transition density is represented by $f_{X_n|X_{n-1}}(x_n| \, x_{n-1}; \theta)$, since $\{X(t)\}$ is Markovian and only relies on the previous state. Meanwhile, the measurement density depends on only the state of the Markov process at that time and so is represented by

$f_{Y_n|X_n}(y_n|\, x_n; \theta)$, where Y_n is a random variable modeling the observation at time t_n [18]. Hence, the entire joint density for a POMP model, including the initial density $f_{X_0}(x_0; \theta)$, is

$$f_{X_{0:N}Y_{1:N}}(x_{0:N}, y_{1:N}; \theta) = f_{X_0}(x_0; \theta) \prod_{n=1}^{N} f_{X_n|X_{n-1}}(x_n|\, x_{n-1}; \theta) f_{Y_n|X_n}(y_n|\, x_n; \theta),$$

and the marginal density for the sequence of measurements, $Y_{1:N}$, evaluated at the data, $y_{1:N}^{*}$, is

$$f_{Y_{1:N}}(y_{1:N}^{*}; \theta) = \int f_{X_{0:N}Y_{1:N}}(x_{0:N}, y_{1:N}; \theta)\, dx_{0:N}.$$

Now that we have a basic understanding of POMP models and compartmental models, we can create a model of mumps dynamics on college campuses.

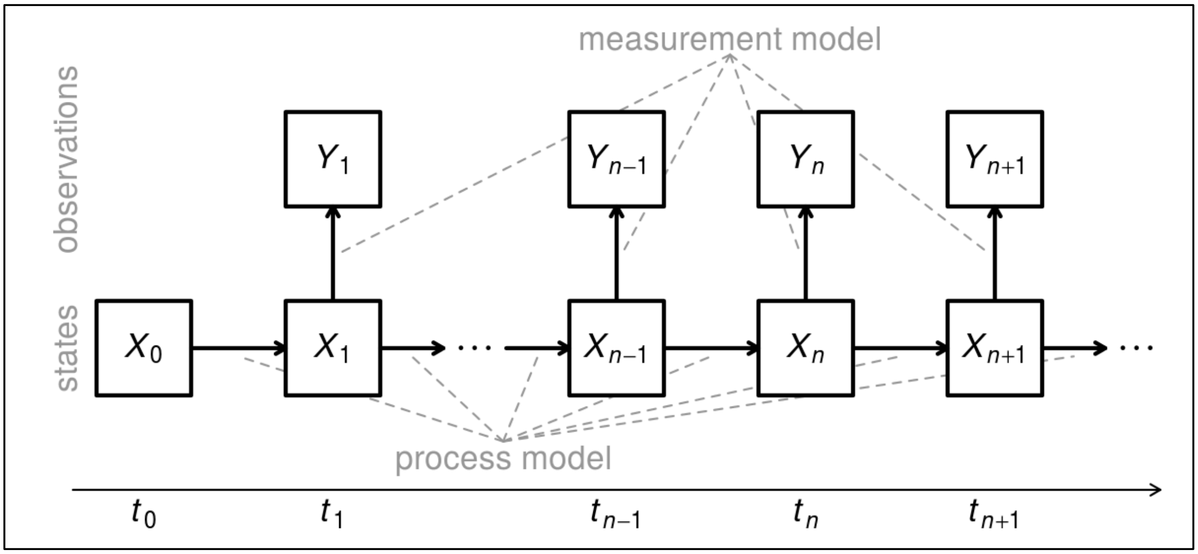

Figure 5: POMP model schematic, with the unobserved process model underlying the measurement model, which then generates the observations [19].

5. METHODOLOGY

5.1 Model Setup

To simulate mumps transmission at Harvard, we construct both the process model (which is unobserved) and the measurement model, the two components of a POMP model. The process model, which we define as a SEIR model, provides the change in true incidence of mumps at every time point, while the measurement model incorporates the fact that not all cases are observed or reported.

5.1.1 Process Model

We first develop the Markovian process model, which counts the true number of cases. As discussed in *Section 4.1.2*, the underlying dynamics of mumps can be aptly captured by a stochastic SEIR compartmental model. Compartmental models are intrinsically Markovian because the future state of each compartment is based on only the present state of the process. As with the basic SIR model, many of the assumptions are still the same with our stochastic SEIR model. The population is considered closed and homogeneous because we are analyzing students on a college campus. However, we add parameters that induce random fluctuations into the population and change the compartments' rates of transfer in response to interventions.

The primary difference between the basic SIR and this stochastic SEIR is that we now use probabilistic densities for the transition of state variables. Moreover, although disease dynamics are technically a continuous Markov process, this is computationally complex and inefficient to model, and so we make discretized approximations by updating the state variables after a time step, δ. We set δ to a small value, such as one day for the Harvard outbreak. The system of discretized equations are shown in Equation 2 below, where $B(t)$ is the number of susceptible individuals who become exposed to mumps, $C(t)$ is the number of newly infectious cases, and $D(t)$ is the number of cases that are removed from the population:

$$S(t+\delta) = S(t) - B(t)$$

$$E(t+\delta) = E(t) + B(t) - C(t)$$

$$I(t+\delta) = I(t) + C(t) - D(t)$$

$$R(t+\delta) = R(t) + D(t)$$

$$S(t) + E(t) + I(t) + R(t) = N \qquad \textbf{(2)}$$

Equation 2 depicts how the sizes of the four compartments (susceptible, exposed, infectious, and removed) change between $(t, t+\delta)$. The susceptible population decreases by the number of susceptibles that become exposed to mumps at time t. Meanwhile, the exposed class increases by the number of newly-exposed people and decreases by the number of newly-infectious people at time t. The infectious population increases by those who are now contagious and decreases by those who have been removed from the population at time t. Finally, the removed compartment increases by the number of infectious people who have recovered or been removed from the population at time t and thus can no longer spread their infection. The model further assumes that the population size N remains constant at every time point.

We add inherent randomness to our model by setting $B(t)$, $C(t)$, and $D(t)$ as binomials. If we assume that the length of time an individual spends in a compartment is exponentially distributed with some compartment-specific rate $x(t)$, then the probability of remaining in that compartment for an additional day is $\exp(-x(t))$ and the probability of leaving that compartment is $1 - \exp(-x(t))$:

$$B(t) \sim \text{Bin}(S(t), 1 - \exp(-\lambda(t)), \text{ where } \lambda(t) = \beta(t)\frac{I(t)}{N}$$

$$C(t) \sim \text{Bin}(E(t), 1 - \exp(-\sigma))$$

$$D(t) \sim \text{Bin}(I(t), 1 - \exp(-\gamma)) \qquad \textbf{(3)}$$

The force of infection, λ(t), is the transition rate between the susceptible and exposed classes, and should increase as the fraction of the population infected increases. Thus, the formula for $\lambda(t)$ is $\beta(t)\frac{I(t)}{N}$, where $\beta(t)$ represents the transmission rate of the disease. We denote $\beta(t)$ as a step function in order to account for potential control interventions added at time τ:

$$\beta(t) = \{ \begin{matrix} \beta, & t < \tau \\ \beta e^{-q(t-\tau)}, & t \geq \tau \end{matrix} \qquad \textbf{(4)}$$

Importantly, once the intervention occurs, $\beta(t)$ does not immediately change to a different constant. Instead, $q > 0$ is the rate that $\beta(t)$ decays for $t \geq \tau$, since the transmission rate should *gradually* decrease as the intervention affects more and more people [15].

Furthermore, in Equation 3, σ is the transition rate between the exposed and infectious classes, and γ is the transition rate between the infectious and removed compartments. σ^{-1} represents the mean length of time a person stays in the latent stage and γ^{-1} represents the mean length of time a person is infectious before being removed from the population (either because of intervention efforts or natural recovery). Unlike λ(t), we would generally expect these two parameters to be constant over the course of the epidemic.

Finally, in evaluating epidemics, it is essential to estimate the basic reproduction number, $R0$, which equals the expected number of secondary cases produced by an infectious person in a completely susceptible population [20]. $R0$ measures the *initial* growth rate of an outbreak and so, if it is less than 1, then the infection will die out and there will be no epidemic. For our stochastic SEIR model, this constant can be expressed as $R0 = \frac{\beta}{\gamma}$ [17]. Meanwhile, the time-dependent *effective* reproduction number is defined as $R_E(t) = \frac{\beta(t)}{\gamma} * \frac{S(t)}{N}$, but because $S(t) \approx N$, we can simplify this expression to $R_E(t) \approx \frac{\beta(t)}{\gamma}$. Both the basic and effective reproduction numbers allow us to understand the strength of an outbreak.

Now consider the measurement model, which maps this stochastic process model to the real data.

5.1.2 Measurement Model

Although it is impossible to directly record the number of people that are susceptible, exposed, infectious, and removed directly, the MDPH data tells us the number of observed cases per day. We expect the mean number of observed cases per day to be the true number of cases multiplied by the reporting rate ρ ($\rho < 1$). However, rather than simply denoting the observed number of cases as a binomial distribution, we should account for greater variability in the measurements than a binomial distribution expects, since the Harvard population is small and prone to randomness [16]. Thus, the number of observed cases, y_t, given the number of true cases, $C(t)$, can be best modelled by an overdispersed binomial distribution defined as a Normal random variable (discretized because case counts must be integer values):

$$y_t \mid C(t) \sim \text{Normal}(\rho C(t), \rho(1-\rho)C(t) + \left(\psi \rho C(t)\right)^2) \qquad \textbf{(5)}$$

For example, the probability that $y_t = y$ is:

$$\mathbb{P}[y_t = y \mid C(t) = \ C] =$$
$$\phi(y + 0.5; \rho C, \rho(1-\rho)C + (\psi \rho C)^2) - \ \phi(y - 0.5; \rho C, \rho(1-\rho)C + (\psi \rho C)^2) \qquad \textbf{(6)}$$

In Equation 5 and 6, the parameter ψ handles the increased variability intrinsic in a small population. If $\psi = 0$, the variance in our measurement model simplifies to the variance for a binomial distribution.

5.1.3 Final POMP Model

We have now formulated expressions for both our process model and measurement model. For each time point, the process model generates the number of new cases based on binomially distributed counts. The measurement model then estimates the observed number of cases based on the true number of cases and reporting rate.

Parameters that will be particularly important to estimate in this model are: (i) q, which measures the effect of an intervention after its introduction at time τ, (ii) γ, which determines the length of the infectious period before removal from the population, (iii) $R0$, the initial growth rate of an outbreak, (iv) ρ, the reporting rate, and (v) ψ, the additional variability in a population.

5.2 Parameter Estimation

In this section, we devise a methodology to estimate parameters that maximize the likelihood of the model, by relying on sequential Monte Carlo (SMC) techniques.

5.2.1 Likelihood

In order to estimate the optimal parameters and diagnose the fit of the model, we first must understand how to calculate the likelihood for POMP models. The likelihood function is the density function evaluated with data at a candidate set of parameter values. It is computationally simpler to work with the log likelihood, $l(\theta) = \log f(y_{1:N}; \theta)$, so that we can deal with sums instead of products.

Usually, for complex statistical models, it is difficult to analytically solve or even determine the density function. Nevertheless, we can take advantage of a simulation-based approach, in which we simulate the random variable $Y_{1:N}$, which implicitly defines the density function. Thus, likelihood evaluation via sequential Monte Carlo is one standard method to obtain the log likelihood for POMP models, because it simulates sample paths rather than requiring explicit forms of the transition probabilities [18].

In order to understand how SMC calculates the likelihood of the model, let us first derive a new expression for the likelihood of a POMP model by factorizing it as the product of conditional likelihoods:

$$L(\theta) = \prod_{n=1}^{N} L_{n|1:n-1}(\theta) \tag{7}$$

where $L_{n|1:n-1}(\theta) = \mathbb{P}[y_n^*|y_{1:n-1}^*;\theta]$ and there are N time points. The structure of a POMP model then implies the representation of $L_{n|1:n-1}(\theta)$ as

$$L_{n|1:n-1}(\theta) = \int \mathbb{P}[y_n^*|x_n;\theta]\mathbb{P}[x_n|y_{1:n-1}^*;\theta]\,dx_n \tag{8}$$

so that the final expression for the likelihood is:

$$\prod_{n=1}^{N} \int \mathbb{P}[y_n^*|x_n;\theta]\mathbb{P}[x_n|y_{1:n-1}^*;\theta]\,dx_n \tag{9}$$

In Equation 9, although $\mathbb{P}[y_n^*|x_n;\theta]$ is simple to calculate (using Equation 6), $\mathbb{P}[x_n|y_{1:n-1}^*;\theta]$ is more difficult to evaluate. We can use the Markov property to determine an expression for this probability, known as the prediction formula:

$$\mathbb{P}[x_n|y_{1:n-1}^*;\theta] = \int \mathbb{P}[x_n|x_{n-1};\theta]\,\mathbb{P}[x_{n-1}|y_{1:n-1}^*;\theta]\,dx_{n-1} \tag{10}$$

We can then use Bayes' Theorem to determine an expression for $\mathbb{P}[x_{n-1}|y_{1:n-1}^*;\theta]$ (in Equation 10), known as the filtering formula:

$$\mathbb{P}[x_n|y_{1:n}^*;\theta] = \mathbb{P}[x_n|y_n^*, y_{1:n-1}^*;\theta] = \frac{\mathbb{P}[y_n^*|x_n;\theta]\,\mathbb{P}[x_n|y_{1:n-1}^*;\theta]}{\int \mathbb{P}[y_n^*|x_n;\theta]\mathbb{P}[x_n|y_{1:n-1}^*;\theta]\,dx_n} \tag{11}$$

The prediction and filtering formulas give us a recursion. Specifically, the prediction formula calculates the prediction distribution, $f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1})$, at time t_n by using the filtering distribution, $f_{X_n|Y_{1:n}}(x_n|y^*_{1:n})$, at time t_{n-1}. Meanwhile, the filtering formula gives us the filtering distribution at time t_n using the prediction distribution at time t_n.

In SMC, we use Monte Carlo techniques to sequentially estimate the integrals in the prediction and filtering recursions, which in turn allows us to estimate $\mathbb{P}[x_n|y^*_{1:n-1};\theta]$. Although a more in-depth algorithm for SMC is shown in the **Appendix**, here we present its basic steps [19]:

1. We generate J points or particles, $\{x^F_{n-1,j}\}$, that are drawn from the filtering distribution at time t_{n-1}.
2. We then obtain a sample of points, $\{x^P_{n,j}\}$, at time t_n, drawn from the prediction distribution by simulating the process model that we defined in *Section 5.1.1*.
3. Once we have a sample of points from the prediction distribution at the next time step, we can find the conditional likelihood (defined in Equation 8). Using the Monte Carlo principle, we can approximate the conditional likelihood to be

$$\hat{L}_{n|1:n-1}(\theta) = \frac{1}{J}\sum_{j} \mathbb{P}\left[y^*_n \middle| x^P_{n,j};\theta\right] \qquad \textbf{(12)}$$

   because $\{x^P_{n,j}\}$ is drawn from the prediction distribution, $f_{X_n|Y_{1:n-1}}(x_n|y^*_{1:n-1})$.
4. Finally, we resample from $\{x^P_{n,j}\}$ with weights proportional to $\mathbb{P}[y^*_n|x_n;\theta]$. The resampled points represent particles drawn from the filtering distribution at time t_n.
5. We repeat steps 1 through 4 for each time point, and then multiply the final vector of estimated conditional likelihoods (Equation 12) to obtain an unbiased estimate of the likelihood. Alternatively, we can take the log of each conditional likelihood and then sum them to retrieve the log likelihood.

SMC is commonly known as the particle filter because the Monte Carlo sample is described as a swarm of J particles that are propagated forward based on the process model and then filtered and altered to more closely fit the next data point [18].

Now that we have determined how to estimate the log likelihood for our model, we outline how we will maximize it to find the optimal θ.

5.2.2 Maximum Likelihood through Iterated Filtering

Iterated filtering is a maximum-likelihood approach devised by King et al. (2016) that depends on SMC [18]. The **Appendix** provides a more detailed algorithm for iterated filtering, but we provide an overview of the technique below.

We begin by defining a set of values for our parameter vector θ and a fixed number of iterations, M. For every iteration, we apply a basic particle filter (defined in *Section 5.2.1*) to the model and add stochastic perturbations to the parameters so that they take a random walk through time. At the end of the time series, we recycle the set of parameters as starting parameters for the next iteration but with a smaller random walk variance than the previous iteration. After completing the M iterations, we obtain the Monte Carlo maximum likelihood estimate, θ_M, and its corresponding log likelihood. Theoretically, if we correctly define the random walk intensity and its cooling schedule (which dictates how quickly the perturbations decrease), this procedure should converge to the region in the parameter space that maximizes the log likelihood. Although the ideal cooling rate cannot be determined beforehand, we can empirically test different cooling schedules and check for which values convergence occurs.

Importantly, we should repeat this entire process a number of times with randomized starting values for the parameters, so that we ensure we do not obtain to a local maximum, but a global maximum. If many of our starting values converge to the same region in the parameter space, we can be more certain of our results.

6. EMPIRICAL RESULTS

In this section, we apply our model to the Harvard data and use iterated filtering to determine the maximum likelihood estimates (MLE). These estimates provide insight into the key characteristics of Harvard's outbreak which made it unique. Notably, we find that the reporting rate, ρ, is significantly higher than expected and that the length of the infectious period, γ^{-1}, is lower than the population average.

6.1 Optimal Parameters

6.1.1 Fixed Primitives

We first consider the parameters that we will fix in the model, rather than estimating through iterated filtering. A summary of these parameters are displayed in **Table 1**.

Symbol	*Description*	*Value*	*Units*	*Source*
τ	Date of intervention	75	—	Harvard records on interventions [12]
σ	Per-capita rate of transition from E to I	$\frac{1}{17}$	day^{-1}	Lewnard and Grad (2018) [6]
N	Total population	$20{,}000$	—	Harvard records on population size [9]

Table 1: List of fixed parameters used in mumps transmission model for Harvard

In **Table 1**, we set $\tau = 75$ because approximately around the 75th day, two interventions are employed by HUHS. The first intervention involves more aggressive diagnoses after receiving an email from the MDPH regarding negative PCR tests. The second involves an email to Harvard's campus raising awareness about a recent spike in mumps. Although interventions are administered at numerous points during the outbreak, the ones around Day 75 are particularly pertinent because (i) the incidence of mumps was highest at this point and (ii) these were the last interventions that occurred before the end of the outbreak.

Meanwhile, we set the flow rate between the exposed and infectious classes to $\sigma = \frac{1}{17}$ because, according to past literature, the average latent period lasts $\sigma^{-1} = 17$ days. While we fix the rate between the exposed and infectious compartments, we choose to estimate γ, the rate between the infectious and removed compartments. Unlike σ, we expect the rate of removal from a population to have some dependence on interventions. Specifically, if Harvard's email campaign and isolation strategy were effective, we should see people reporting their symptoms earlier to HUHS and being removed from the population immediately after.

Finally, we set $N \approx 20{,}000$ people based on records of Harvard's enrollment and employment.

6.1.2 Maximum Likelihood Estimates

Once we have our fixed parameters defined, we run global likelihood maximization with 20 different, randomized starting points, using the method defined in *Section 5.2.2*. Each of the 20 starting points should converge to comparable values of the log likelihood as well as similar regions in the parameter space, as this suggests we have found a global, not local, optimum. Indeed, we find that the final log likelihoods for all 20 iterations stabilize in the same interval. To demonstrate the stability of our iterated filtering algorithm, **Figure 6** displays convergence diagnostic plots for one randomized starting point.

Given that the global likelihood maximization technique successfully converges, we pull the maximum likelihood estimates for the nine parameters from the optimization round with the highest likelihood, -233.7943, and log likelihood standard error of 0.2582. The results are shown in **Table 2**.

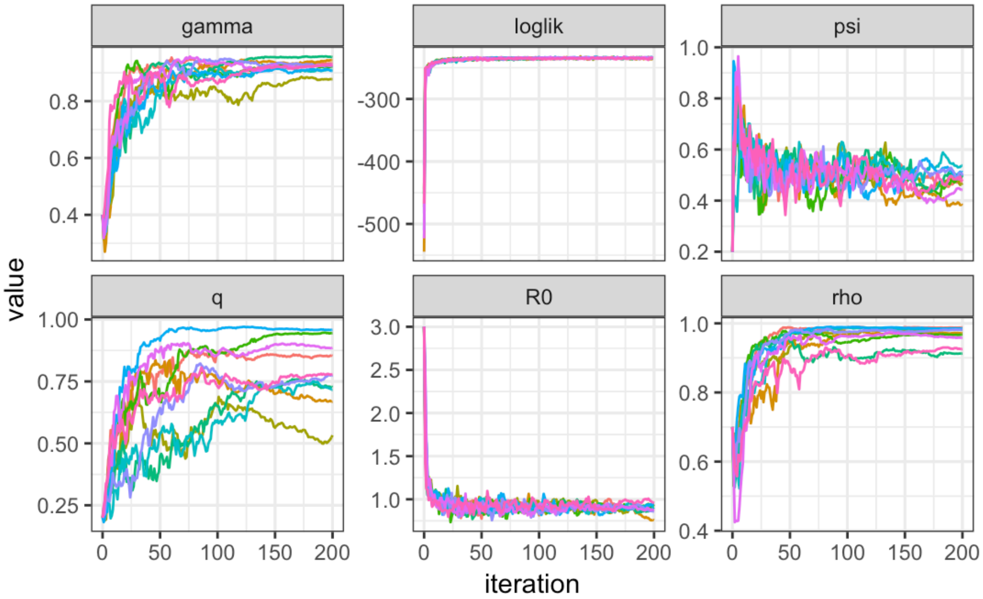

Figure 6: Convergence plots for a single starting point. We apply ten particle filters to the model with the same starting set of parameters and find that the algorithm converges to similar values of log likelihood. This suggests that our maximization algorithm is stable.

Symbol	*Description*	*Value*	*Units*	*Source*
$R0$	Basic reproduction number	0.8710		Estimated in *Section 6.1.2*
$\beta(t)$	Transmission rate	Time-dependent	day^{-1}	Calculated: $\beta(t) = \begin{cases} R_0\gamma, & t < \tau \\ R_0\gamma e^{-q(t-\tau)}, & t \geq \tau \end{cases}$
q	Effect of intervention	0.9422	—	Estimated in *Section 6.1.2*
$\lambda(t)$	Force of infection: rate at which susceptibles acquire disease	Time-dependent	day^{-1}	Calculated: $\lambda(t) = \beta(t)\frac{I(t)}{N}$
γ	Per-capita rate of transition from I to R	0.9641	day^{-1}	Estimated in *Section 6.1.2*
ρ	Proportion of infections reported	0.9698	—	Estimated in *Section 6.1.2*
ψ	Overdispersion parameter	0.5093	—	Estimated in *Section 6.1.2*
S_0	Initial proportion of susceptible persons	0.9995	—	Estimated in *Section 6.1.2*
E_0	Initial proportion of exposed persons	0.0004	—	Estimated in *Section 6.1.2*
I_0	Initial proportion of infectious persons	0.0001	—	Estimated in *Section 6.1.2*
R_0	Initial proportion of recovered persons	0	—	Estimated in *Section 6.1.2*

Table 2: List of parameters used in mumps transmission model that are estimated via iterated filtering or calculated using the estimated parameters. The values of these parameters help us understand certain characteristics of the Harvard outbreak.

We also construct 95% confidence intervals via profile likelihoods for these parameters in order to determine how sensitive the conclusions in **Table 2** are. The profile likelihood function for a parameter involves varying that parameter over a range of values and then maximizing the likelihood over the remaining parameters. For instance, as seen in **Figure 7**, we let ρ take on all values between $(0,1)$ and then, with one fewer parameter to optimize, find the new maximized likelihood estimates and associated log likelihood. We then compare the profile log likelihoods for ρ, expressed as $l_{profile}(\rho)$, to the log likelihood for the original MLEs, expressed as $l(\hat{\theta})$. The approximate 95% confidence interval is derived by checking which values of ρ satisfy the following inequality:

$$\{\, \rho: l(\hat{\theta}) \text{ - } l_{profile}(\rho) \,\} < (\tfrac{\chi_1^2}{2} = \tfrac{3.84}{2} = 1.92), \qquad \textbf{(13)}$$

where 1.92 is the cutoff determined using Wilks' Theorem [21]. We repeat this method for all other estimated parameters (other than the initial conditions: S_0, E_0, I_0, R_0) listed in **Table 2**.

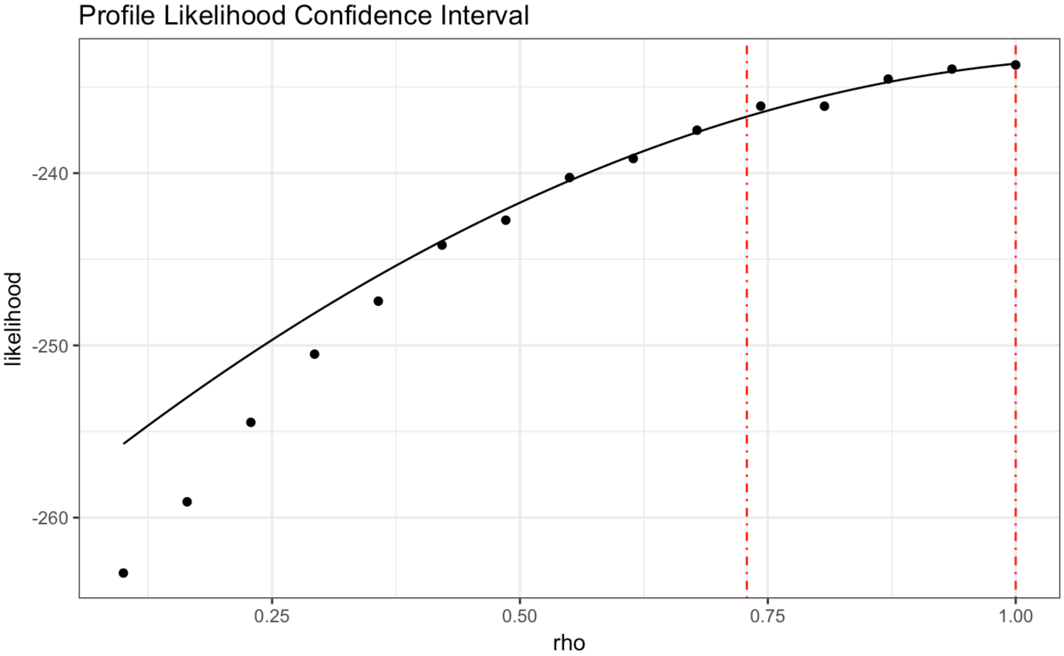

Figure 7: 95% confidence interval (denoted by dotted red lines) constructed via the profile likelihood method for ρ. Thus, we are 95% confident that the true value for ρ lies between (0.735, 1.000).

6.2 Parameter Interpretation

The maximum likelihood estimates give us insight into the different characteristics of Harvard's outbreak. In this section, we analyze the implications of the estimates for the initial conditions, basic reproduction number, intervention parameter, rate of removal, reporting rate, and overdispersion parameter.

6.2.1 Initial Conditions

We first begin with the estimated initial conditions, which represent the fraction of the total population in each compartment:

$$20000 * 0.9995 = 19990 \text{ susceptibles at } t_0$$

$$20000 * 0.0004 = 8 \text{ exposed at } t_0$$

$$20000 * 0.0001 = 2 \text{ infectious at } t_0$$

$$20000 * 0 = 0 \text{ removals at } t_0$$

These results are consistent with the data, given that on the first official day of the outbreak at Harvard, two mumps cases had been reported.

6.2.2 Reproduction Number

The basic reproduction number is 0.8710, with a 95% confidence interval of $(0.7241, 1.2408)$. As a reminder, the basic reproduction number is defined as the average number of secondary infections caused by an infectious individual when the entire population is susceptible. In other words, $R0$ is the growth rate of the disease at t_0 and thus if $R0 \geq 1$, an outbreak will occur. Although the MLE for $R0$ is below 1, the 95% confidence interval includes 1, and hence our results are indeed compatible with what happened on Harvard's campus.

6.2.3 Intervention Parameter

The parameter, q, evaluates the effect of an intervention on the outbreak. The transmission rate, $\beta(t) = R_0\gamma e^{-q(t-75)}$, decreases faster over time for larger values of q. As seen in **Table 2**, the estimate for q is 0.9422, with a 95% confidence interval of $(0.4715, 1.000)$. The wide interval implies that, given the data and small sample size, it is difficult to precisely estimate the impact of the interventions administered around Day 75. Nevertheless, because the confidence interval does not include 0, we can reject our null hypothesis that the interventions had no effect. We shall analyze these effects further in *Section 7*.

6.2.4 Rate of Removal

In *Section 6.1.1*, we hypothesized that γ^{-1} – the number of days an infectious person remains in the population – would be smaller at Harvard than the average period of infectiousness for mumps because of Harvard's isolation strategy and email awareness campaign.

Indeed, while the expected duration of infectiousness is five days, the maximum likelihood estimate reports that a Harvard case is removed after $\frac{1}{0.9641} = 1.0372$ days of infectiousness. The confidence interval for γ is $(0.3213, 1.000)$, implying that infectiousness can last anywhere from 1 to 3.1124 days. These results are consistent with what we found in the section on clustering, because the mean lag time (between experiencing symptoms and getting treated and isolated) for the Harvard cluster was 1.568 days.

6.2.5 Reporting Rate

The fraction of cases that were actually reported and accounted for, represented by ρ, is 0.9698, with a narrow 95% confidence interval of $(0.735, 1.000)$. Lewnard and Grad (2018) estimate that approximately 4.0% of all mumps cases are reported in the US, and thus, the reporting rate at

Harvard, even if slightly inflated, is truly exceptional [6]. There are likely two reasons for such a high reporting rate. First, the email awareness campaign by HUHS encouraged people to visit health services at the first sign of symptoms and keep an eye on their surroundings. A network of people – from resident deans to athletic coaches – were crucial in reporting students and employees who seemed at-risk for mumps. Second, more aggressive diagnoses by HUHS, particularly towards the end of the outbreak, ensured that more cases were detected than usual.

6.2.6 Overdispersion Parameter

Finally, our overdispersion parameter, ψ, is 0.5093, implying that the actual data has more variability than expected under the assumed distribution. As a reminder, had ψ been approximately 0, the variance in our measurement model would have simplified to the variance for a binomial distribution. However, because the 95% confidence interval for ψ is $(0.5065, 0.8569)$ and thus does not include 0, we justify the modelling decision of representing the number of cases as an overdispersed binomial.

We can hypothesize reasons for the overdispersion. Demographic and environmental stochasticity can vary the number of reported cases. A student in the midst of midterm season may be less likely to report symptoms, afraid that it will prevent him from studying for an exam. Alternatively, overdispersion can be the result of interventions. After an email was sent out about the increasing spread of mumps, vigilance and reporting likely increased temporarily before returning to the average.

6.3 Simulations at the MLE

In order to visually check the fit of our model to the data, we run stochastic simulations of Harvard's outbreak using the parameter values from **Table 2**.

Figure 8 proves that our data is similar in size and pattern to many of our simulations, such as Simulation 1, 2, and 4. Additionally, shortly after day 75 (the time of the primary intervention, as explained in *Section 6.1.1*), we consistently see a decrease in the number of observed cases.

The variability in the simulations (such as Simulation 2 versus Simulation 8) can partly be attributed to the randomness inherent in a stochastic model as well as the overdispersion parameter. However, variability also exists because the maximum likelihood estimate for the basic reproduction number is slightly below 1, which implies that sometimes, no outbreak will occur. The number of cases will remain around 0, thus explaining Simulation 8.

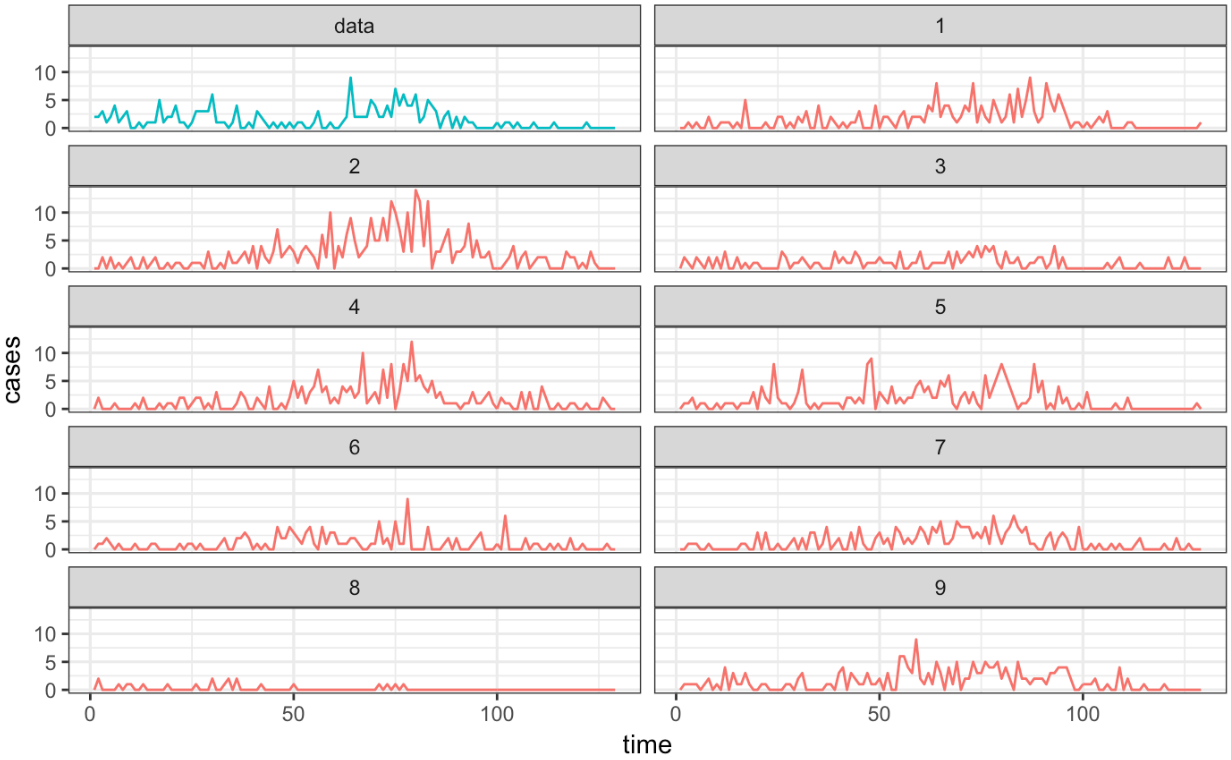

Figure 8: Nine simulations of the final model evaluated at the maximum likelihood estimates. Comparisons to the actual data show that many of the simulations (particularly Simulation 1, 2, and 4) have similar patterns.

7. INTERVENTION ANALYSIS

We now analyze the effects of the three interventions – the email campaign, more aggressive diagnoses, and isolation of infectious persons – on outbreak duration and size. We determine that the first two techniques, in combination, decrease the size of the outbreak by approximately two-thirds. Meanwhile, by comparing the Harvard outbreak to an outbreak at Ohio State University (OSU) in 2014, we find that the isolation policy seems to lead to a smaller average infectious period for Harvard patients, which in turn shortens the length of the outbreak.

7.1 Effect of Vigilance

The vigilance at Harvard University was unparalleled during the mumps outbreak. Anecdotally, students and faculty alike were careful of both catching and spreading symptoms. The model confirms this with the high reporting rate at Harvard ($\rho = 0.9698$) and low period of infectiousness ($\gamma = 0.9641$). The most obvious reason for this can be attributed to the HUHS email campaign as well as the constant dialogue around mumps at Harvard, from newspapers like *The Crimson* and *Boston Globe* to national news coverage on NBC News [22, 23]. Additionally, the criteria that HUHS used to diagnose and isolate potentially infectious people became increasingly more expansive as their understanding of the disease and the shortcomings of the PCR tests improved. This ensured that infectious people who normally may not have been detected were still being removed from the population.

In order to understand to what extent these interventions made a difference on the trajectory of the outbreak, we can perform detailed analysis of the parameter q, which quantifies the effect of the two interventions occurring around Day 75.

7.1.1 Outbreak Size

To determine the effect of these vigilance-increasing interventions on outbreak size, we perform a comparative analysis of a scenario with the interventions versus a scenario without the interventions.

Controlling for all other parameters, we run two sets of simulations at the maximum likelihood estimates. The first set of simulations fixes q at 0.9422 (value obtained from **Table 2**) while the second set of simulations sets q to 0, assuming that no interventions occurred around Day 75. We then choose the simulation from each set with the median outbreak size and compare their cumulative number of cases over time (see **Figure 9**).

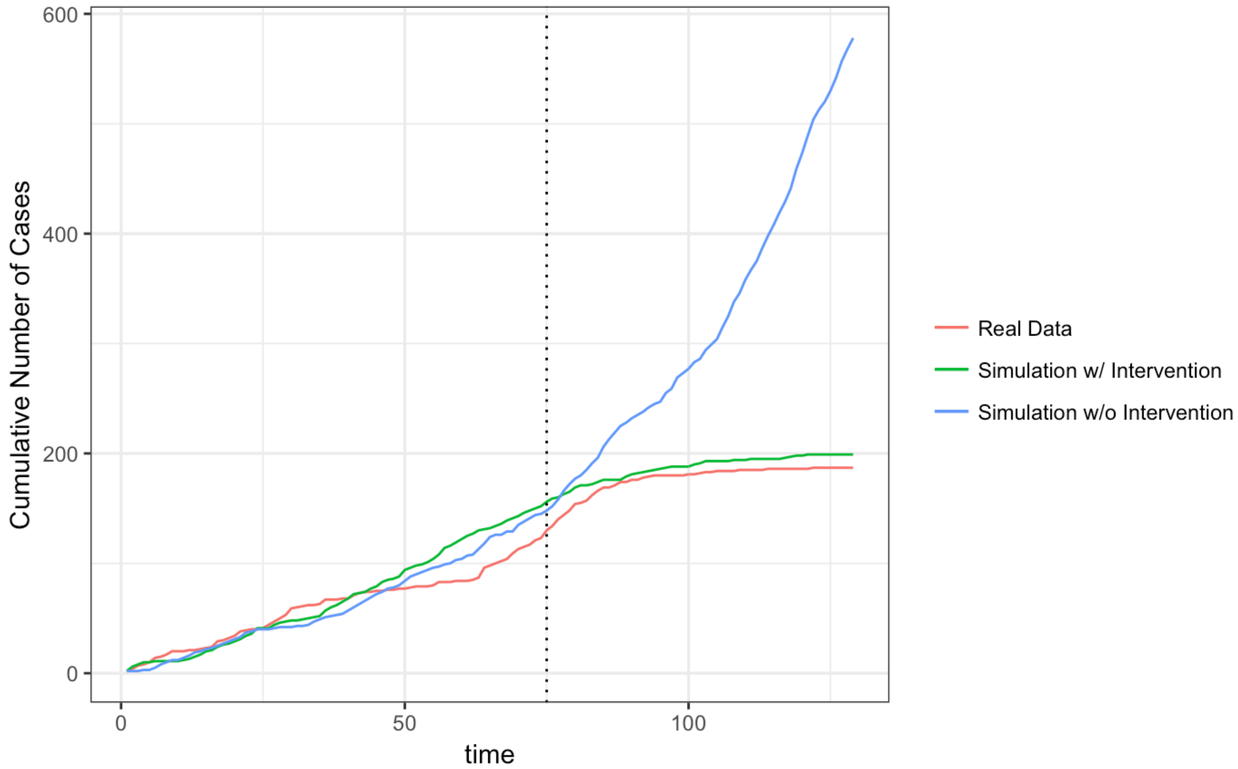

Figure 9: Comparison of the cumulative number of cases over time for the true Harvard data, the simulation with interventions, and the simulation without interventions. Dotted line represents the timing of the interventions, Day 75. The outbreak size is approximately three times as large without the interventions administered around Day 75.

Figure 9 demonstrates how, by the final day of the outbreak (Day 130), the simulation without the interventions is approximately three times the size of Harvard's actual outbreak. These results also indicate that the outbreak would have lasted much longer, if not for these vigilance-increasing strategies.

Because interventions seem to have drastically affected outbreak size, we perform additional analysis to determine if administering them earlier could have further reduced the number of mumps cases. We alter the day of the intervention (recall that we had earlier fixed τ to 75) to take on values between 1 and 74. Subsequently, we run numerous simulations for each of these 74 cases, pull the final outbreak size from the median simulation, and calculate the reduction in outbreak size.

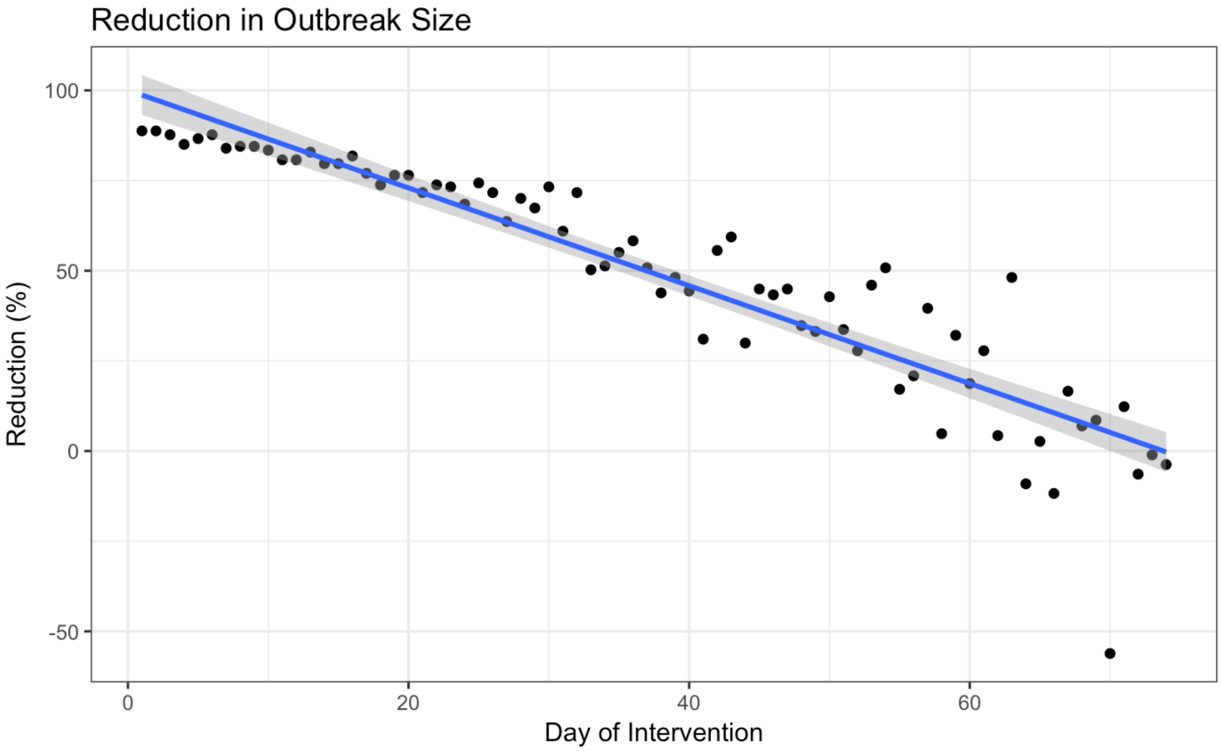

Figure 10: The percentage we expect outbreak size to decrease by if the date of intervention is moved up. There is a significant linear relationship between the predictor and response variable.

The linear regression in **Figure 10** determines the exact relationship between the day the intervention is administered and the reduction of the outbreak:

$$Reduction\,\% = \beta_0 + \beta_1 * day_{intervention} \tag{14}$$

We obtain statistically significant results for this regression, in which the intercept, β_0, is 100.0677, and the coefficient for the intervention date, β_1, is -1.3558, with p-values well below the significance level of 0.05. For every day we delay the interventions, outbreak size increases by 1.3558 percentage points, a non-trivial amount. So, for example, if we arbitrarily decide that the interventions had been administered on Day 40 instead, then Harvard would have seen approximately 90 fewer cases.

7.1.2 Limitations

There are, however, limitations in this analysis. Without conducting a randomized control trial, it is infeasible to understand the true effects of interventions on outbreak size. Part of the reason q may have such a large value is not because of the interventions, but rather because of confounding factors that we cannot control for in this analysis. For instance, Day 75 falls in late April, shortly before students finish the semester and leave campus, which would naturally decrease the number of potential infections. Moreover, with graduation approaching and Harvard commencement at risk because of the outbreak, both HUHS and students probably began taking extra precautions to prevent the spread of mumps.

An additional limitation in this analysis is the difficulty in differentiating between the effects of the two vigilance-increasing interventions – the formal change in HUHS procedures, which occurred on Day 60, and the email that was sent on Day 75. Intuitively, we can assume that the revised HUHS procedures would have a greater effect, but there is also mathematical reasoning to back this assumption. Given that the incubation period for mumps ranges between two to three weeks, we should expect q to be relatively small if we estimate it directly after an intervention (since the number of newly infectious cases are still a result of the “old” behavior). Only after a few weeks should we begin seeing the impact of the “new” behavior induced by the intervention. However, on Day 75 itself, q is relatively large (at 0.9422). This could either mean that the HUHS email had an immediate and drastic effect on the campus *or* that q mainly measures the effect of the intervention from Day 60. Given that the previous five emails from HUHS did not *immensely* impact the size of the outbreak (although they did improve reporting rates), it is unlikely that the final email could have, on its own, changed the trajectory of the outbreak. Hence, we conclude that the new HUHS procedures primarily explain the magnitude of q.

7.2 Effect of Isolation

Arguably the most critical intervention utilized by HUHS was its isolation requirement. Infectious people were physically removed from the population and placed in separate Harvard housing. Most schools require that students remain in their rooms for the course of the infectious period, but there is no reliable way to ensure that patients actually adhere to these rules. It is likely that they still engage in some contact with roommates or close friends.

Because the isolation requirement was implemented throughout the course of the outbreak, it is difficult to perform a before and after comparison as we do in *Section 7.1.1* in order to understand the impact of this intervention. One alternative, however, is to evaluate the characteristics of the Harvard mumps outbreak against that of another college campus, like Ohio State University, where mumps cases were not formally isolated.

7.2.1 Outbreak at Ohio State University

7.2.1.1 Data

In 2014, a massive outbreak of mumps occurred in central Ohio, with the majority of cases linked to Ohio State University in Columbus. The outbreak began at OSU in February 2014, and by March 21st, health authorities had reported 63 mumps cases, with 45 tied to the university [24].

Although most universities do not release case data to the public, they do need to report the numbers to the CDC. The CDC then publishes weekly reports of mumps cases across each state. We thus access the CDC's *Morbidity and Mortality Weekly Report* from 2014 and filter the dataset for cases originating in Ohio [25]. One drawback of this dataset is that, unlike the Harvard dataset, we only know the number of cases per week, which will make our analysis and parameter estimations less precise. Furthermore, we cannot guarantee that all the cases in this dataset are linked to the university itself. Fortunately, we know from news reports that most cases in Ohio occurred on campus during the first half of 2014 [24].

7.2.1.2 Characteristics of Outbreak

The outbreak at the university commenced in February 2014 and peaked in early April with 96 cases in one week. By summer and early fall, the number of cases had dramatically dropped and stabilized. We therefore restrict our analysis of the outbreak to the time between Week 1 and Week 40 of 2014, in which there were a total of 528 cases (see **Figure 11**).

Because we were unable to speak to public health officials at the university, the exact timeline and range of interventions administered over this period are not known. Like at Harvard, advisories were published by the university, notifying students of the issue and how to prevent its spread; these probably had a similar effect of increasing vigilance on campus. However, OSU did not formally isolate infectious persons. One notice published by OSU's medical center reads: "Stay at home for five days after symptoms (salivary gland swelling) begins (required by Ohio law OAC 3701-3-13, (P)); avoid school, work, social gatherings, and other public settings" [26].

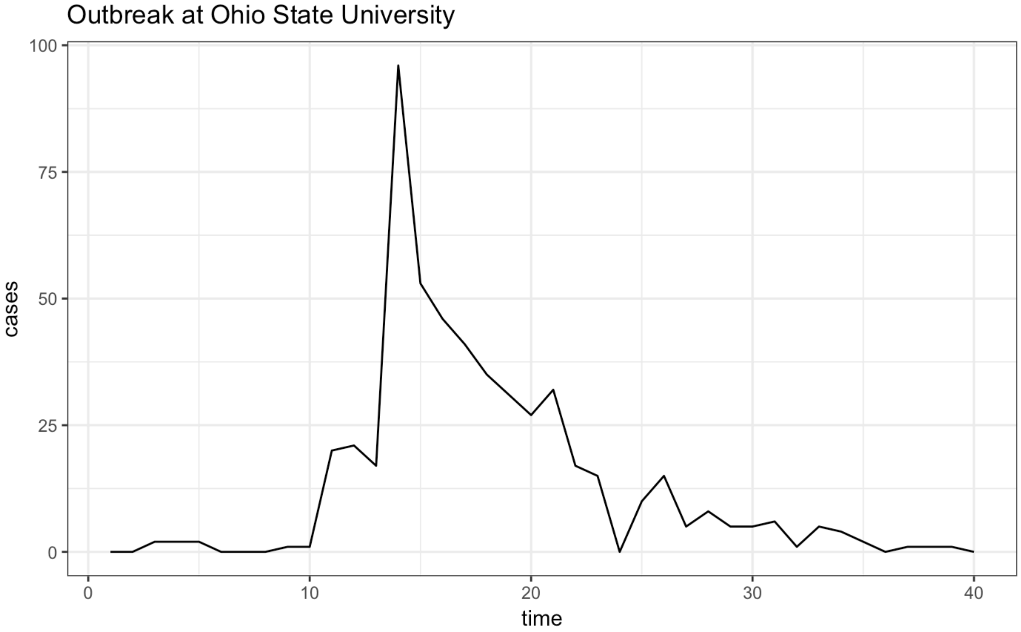

Figure 11: Number of weekly mumps cases in Ohio (particularly Ohio State University) between January and September 2014. There were a total of 528 cases during this time period.

7.2.1.3 Model and Results

To simulate the OSU outbreak and infer the underlying parameters, we use the same POMP model structure – with a stochastic compartmental model as the process and an overdispersed binomial as the measurement – that we did for Harvard. There are, however, a few significant differences.

First, because no publicly available knowledge exists of unique measures that OSU took to mitigate the spread of mumps, we do not include an intervention effect parameter, q, in the model. This also suggests that the transmission rate, β, no longer changes over time.

Second, recall that the unit for time is in weeks, rather than days. In *Section 5.1.1*, we discussed that, although stochastic compartmental models are technically continuous, we simplify the computation by making discretized approximations, with a small value for the time step δ. However, if we fix δ to 1, as we do for the Harvard model, then the transition rates and compartment sizes will be held constant over an entire week before being updated. This is unrealistic, and thus we now fix δ to $\frac{1}{7}$, so that the time step occurs each day.

Finally, we change the fixed parameters to reflect the qualities of this dataset. N, the total population, becomes $60{,}000$, to represent the size of the OSU campus. Meanwhile, because σ^{-1} equals the length of the latent period in days, we must change the units for σ:

$$\sigma = \frac{1}{17\,days} * \frac{7\,days}{1\,week} = \frac{7}{17}$$

With these adjustments, we finalize our model and apply iterated filtering to it to find the optimal parameters. We find that the maximized log likelihood is -128.5178, with a standard error of 0.1159. The MLEs are shown in **Table 3**:

Symbol	*Description*	*Value*	*Units*	*Source*
$R0$	Basic reproduction number	3.3307		Estimated in *Section 7.2.1.3*
β	Transmission rate	0.5399	day^{-1}	Calculated: $\beta = R_0\gamma$
$\lambda(t)$	Force of infection: rate at which susceptibles acquire disease	Time-dependent	day^{-1}	Calculated: $\lambda(t) = \beta\frac{I(t)}{N}$
γ	Per-capita rate of transition from I to R	0.1621	day^{-1}	Estimated in *Section 7.2.1.3*
ρ	Proportion of infections reported	0.0148	—	Estimated in *Section 7.2.1.3*
ψ	Overdispersion parameter	0.7410	—	Estimated in *Section 7.2.1.3*
S_0	Initial proportion of susceptible persons	0.9998	—	Estimated in *Section 7.2.1.3*
E_0	Initial proportion of exposed persons	0.00019	—	Estimated in *Section 7.2.1.3*
I_0	Initial proportion of infectious persons	0.00001	—	Estimated in *Section 7.2.1.3*
R_0	Initial proportion of recovered persons	0	—	Estimated in *Section 7.2.1.3*

Table 3: Maximum likelihood estimates obtained using iterated filtering techniques for the Ohio outbreak. These estimated parameters help us understand the characteristics of this outbreak.

We run stochastic simulations of the model evaluated at the MLEs in order to visually determine its fit to the OSU data (see **Figure 12**). While Simulations 1 and 2 seem to emulate the patterns of the data quite well, Simulation 3 greatly overestimates the outbreak size. The variance in results is a natural side effect of using a stochastic model. Moreover, we see that the infectious curve for the data is smoother than the jagged simulations; this can be attributed to the high value of ψ, which increases the variability in the number of weekly observed cases.

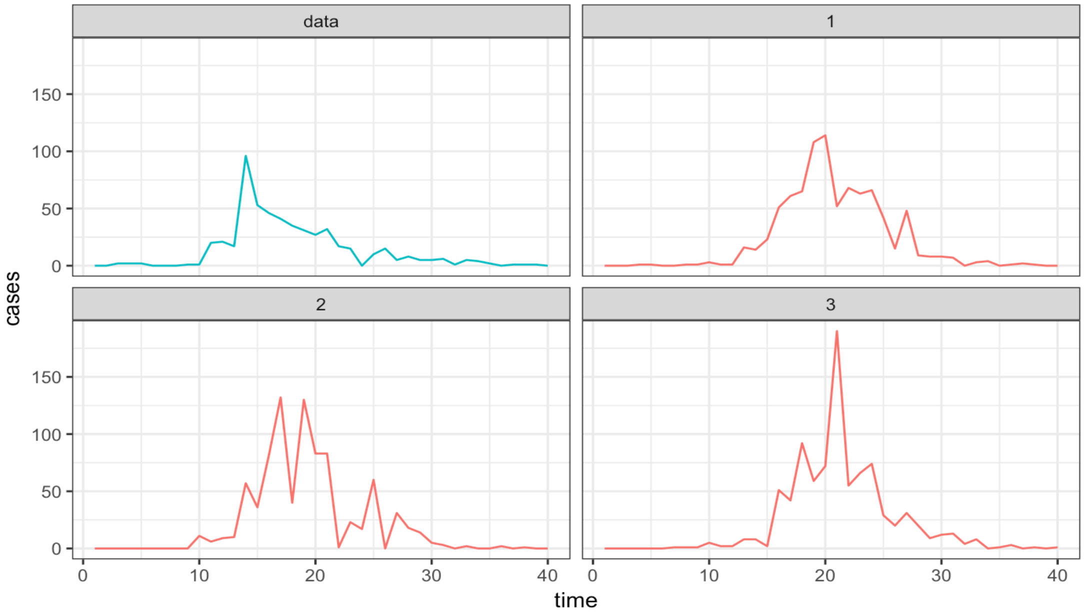

Figure 12: Three simulations of the final OSU model evaluated at the maximum likelihood estimates. Simulation 1 and 2, in particular, have similar patterns as the data. Note that the variability in weekly observed cases is due to a high value of the overdispersion parameter, ψ.

7.2.2 Comparison of Harvard and OSU Parameters

The maximum likelihood estimates for Harvard and OSU are different on multiple accounts. In this section, we explore the three parameters (basic reproduction number, reporting rate, and rate of transition from the infectious to removed class) that are most dissimilar between the two schools, and propose potential explanations for their differences.

Firstly, OSU's basic reproduction number, which indicates the initial growth rate of the outbreak, is over three times that of Harvard. Although wide, its confidence interval, $(2.068, 7.000)$, does not overlap with the confidence interval for Harvard's $R0$, confirming this difference. Harvard's isolation policy best explains this difference because it physically prevents infectious persons from causing multiple secondary infections, thus suppressing the growth of the outbreak.

Secondly, OSU's reporting rate is extremely low, at approximately 1.5%, compared to Harvard's 98%. In fact, the reporting rate at OSU is closer to population-wide estimates of this parameter (4.0%, according to Lewnard and Grad) [6]. Although ρ's 95% confidence interval of $(0, 0.7143)$, constructed via profile likelihoods, is quite large, the difference in Harvard and OSU's values is still statistically significant, with non-overlapping confidence intervals. We do not have access to OSU's diagnostic procedures nor do we know the extent of their email awareness campaign, but we hypothesize that a lack of one or both of these may explain at least a portion of the dissimilarity in the two schools' reporting rates.

Finally, the rate of recovery, γ, is 0.1621, so that the period of infectiousness is $\frac{1}{0.1621} = 6.169$ days, approximately six times as large as Harvard's period of infectiousness. γ is by far the most critical parameter in evaluating the effectiveness of isolation. Without a formal isolation policy, an infectious person cannot enter the removed class until fully recovered, which happens between five and seven days after the onset of symptoms. With an isolation policy, on the other hand, an infectious person enters the removed class the moment they have been isolated because they can no longer spread their infection to anyone else in the population.

With a much shorter period of infectiousness, it is no surprise, then, that Harvard's outbreak duration is less than OSU's. While Harvard's lasted 18 weeks, the OSU outbreak lasted over 25 weeks. Moreover, Harvard's isolation strategy allowed, for the most part, containment of the disease on campus. In contrast, infectious students at Ohio State University could still interact with the greater Columbus population if they chose not to self-isolate, which likely led to the rapid spread of the disease throughout both campus and central Ohio.

7.2.3 Limitations

Although there are clear differences in the OSU and Harvard parameters, we must be cautious in taking the OSU estimates at face value.

Given that the OSU data consists of weekly reports rather than daily reports of cases, we should expect the estimates for the parameters to be less accurate. Indeed, this explains the wide confidence intervals for estimates of ρ and $R0$ in the OSU model.

Furthermore, as covered in our discussion of the dataset, the cases are not solely linked to the university. Numerous cases in the data occurred in the greater Columbus area, suggesting that the parameter estimates do not only account for the dynamics of mumps on campus. They also incorporate the dynamics of mumps across a much less-structured area. So, for instance, even if awareness and reporting of mumps on campus are high, an unsuspecting person in Columbus may not know to report his symptoms; this would lead to a lower estimated reporting rate than expected.

Lastly, OSU's population size, N, is three times that of Harvard. Because the force of infection is a function of raw population size (as we see from the equation for $\lambda(t)$ in **Table 3**), interventions used at Harvard simply may not have worked as well at OSU. To more accurately quantify the efficacy of interventions like isolation, we would need to conduct a comparative analysis between Harvard and a school of similar size. Moreover, we would have to understand how such a policy scales with population size.

Thus, while this investigation is useful in gaining a broad idea of the characteristics of other universities' outbreaks, we should be careful not to over-generalize and overestimate the impact of Harvard's isolation efforts. As discussed in *Section 7.1.2*, the most promising method to determine the exact effect of isolation strategies is through a randomized control trial.

8. CONCLUSION

8.1 Summary

Throughout this paper, we construct and estimate a mathematical model for the transmission of mumps on college campuses. Unlike most models of infectious disease, which opt for deterministic representations, the stochastic model that we build is adaptable to small populations and accounts for the noisiness and incompleteness of data in its structure. Moreover, it incorporates a parameter that measures the effect of interventions added after time τ. Importantly, while most literature today focuses on mumps prevention – such as administering third MMR doses to college-age students – this paper provides quantitative backing for more immediate and less costly approaches to mitigating the spread of mumps. Future work should pursue understanding the precise effects of each control intervention in isolation.

8.2 Recommendations

We conclude with a set of recommendations that are highly effective, at least in combination with one another, in reducing outbreak size and duration.

We determine that, although the HUHS email awareness campaign increased the reporting rate of symptoms, it alone could not have ended the outbreak. Nevertheless, universities should not undervalue the importance of raising awareness and vigilance and should consider sending weekly updates on the spread of disease across campus.

We also find that more informed HUHS diagnoses, beginning Day 60, helped decrease outbreak size by one third and seemed to most directly lead to the end of the outbreak. Therefore, the CDC should provide more detailed instructions regarding the diagnosis of mumps in vaccinated persons and the shortcomings of PCR tests. For each passing day that medical centers did not have this information, we estimate that outbreak size increased by approximately one percentage point.

Finally, although difficult to quantify the impact of HUHS isolation policies, Harvard's period of infectiousness was approximately $\frac{1}{6}^{th}$ that of other universities, leading to a shorter and more contained outbreak. Universities with less resources to formally isolate infectious cases should consider devising stricter guidelines for what self-isolation looks like for students. Additionally, to ease this process of self-isolation, they should consider setting up a system of meal delivery to the infectious person's room and providing moral and emotional support through daily phone check-ins [11].

Ultimately, although infectious disease outbreaks at universities pose a unique public health problem because of the increased transmission rates, the Harvard mumps outbreak provides insight into the power of straightforward interventions. Perhaps this public health challenge is more manageable than it appears. Simple awareness and understanding can make all the difference.

REFERENCES

[1] Costill, David. “College Campus Outbreaks Require Timely Public Health Response.” *Infectious Diseases in Children*, Oct 2015. Accessed December 1, 2017.

[2] Massachusetts Immunization Program. “Mumps Update for Higher Education Student Health Services.” Oct 6, 2016.

[3] Centers for Disease Control and Prevention. “Mumps Cases and Outbreaks.” June 1, 2016.

[4] Barskey, Albert E., John W. Glasser, and Charles W. LeBaron. “Mumps Resurgences in the United States: A Historical Perspective on Unexpected Elements.” *Vaccine* 27, no. 44 (October 19, 2009): 6186–95.

[5] Li, Yong, Xianning Liu, and Lianwen Wang. “Modelling the Transmission Dynamics and Control of Mumps in Mainland China.” *International Journal of Environmental Research and Public Health* 15, no. 1 (December 26, 2017).

[6] Lewnard, Joseph A., and Yonatan H. Grad. “Vaccine Waning and Mumps Re-Emergence in the United States.” *Science Translational Medicine* 10, no. 433 (March 21, 2018): eaao5945.

[7] Shah, Minesh, Patricia Quinlisk, Andrew Weigel, Jacob Riley, Lisa James, James Patterson, Carole Hickman, et al. “Mumps Outbreak in a Highly Vaccinated University-Affiliated Setting Before and After a Measles-Mumps-Rubella Vaccination Campaign—Iowa, July 2015–May 2016.” *Clinical Infectious Diseases* 66, no. 1 (January 6, 2018): 81–88.

[8] Harvard University Health Services. “Isolation Expenses (2016-2017).” Accessed January 24, 2018.

[9] Harvard University Health Services. “Mumps List for CDC.” Accessed January 24, 2018.

[10] “University Quarantines Students Infected with Mumps.” *The Harvard Crimson*, March 7, 2016. Accessed March 24, 2018.

[11] Barreira, Paul and Susan Fitzgerald. Personal interview. January 17, 2018.

[12] Harvard University Health Services. “Correspondence re Mumps 2016, Spring to Summer.” Accessed January 24, 2018.

[13] Bitsko, Rebecca H., Margaret M. Cortese, Gustavo H. Dayan, Paul A. Rota, Luis Lowe, Susan C. Iversen, and William J. Bellini. “Detection of RNA of Mumps Virus during an Outbreak in a Population with a High Level of Measles, Mumps, and Rubella Vaccine Coverage.” *Journal of Clinical Microbiology* 46, no. 3 (March 2008): 1101–3.

[14] Massachusetts Department of Health. Massachusetts Mumps Case Data between 2015 and 2017. Accessed March 8, 2018.

[15] Lekone, P. E. and Finkenstädt, B. F. “Statistical Inference in a Stochastic Epidemic SEIR Model with Control Intervention: Ebola as a Case Study.” Biometrics, 62: 1170-1177 (June 2006).

[16] He, Daihai, Edward L. Ionides, and Aaron A. King. “Plug-and-Play Inference for Disease Dynamics: Measles in Large and Small Populations as a Case Study.” *Journal of The Royal Society Interface* 7, no. 43 (February 6, 2010): 271.

[17] Brauer, Brauer, Fred, Van den Driessche, Pauline, & Wu, Jianhong. (2008). *Mathematical epidemiology* (Lecture notes in mathematics (Springer-Verlag) ; 1945). Berlin: Springer.

[18] Aaron A. King, et al. “Statistical Inference for Partially Observed Markov Processes via the R Package Pomp.” *Journal of Statistical Software*, vol. 69, no. 1, 2016, pp. 1–43.

[19] King, Aaron A. and Edward L. Ionides. “Likelihood-based inference for POMP Models.” Lecture notes. Accessed February 4, 2018.

[20] Martcheva, M. *An Introduction to Mathematical Epidemiology*. 1st ed. 2015. ed., Springer US : Imprint: Springer, 2015.

[21] Ionides, Edward L., Carles Breto, Joonha Park, Richard A. Smith, and Aaron A. King. “Monte Carlo Profile Confidence Intervals.” *ArXiv:1612.02710 [Stat]*, December 8, 2016.

[22] Freyer, Felice J. “Harvard Gets the Mumps, but Students Fret and Joke.” *The Boston Globe*, May 8, 2016. Accessed March 20, 2018.

[23] Holt, Lester. “Harvard Commencement at Risk as Mumps Outbreak Grows.” *NBC Nightly News*. New York City, New York. April 28, 2016.

[24] Bixler, Jennifer and Greg Botelho. “Over 360 Cases of Mumps in Central Ohio, Most of Them Tied to OSU.” *CNN*, March 16, 2016. Accessed March 18, 2018.

[25] Centers for Disease Control and Prevention. *Morbidity and Mortality Weekly Report (2014)*. Accessed February 2, 2018.

[26] Ohio State University. “Mumps Outbreak FAQs.” January 2014. Accessed March 20, 2018.

[27] Rosenbaum, Lars, et al. “Optimization and Visualization of the Edge Weights in Optimal Assignment Methods for Virtual Screening.” *BioData Mining*, vol. 6, 2013, p. 7.

APPENDIX A: Diagnostic Procedures for Mumps

A.1 PCR Test

The Massachusetts Department of Public Health recommends the PCR test as the gold-standard diagnostic for mumps. But, while the positive PCR test indicates the presence of mumps virus RNA, negative PCR tests do not necessarily rule out mumps as a diagnosis. In particular, vaccinated individuals shed smaller amounts of virus for a shorter period of time. Thus, in outbreaks among two-dose vaccine recipients, mumps virus was only detected in samples from approximately 30-35% of case patients if the samples were collected within the first three days following onset of parotitis.

A.2 Harvard Diagnoses

At the beginning of the outbreak, Harvard struggled to diagnose mumps. Many nurses and clinicians were witnessing mumps for the first time and struggled to identify it, especially in vaccinated individuals. Occasionally, they took improper samples from infectious individuals, leading to negative PCR tests. Moreover, certain cases that came into the clinic too late often no longer had mumps in their upper respiratory tracts, also leading to negative results.

Harvard did not have a clear policy regarding negative PCR tests at the beginning of the outbreak. Many cases that were infectious were likely still ruled out and released back into the susceptible population simply because of their test results. However, upon recommendations from the MDPH mid-outbreak, HUHS developed stricter guidelines on how to handle negative PCR results and diagnoses of students.

Appendix B: Cluster Analysis

B.1 Gower Distances

Gower distances is a popular measure for proximity between mixed data types (such as continuous, binary, nominal, and ordinal). It computes the dissimilarity value between individuals by each variable, taking the variable type into account, and then averages the similarity values across all columns [27].

B.2 Elbow Plot

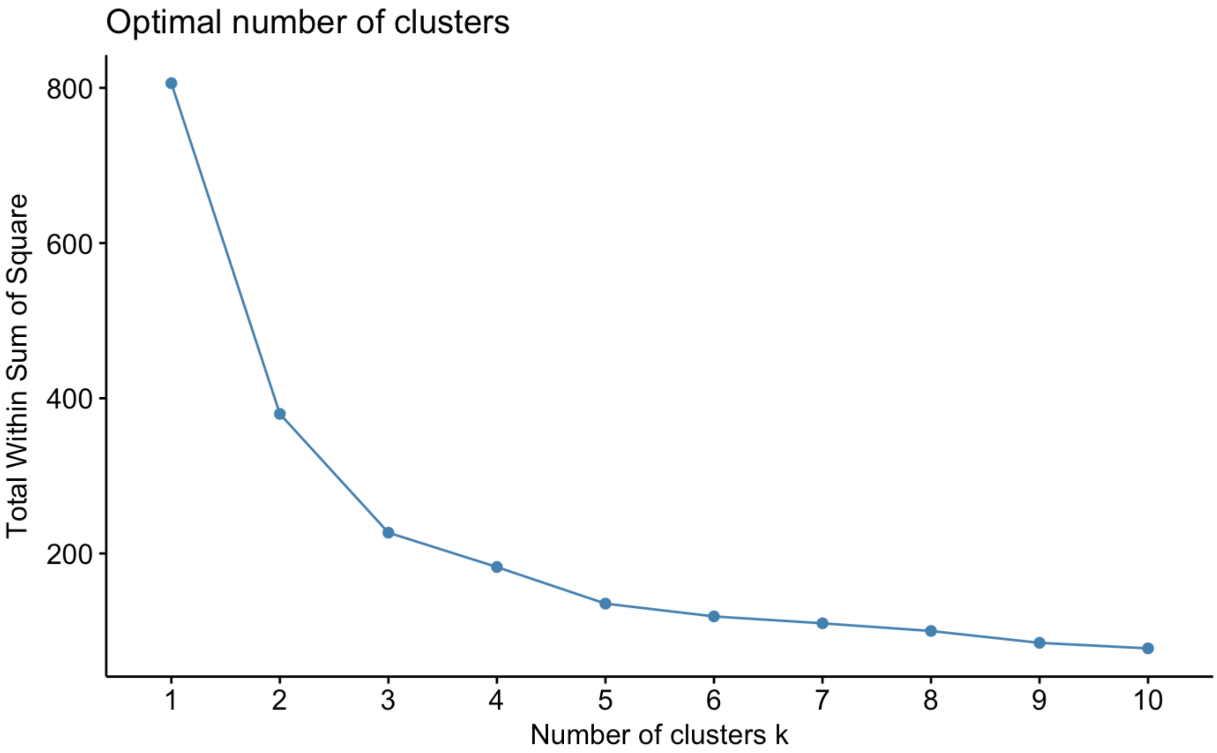

Figure 1: An elbow plot demonstrating that $k = 3$ is the optimal number of clusters to group the MDPH data. At $k > 3$, the gradient of the total intra-cluster variation, T_K, noticeably changes. After that point, the increase in dimensionality only minimally improves the accuracy of our clustering.

B.3 Cluster Visualization

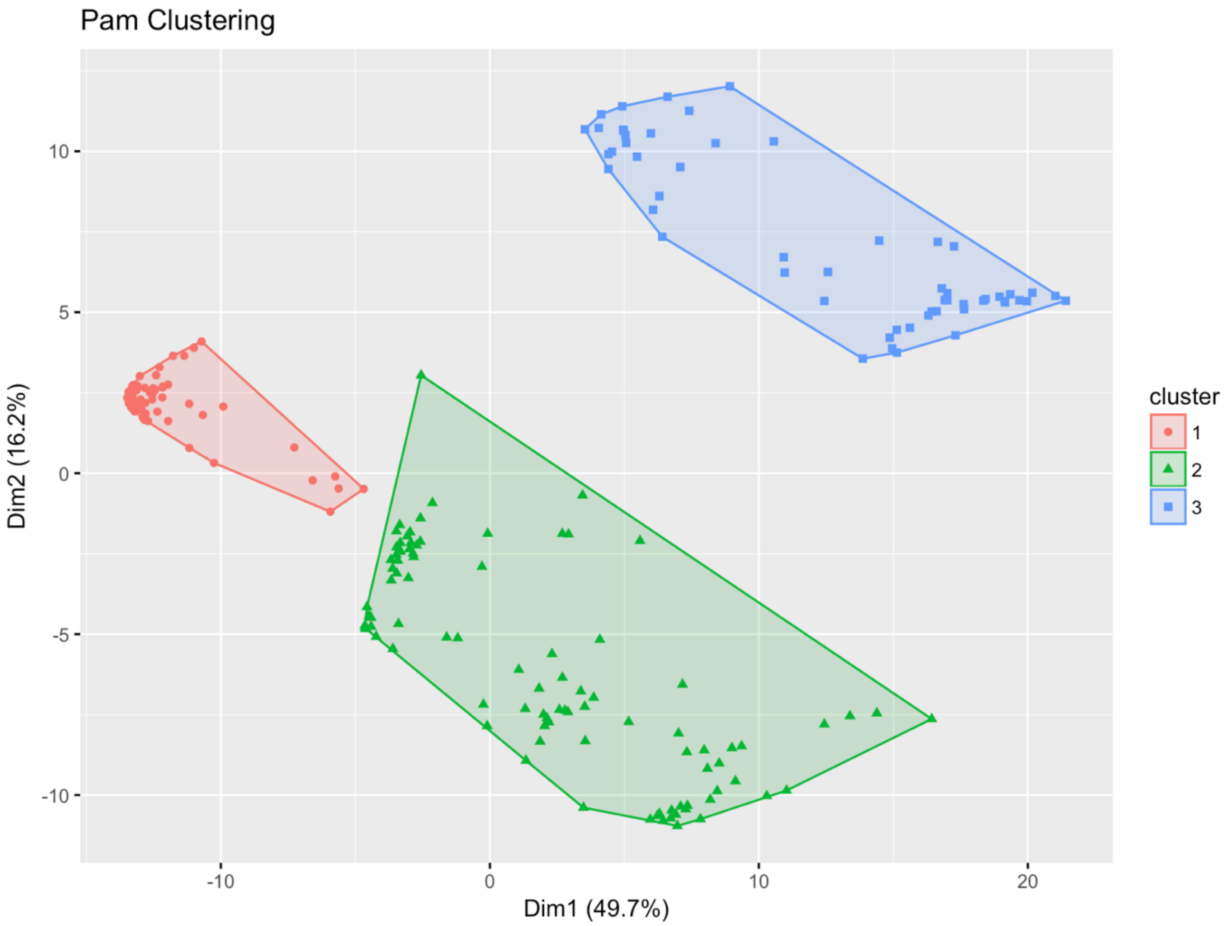

Figure 2: A visualization of the results of PAM clustering, in which Cluster 1 consists of Harvard-only data points. The clusters are graphically represented by the two dimensions that explain the most variance between the three groups.

Cluster 1 has the lowest within-cluster variance, given that its sum of squared distances between all pairs of points in the cluster are lower than that of Cluster 2 and 3. This suggests that there are unique features about Harvard's outbreak.

B.4 Cluster Summary Statistics

	Age	*Lag Time*	*Time of Cases*
	Cluster 1		
Min	18.00	0.000	233.0
1st Quartile	19.00	1.000	263.8
Median	20.00	1.000	294.5
Mean	21.16	1.568	302.6
3rd Quartile	22.00	2.000	310.0
Max	29.00	10.000	517.0
	Cluster 2		
Min	15.00	0.000	0.0
1st Quartile	19.00	1.000	285.5
Median	20.00	1.000	486.5
Mean	22.53	1.677	450.2
3rd Quartile	21.00	2.000	613.5
Max	69.00	12.000	719.0
	Cluster 3		
Min	14.00	0.000	71.0
1st Quartile	23.00	1.000	331.8
Median	27.00	2.000	550.0
Mean	30.14	2.357	506.1
3rd Quartile	35.00	3.000	698.5
Max	57.00	9.000	725.0

Table 1: Summary statistics of each cluster, providing insight into the distinctive characteristics of Harvard's outbreak.

APPENDIX C: Model Algorithms

C.1 Likelihood Estimation via Sequential Monte Carlo

Algorithm 1: Sequential Monte Carlo [18]

```
Input: simulator for $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$, the process model; evaluator for $f_{Y_n|X_n}(y_n|x_n;\theta)$, the measurement model; simulator for $f_{X_0}(x_0;\theta)$; parameter, $\theta$; data, $y^*_{1:N}$; number of particles, $J$.
1  Initialize filter particles: simulate $X^F_{0,j} \sim f_{X_0}(\,\cdot\,;\theta)$ for $j$ in $1{:}J$.
2  for $n$ in $1{:}N$ do
3      Simulate for prediction: $X^P_{n,j} \sim f_{X_n|X_{n-1}}(\,\cdot\,|X^F_{n-1,j};\theta)$ for $j$ in $1{:}J$.
4      Evaluate weights: $w(n,j) = f_{Y_n|X_n}(y^*_n|X^P_{n,j};\theta)$ for $j$ in $1{:}J$.
5      Normalize weights: $\tilde{w}(n,j) = w(n,j)/\sum_{m=1}^{J} w(n,m)$.
6      Apply Algorithm 2 to select indices $k_{1:J}$ with $\mathbb{P}[k_j = m] = \tilde{w}(n,m)$.
7      Resample: set $X^F_{n,j} = X^P_{n,k_j}$ for $j$ in $1{:}J$.
8      Compute conditional log likelihood: $\hat{l}_{n|1:n-1} = \log\left(J^{-1}\sum_{m=1}^{J} w(n,m)\right)$.
9  end
Output: Log likelihood estimate, $\hat{l}(\theta) = \sum_{n=1}^{N} \hat{l}_{n|1:n-1}$; filter sample, $X^F_{n,1:J}$, for $n$ in $1{:}N$.
```

Algorithm 2: Systematic resampling (Line 6 of Algorithm 1, Line 11 of Algorithm 3) [18]

```
Input: Weights, $\tilde{w}_{1:J}$, normalized so that $\sum_{j=1}^{J} \tilde{w}_j = 1$.
1   Construct cumulative sum: $c_j = \sum_{m=1}^{j} \tilde{w}_m$, for $j$ in $1{:}J$.
2   Draw a uniform initial sampling point: $U_1 \sim \text{Uniform}(0, J^{-1})$.
3   Construct evenly spaced sampling points: $U_j = U_1 + (j-1)J^{-1}$, for $j$ in $2{:}J$.
4   Initialize: set $p = 1$.
5   for $j$ in $1{:}J$ do
6       while $U_j > c_p$ do
7           Step to the next resampling index: set $p = p + 1$.
8       end
9       Assign resampling index: set $k_j = p$.
10  end
Output: Resampling indices, $k_{1:J}$.
```

C.2 Iterated Filtering

Algorithm 3: Iterated filtering [18]

Input: starting parameter θ_0; simulator for $f_{X_0}(x_0;\theta)$; simulator for $f_{X_n|X_{n-1}}(x_n|x_{n-1};\theta)$; evaluator for $f_{Y_n|X_n}(y_n|x_n;\theta)$; data, $y^*_{1:N}$; number of particles, J; number of iterations, M; length of parameter vector, p; cooling rate, $0 < a < 1$; perturbation scales, $\sigma_{1:p}$; initial scale multiplier, $C > 0$.

```
for $m$ in $1{:}M$ do
    Initialize parameters: $[\Theta^F_{0,j}]_i \sim Normal([\theta_{m-1}]_i, (Ca^{m-1}\sigma_i)^2)$ for $i$ in $1{:}p$, $j$ in $1{:}J$.
    Initialize states: simulate $X^F_{0,j} \sim f_{X_0}(\,\cdot\,;\Theta^F_{0,j})$ for $j$ in $1{:}J$.
    Initialize filter mean for parameters: $\bar{\theta}_0 = \theta_{m-1}$.
    Define $[V_1]_i = (C^2+1)(a^{m-1}\sigma_i)^2$.
    for $n$ in $1{:}N$ do
        Perturb parameters: $[\Theta^P_{n,j}]_i \sim Normal([\Theta^F_{n-1,j}]_i, (a^{m-1}\sigma_i)^2)$ for $i$ in $1{:}p$, for $j$ in $1{:}J$.
        Simulate prediction particles: $X^P_{n,j} \sim f_{X_n|X_{n-1}}(\,\cdot\,|X^F_{n-1,j};\Theta^P_{n,j})$ for $j$ in $1{:}J$.
        Evaluate weights: $w(n,j) = f_{Y_n|X_n}(y^*_n|X^P_{n,j};\Theta^P_{n,j})$ for $j$ in $1{:}J$.
        Normalize weights: $\tilde{w}(n,j) = w(n,j)/\sum_{u=1}^{J} w(n,u)$.
        Apply Algorithm 2 to select indices $k_{1:J}$ with $\mathbb{P}[k_j = u] = \tilde{w}(n,u)$.
        Resample particles: $X^F_{n,j} = X^P_{n,k_j}$ and $\Theta^F_{n,j} = \Theta^P_{n,k_j}$ for $j$ in $1{:}J$.
        Filter mean: $[\bar{\theta}_n]_i = \sum_{j=1}^{J} \tilde{w}(n,j)[\Theta^P_{n,j}]_i$ for $i$ in $1{:}p$.
        Prediction variance: $[V_{n+1}]_i = (a^{m-1}\sigma_i)^2 + \sum_j \tilde{w}(n,j)([\Theta^P_{n,j}]_i - [\bar{\theta}_n]_i)^2$ for $i$ in $1{:}p$.
    end
    $[\theta_m]_i = [\theta_{m-1}]_i + [V_1]_i \sum_{n=1}^{N} [V_n]_i^{-1}([\bar{\theta}_n]_i - [\theta_{n-1}]_i)$ for $i$ in $1{:}p$.
end
```

Output: Monte Carlo maximum likelihood estimate, θ_M.