CHARACTERIZING RISK OF TRANSMISSION OF RESPIRATORY VIRUSES AMONG SCHOOL CHILDREN USING PROXIES OF INFECTIOUS CONTACT AND GENETIC DATA

by
Jacob Carey

A thesis submitted to Johns Hopkins University in conformity with the requirements for the degree of Master of Science

Baltimore, MD

April 2016

Abstract

School-aged children play an important role in influenza. In this thesis we present a social network analysis of contacts among 746 students in 3 different schools and a genomic analysis of influenza viruses from 180 students in 9 schools. These schools are located in urban and suburban areas in and near Pittsburgh, Pennsylvania, USA and include elementary, middle, and high schools. We collected a proxy for social contact information using wireless sensor devices worn by the students, programmed so that they regularly record other nearby devices if they are closer than 4 meters. We analyzed these networks to identify the patterns of proximal interactions of children in different classes and grades, identify community structure within schools and examine the impact of the physical environment on proximal contacts between students. We created undirected weighted networks from the data recorded by these devices and conducted social network analyses of these networks. In elementary and middle schools we observe high number of intra-grade and intra-classroom contacts, and relatively low number of inter-grade contacts. However, in high schools, contact networks are well connected and mixed and are difficult to separate into specific grades or classrooms. The high modularity of lower grades suggests that assumptions of homogeneous mixing within schools in epidemic models may be inappropriate whereas lower modularity in high schools suggest that homogenous mixing assumptions may be more acceptable in these settings. Genomic data of the influenza viruses was analyzed via phylogenetic trees. This thesis was advised by Derek Cummings and read by Robert Scharpf.

Acknowledgements

Contents

List of Tables

List of Figures

Quantifying proximal contacts between school children during school, outside of school and during school closures

Introduction

The 2009 novel H1N1 pandemic disrupted schools, businesses, governmental entities and the general public with real and uncertain health risks of the virus and by debate regarding the most effective interventions. The H1N1 pandemic also reinforced the previous experience that "the fires of the epidemic are carried by healthy school-age children" (Glezen 1996). However, Influenza causes great economic damage every year due to lost productivity, medical treatment and preventative measures (Zhou et al. 2012). School summer holidays apparently helped reduce influenza transmission after the pandemic initial wave (Earn 2012). School reopening dates (fall, 2009) in the US coincided with a large second pandemic wave (Chao, Halloran, and Longini 2010). Numerous reports from CDC (2016) and European countries (WHO 2016) and studies (Cauchemez et al. 2011), document the central role of school-age children in spreading the pandemic virus. Children experience higher rates of infection, shed influenza virus for approximately twice as long as adults (Esposito et al. 2011), and are thought to have much higher rates of contacts than the rest of the population (Mossong et al. 2008).

Mixing patterns among school-children may be important to flu and other respiratory and close-contact infections (READ et al. 2012). A key unknown is the mixing rates and patterns of encounters relevant to the spread of infections during normal school times, and during both planned and unplanned school closures. The statistical properties of social interaction as characterized by social networks are crucial in determining patterns of epidemic spread. Knowing the structure of social contact networks enables us to test and assess the effect of different interventions that may change the dynamics of epidemics or stop it.

Vaccination has been shown to be the most effective way to mitigate the impact of influenza across all ages. Antivirals also have their place in treatment (Ferguson et al. 2006). However, flu vaccine is not always available in a timely manner, as was the case with the H1N1

pandemic. Vaccine is not always effective, as in the current (2014-15) season, where vaccine effectiveness is 22% overall and 26% in the 6 mo. − 18 yr. age group which includes school children. (Cdc.gov 2015a) Non-pharmaceutical interventions are the only things available when vaccine is absent or ineffective. These reduce the spread of disease in school children by reducing the risk of transmission or reducing the number of contacts. Hygiene programs (hand sanitizer/washing/cover coughs and sneezes) have been shown to reduce influenza A by 46%. (Stebbins et al. 2011) Keeping children home when sick with influenza is recommended to parents by the CDC (Cdc.gov 2011). Distancing efforts in school, such as reducing mixing during recess, lunch etc. reduces the number of interactions, and should be helpful. None of this has been specifically proven to reduce the incidence of influenza.

It has been suggested that school closure is an effective means to reduce the spread of influenza. Markel showed in the 1918–19 influenza pandemic that cities closing schools early and keeping them closed for a long time reduced the impact of the pandemic. Median closure time in the 43 cities Markel studied was 6 weeks. There was a statistically significant association between increased duration of nonpharmaceutical interventions and a reduced total mortality burden (Spearman r=-0.39, P=.005) (Markel, Stern, and Cetron 2008). A study in Israel in 2000 during a 12 day work stoppage found an impact on respiratory morbidity. (Heymann et al. 2004) Similarly Cauchemez found reductions in disease over school holidays, which are typically 10-14 days. (Cauchemez et al. 2008) Lee et al (2010) simulated a school closure and found that it could have an impact on a pandemic if maintained for 8 weeks. Jackson et al (2013) reviewed 79 studies and concluded "School closures appear to have the potential to reduce influenza transmission, but the heterogeneity in the data available means that the optimum strategy (e.g., the ideal length and timing of closure) remains unclear."

It seems that school closure was effective in 1918–19, however, the world has changed significantly since then. School children have vastly larger social networks. All of the studies cited above share the concern that school children recongregate significantly outside of school and during school holidays, both planned and unplanned. Children (89%) in a school closure in NC children recongregated in a large public setting even though instructed not to do so.

Many were observed by the school superintendent at the mall. (Johnson et al. 2008)

In 21st century America, school closure may be a problem. Short term school closure is manageable, but may not be of sufficient length to be effective. Longer term closure which might be effective would be disruptive, impacting economic and social costs. This would involve disruption of adults' lives, loss of income, an disruption in Kids' Lives, including education, meal programs, internet access, etc. (Cauchemez et al. 2009; Cdc.gov 2015c) Research suggests that some workers in the health care system would have to care for children and not be able to come to work when the system is stressed with influenza. Estimates range form 6-19% of workers (Lempel, Epstein, and Hammond 2009), 30% (Sadique, Adams, and Edmunds 2008) and 38% (Dalton, Durrheim, and Conroy 2008). CDC seems to be undecided about school closure at this time. Their Guidance for School Administrators to Help Reduce the Spread of Seasonal Influenza in K-12 Schools does not even mention school closure. (Cdc.gov 2015b)

This discussion on school closure involves student's social networking patterns. There are numerous ways to measure the social mixing of school children, including diaries, surveys and observation. (Wallinga, Teunis, and Kretzschmar 2006; Mossong et al. 2008). New technologies offer automatic collection of high-resolution interaction data over a short or long period of time (Lazer et al. 2009; Waber et al. 2010; Read, Eames, and Edmunds 2008). Recent advances in digital electronics have enabled the development of low-cost, low-power, multifunctional sensor nodes (also known as sensor motes or simply motes) that can measure proximity between devices over time (Akyildiz et al. 2002; Laibowitz, Pentland, and al. 2006). Having children wear these motes over a given time period can measure their interactions. Social network analysis, as a fundamental tool to study social structures, has shown great progress during the last decade parallel to explosion in data and advancements in the computational methods.

Studies have used diaries, survey and observation. One study (Wallinga, Teunis, and Kretzschmar 2006) demonstrated that explanation of observed infection incidence of mumps and influenza was improved if models accounted for the contact patterns. The POLYMOD study (Mossong et al. 2008) quantified mixing patterns for 8 European countries. This study

found strong assortative mixing of age groups, and particularly high rates of assortative mixing among school-aged children. Although the POLYMOD study found relatively few differences in mixing patterns across the countries, it is unknown if the information generated is appropriate for use for public health purposes within the USA, particularly in school-aged children, given differences in education systems and other establishments. These studies rely on human observations, which make them labor intensive, and provide a human a source of error with faulty memories and observations.

It would be more reliable to be able to track school children electronically, without any human influence. A study of school based contacts was carried out in a US high school (grades 9–12) consisting of 800 students, teachers, and staff using motes (Salathe et al. 2010). Their main finding is that the social network formed by connecting individuals who were in close contact (3 m), was a very dense network (about 750,000 close contacts) with low mean network distances between individuals and a relatively homogeneous connectivity distribution. Our study - the Social Mixing and Respiratory Transmission (SMART) study - conducted multi-trial studies with school children wearing motes during the school day and taking them home overnight and over school holidays (planned and unplanned). This report will show school day social networks and compare them to evening and non-school day networks.

Study Design and Methods

Participants

SMART was conducted in two school districts in Western Pennsylvania (Pittsburgh, PA Standard Metropolitan Statistical Area). Canon-McMillan is a public school district with 10 schools and 4700 students in grades K to 12. The district has an urban core, but is mostly suburban, with some areas classified as rural (Education 2016). Propel Charter School System has 8 schools in urban areas with 2700 K-12. SMART worked with 7 schools in these two districts on the facets of the research described herein.

The demographics of the study population differed from those of the Pittsburgh SMSA

(population >2.6 million; 89.8% White, 7.7% Black, 1.1% Asian, and 0.7% Hispanic). The subject population was less white (70.5%), more African American (25.8%), and less Asian (0.9%) than originally projected, reflecting a more urban population.

Table 1: Population Description

School	Grade	Motes	Male Students	Female Students	Classes
Elementary School A	0	42	19	23	2
Elementary School A	1	42	13	29	3
Elementary School A	2	39	21	18	3
Elementary School A	3	55	34	21	3
Elementary School A	4	40	20	20	3
Elementary School A	5	157	79	78	8
Elementary School B	0	16	8	8	1
Elementary School B	1	31	21	10	2
Elementary School B	2	43	22	21	2
Elementary School B	3	34	24	10	2
Elementary School B	4	35	18	17	2
High School A	9	75	37	38	8
High School A	10	74	32	42	8
High School A	11	52	27	25	9
High School A	12	11	7	4	6

Table 2: Mote Deployment Description

School	Grade	Contacts	Within Class Contacts	Within Grade Contacts
Elem. School A	0	209,520	188,985	190,656
Elem. School A	1	150,579	113,005	123,802
Elem. School A	2	113,498	72,908	87,390
Elem. School A	3	277,463	200,945	230,786
Elem. School A	4	114,660	74,678	93,712
Elem. School A	5	760,957	637,220	751,199
Elem. School B	0	24,029	12,458	12,458
Elem. School B	1	64,405	39,420	43,833
Elem. School B	2	151,025	90,295	101,353
Elem. School B	3	70,860	39,675	46,368
Elem. School B	4	105,174	52,459	59,867
High School A	9	371,803	91,470	272,385
High School A	10	493,024	147,265	367,372
High School A	11	239,052	35,428	160,036
High School A	12	32,377	5,164	12,854

Recruitment

School districts were recruited based on predetermined interest, identified by prior contact with the University of Pittsburgh School Based Research and Practice Network. Investigators met with district and school administrators and provided a detailed summary of the research, along with the opportunity to ask questions and discuss participation. School personnel who were authorized to approve participation did so using a formal letter.

Project staff met with school boards, parent-teacher organizations, school staff, and school nurses in order to introduce the project and distribute written explanatory materials. Students and parents were provided with a concise, readable summary of the study including

disclosure of potential risks and a signature section for opting out. After reviewing this material, students and parents could opt out of the study by signing and returning the form. A student could opt out of all surveillance at any time by refusing to participate, though no student refused to wear a mote. Parents and students had access to investigators via telephone or email to answer questions. This study was approved by the University of Pittsburgh IRB #PRO11020500 and CDC IRB# IRB00000319.

Across all schools, SMART had a 93.5% participation rate in year 1. The opt-out rates by school ranged from 1.0% to 18%. The 18% opt-out occurred in a school in which nearly one whole third grade class opted out. Participation was 88.9% in year 2; opt-out rate ranged from 0.4 to 23.5%.

Study Procedures

Schools provided a roster with the name, grade, and class of each participating student. An anonymous ID was assigned to each student in a database prior to any interaction with individual students. IDs were not assigned to individuals who opted out of the study. The link between this anonymous ID and the student's name were kept in secure study computers at the University of Pittsburgh. These IDs were used to link motes to specific children.

Motes

All consented students in all participating schools were given motes to wear. Motes are small electronic devices about the size of their two AA battery pack (1"x3"x3/4") worn around an individual's neck. Each mote is programmed to send and receive a signal from other motes when in proximity to one another. SMART used TelosB (made previously by Crossbow Technologies, now by Memsic Inc) wireless sensor motes (Polastre, Szewczyk, and Culler 2005). TelosB sensors utilize an IEEE 802.15.4 compliant RF transceiver, 2.4 GHz globally competitive ISM band, 8 MHz TI MSP430 microcontroller with 10 KB RAM, and 1 MB external flash memory for logging contacts. Motes are considered very low energy, operating on 2400 MHz to 2483.5 MHz frequency, similar to cordless phone and Wi-Fi. There is no

known exposure risk.

The sensor motes were programmed to continuously receive a signal from other nearby motes, and transmit a signal every 20 seconds, balancing the time resolution of data with the battery-life.. Whenever a mote receives a signal from another mote, it records the other mote's unique ID, the current time stamp and a radio signal strength indicator (RSSI). Signal strength provides a measure of proximity of the sensor motes and hence the individuals wearing the motes. Initial pilot investigation found that the signal strength between two motes dropped to about -80 dBm when they are face-to-face and about 3–4 meters away from each other. This distance was assumed to be meaningful to influenza transmission, and densities of signal strength were investigated by school and time window. Signal strength depends on many factors, including the line-of-sight and the presence of obstructions between the motes. Each recording is assumed to correspond to a continuous 20-second contact between the students. For example, x recordings between a pair of students throughout the day indicate x/3 minutes of aggregated contact time between them on that day. Mote programming uses Tiny OS 2.1.2, an open source system, running in a Linux environment.

During a school day, motes were worn by individuals, primarily students, but also teachers and staff who volunteered to wear them. The students wore motes inside in a waterproof pouch and attached to an adjustable lanyard. The motes were light and compact and so did not interfere with the student's activities. The entire unit, as worn, weighted 2.8 oz. Stationary sensor motes were also deployed in classrooms (1-2 motes per classroom) and common areas (1-5 motes each) to determine the spatial location of contacts. The locations of stationary motes were determined by studying floor plans to determine where common points occur. Finally, a master mote was created to send a signal to other motes to begin recording data. This saved battery life, and insured that all motes were synchronized.

Mote distribution started with compilation of class lists containing only participating students. Batteries were placed in motes as close to the time of deployment as possible to reserve battery life. Motes were first set at idle and did not record data until receiving a signal from a master mote. Bags of motes, along with a distribution list of consented students, were delivered to each classroom, typically during the first period. The motes

were distributed, and the mote number recorded next to the student number by the study staff. A SMART staff member stayed in the building all day to answer questions and resolve problems. Students wore the motes all day, except during vigorous activity where they might pose a hazard for the students. Students left school still wearing the motes, with the instruction:

> Keep the mote with you at all times. Wear the mote to all normal activities. Remove the mote only if playing rough sports, sleeping, getting wet or if it is otherwise not practical to wear the mote.

After two nights of wearing the motes, students returned the motes at school (typically in the last period). The motes were collected and the batteries removed to stop data collection. Motes were later plugged into a computer via usb port to download data.

In the three instances where two schools are shown, the K-4 elementary school is a feeder school for the intermediate school (5-6), so that the students would be riding the same school us system, and siblings may attend both schools.

Analysis

In this analysis, we compare contacts by three different time windows:

1. "Normal school hours" (8am – 3pm)
2. "Awake, non-school hours" (3pm-9pm, 6am-8am)
3. "Asleep, non-school hours" (9pm-6am)

Comparisons are made between in-school days and out-of-school days. We use the in-school days as a reference to out-of-school days to estimate the proportional reductions of contacts by time window.

Additionally, we constructed "student-pairs" from the data to determine factors associated with number of contacts. Characteristics such as gender-gender, age difference, and others are investigated.

In order to visually inspect how contacts varied over time, we calculated the average contacts per mote in a given hour. In this graph, the sets of vertical lines indicate the start of school (8 AM) and the end of school (3 PM). Note that the studies were not concurrent but, for comparison, are shown on the same time axis.
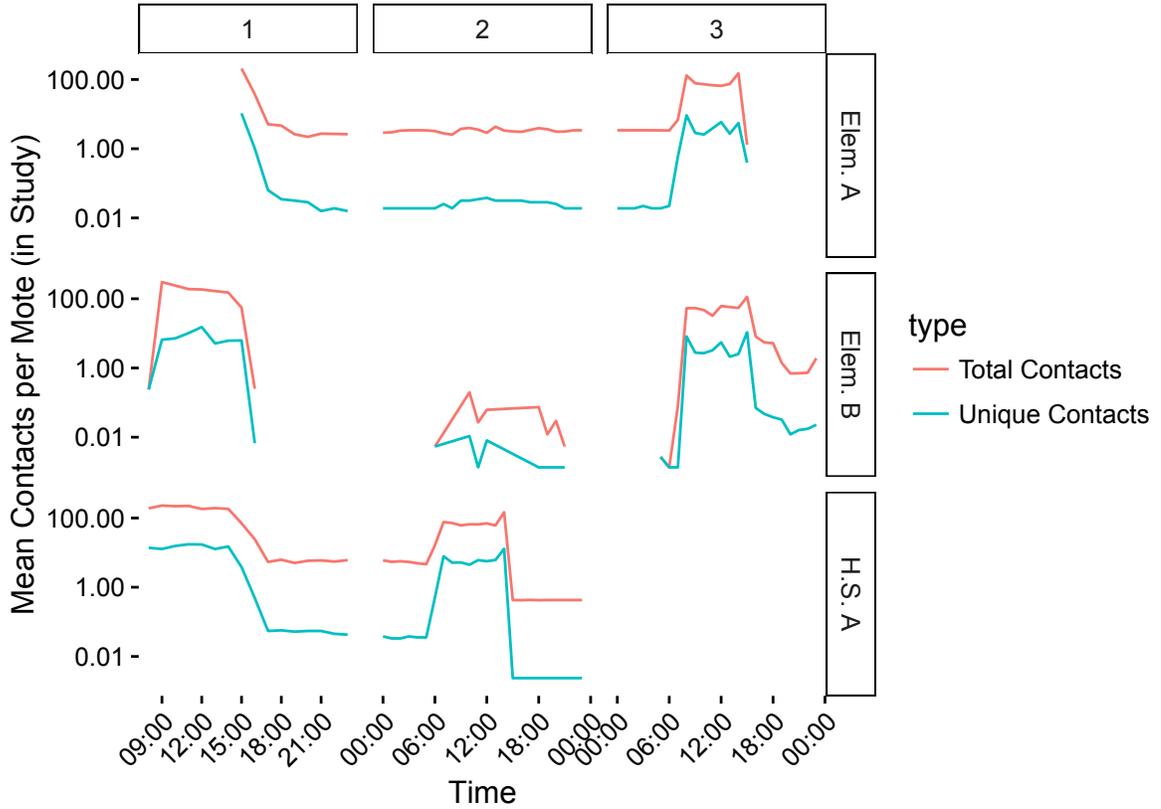

Figure 1: Average Contacts per Mote over Time

A bootstrap analysis to compare the proportion reduction in *mean hourly contacts* between the school closure day and normal school day was performed for all studies that had multiple consecutive days of contact data. Since some studies included data from two days of school, the school day with the most complete temporal data was chosen for comparison. Both days are broken out into the three categories mentioned above. Comparison is made within these categories between days. These estimates represent a crude measure of the effect of school closure on contact between children not adjusting for differences between hours besides designation into each of these categories.

We performed a regression analysis, regressing pairwise contacts between school children

on various individual characteristics. Information collected by study coordinators included *grade*, *class*, and *sex*. From the mote data, we calculated the number of *pairwise contacts* between a given pair of school children. The dependent variable in the regression was pairwise contacts in order to meet the independence assumption of linear models. Treating number of contacts by student as the response variable would introduce dependence between observations. Covariates for the regression were included in a pairwise fashion as well, such as *same grade* and *same class*. Linear regression was performed, regressing school contacts on the above covariates for each study. Additionally, logistic regression was performed, with the outcome of at least one pairwise contact and – separately – at least 15 pairwise contacts.

Results
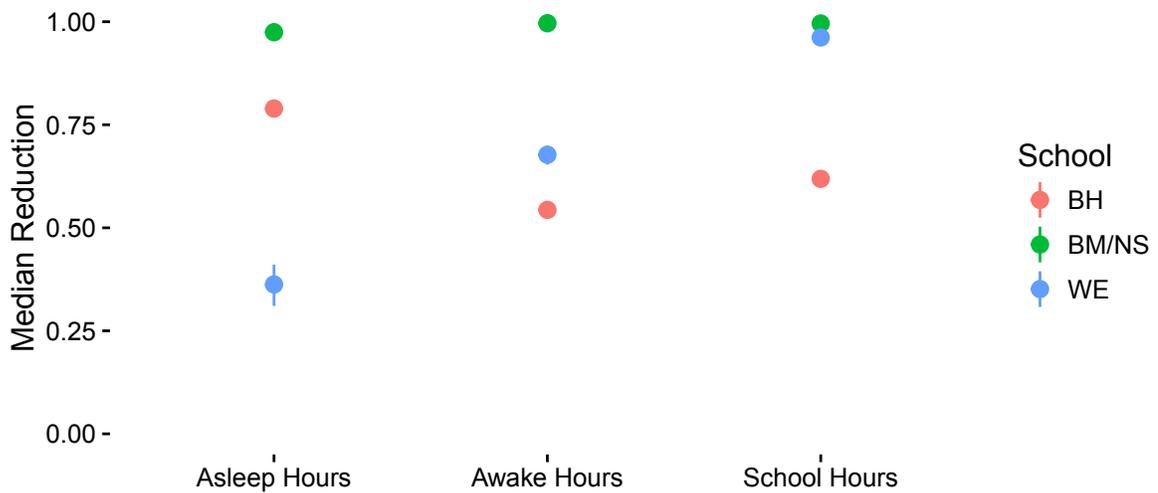
Figure 2: Percent Reduction of off-school day Mean hourly contacts to in-school day (2000 Bootstraps)

Signal Strength

We plotted density estimates of the by-mote signal strengths to compare contact patterns across studies and time windows. A lower number indicates closer proximity between motes, and a higher number indicates more distal proximity between motes. Indicated on each distribution is the median in blue and mean in red. It should be noted that signal strength is only a correlation of distance and is not precise.

From this density estimate, we can visualize how contact patterns vary by age group (using school level as a proxy) and time of day. While this density plot only provides marginal estimates, the shapes of the density – especially for out-of-school hours – provides insight into how children of different age groups interact.
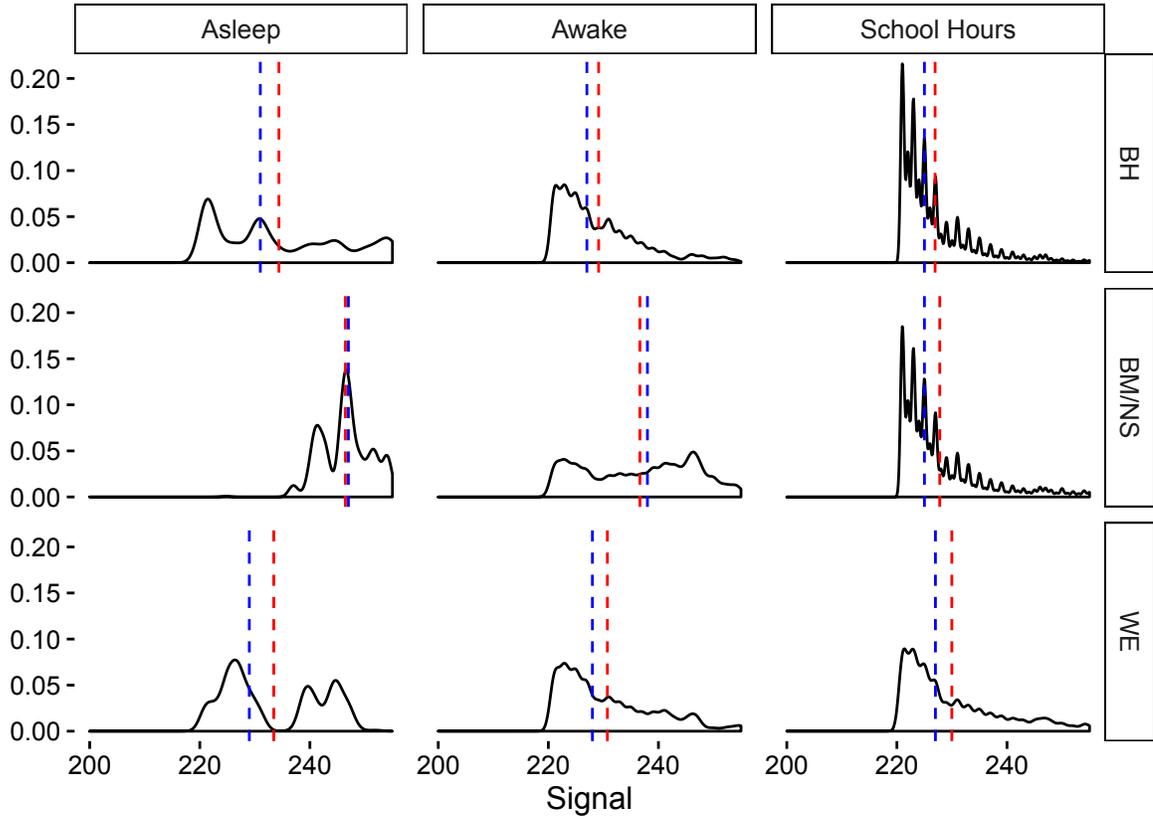

Figure 3: Signal Strength by Time Window and School

Pairwise Contacts

Here we show three different regression analyses. In the first model, we regressed the number of log-adjusted pairwise contacts on several pairwise characteristics. In the second model, we conducted a logistic regression of greater than or equal to one pairwise. In the final model, we used the same logistic regression model, but with an outcome of at least 15 pairwise contacts.

In the normal model, we included pairwise

- pairwise gender (relative to Male/Male)
- Grade Difference (abs(grade of student a – grade of student b))
- Pairwise grade/class indicator (relative to different grade/different class)

In the logistic models, we included all of these variables as well as an interaction between pairwise grade/class and pairwise gender.

For each of these three models, three regressions were conducted – one for elementary school A, one for elementary school B, and one for the high school. Below we show the estimates and confidence intervals for the coefficients estimates. Each model (normal, two logistic models) is represented as individual plots, with the coefficient estimates shown by point and the 95% confidence intervals represented as lines.
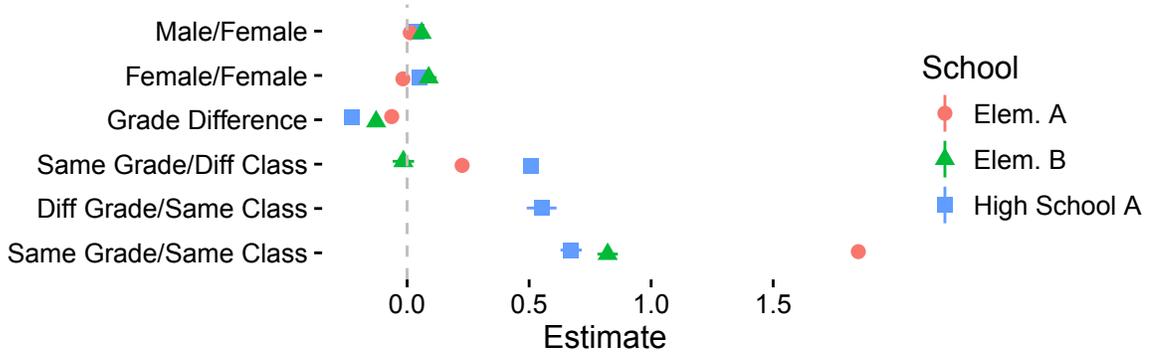
Figure 4: Linear Model Predicting $\log_{10} +1$ Pairwise School Contacts
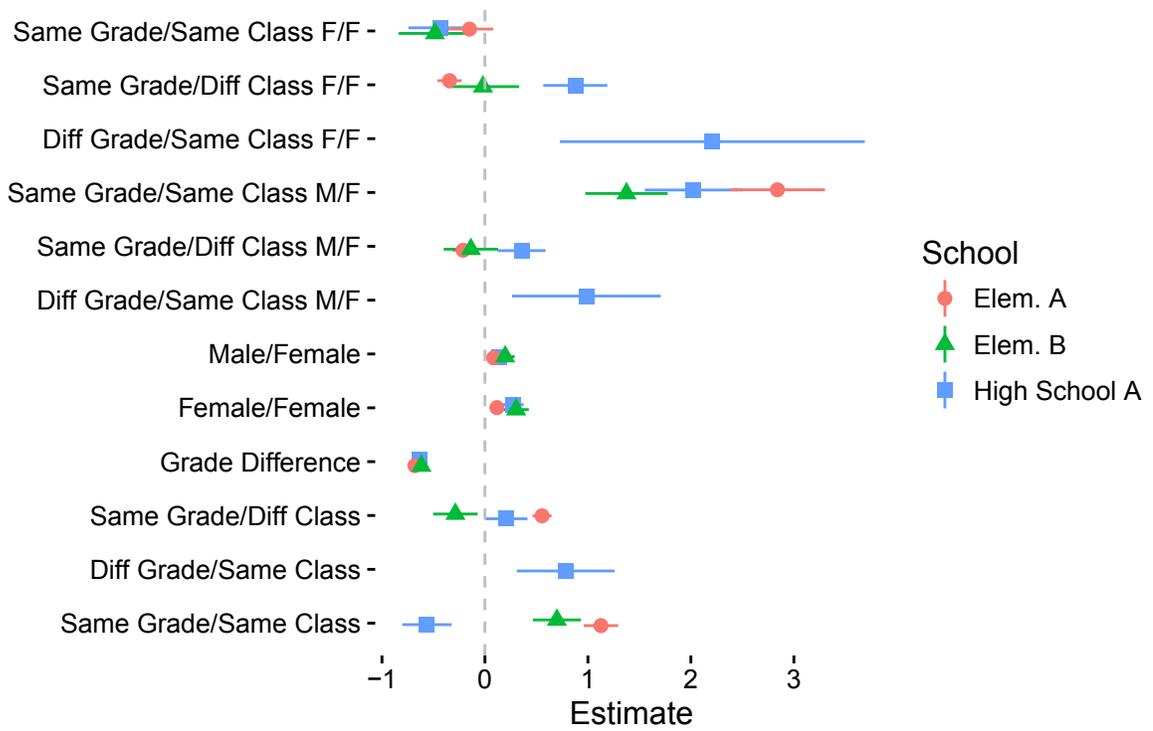
Figure 5: Logistic Model Predicting ≥ 1 Pairwise School Contacts
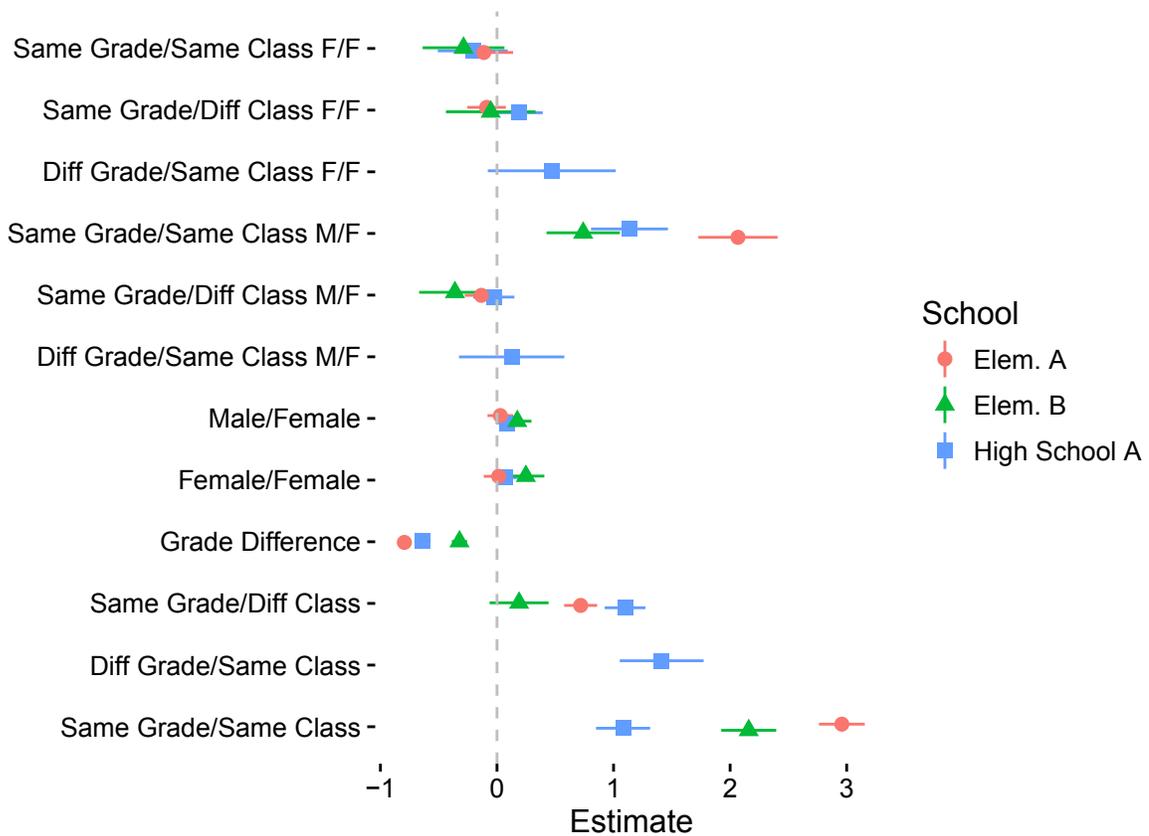
Figure 6: Logistic Model Predicting ≥ 15 Pairwise School Contacts

Discussion

By the three times of days, we recorded

1. 7,124.889 contacts per hour (Non-school Sleeping hour)

2. 49,500.25 contacts per hour (Non-school Waking hours)

3. 391,082.714 contacts per hour (School hours)

Based on our bootstrap analysis we estimate that closing a high school will reduce contacts between students by about 61.8% (CI 61.4-62.3%) during school hours. In elementary schools, we estimate that contacts between students will be reduced by nearly *100%* (99.6% in elementary school A, 96.2% in elementary school B)

Two of the three schools with multiple days of data were elementary schools. In contrast to the high school, the younger students showed a much greater reduction in contacts on off

days relative to school days. In particular, the reduction in contacts between high schoolers during school hours of off days was a nearly half of that of the younger students. This suggests that a school closure may prevent many contacts among younger students, but high schoolers will still have some contact with their peers.

Additionally, the distributions of signal strength show that interaction between students was more likely to be at a close proximity during school hours. This observation shows the importance of limiting contact during the school hours in the wake of an outbreak.

The regression analyses quantitatively verify some expected results. The logistic regression modeling the log odds of at least 15 contacts between students shows that being in the same class and grade is a very strong indicator of multiple contacts. The linear model again repeats this finding. In all three models, the high school students show a weaker difference of interacting with students in the same grade and same class relative to the elementary school students.

In both logistic models, increasing grade difference in pairs is a strong predictor of having a *lower* probability of contact between pairs. For every grade difference between a pair, there is a roughly 35% lower probability of having contacts or contacts. The linear model also shows a negative relationship between increasing grade difference and number of contacts, though to a lesser degree (about one less contact per grade difference).

Limitations

While the motes provide a way to quantify contacts between children, the motes themselves have limitations. Foremost among these limitations is the assumption that children wore the motes as instructed (i.e. they did not remove the motes during school, forget the motes at home, etc). Additionally, for the regression analysis, we have treated all contacts as being equal. Instead, as demonstrated in the density graph, there is variation between signal strengths, which is not accounted for in the regressions. Finally, the measure of distance between motes is only *correlation* of distance. While the density plots gives a good guess as to the distance between students by time window and school, it is not exact.

Genomic Analysis of Influenza Infected Students

Introduction

Previously, we considered only data collected via the motes. After this information was collected, students were followed throughout the influenza season with dates of symptom onset from December 13, 2012 to March 24, 2013. Students who were diagnosed with an influenza like illness (ILI) had the date of symptom onset recorded. In addition, a sample was collected and a deep sequence of the whole genome of the influenza virus was collected. This sequence contains the count of each nucelotide (A, T, C, G) at each nucleotide position.

This influenza data is combined with the previously collected epidemiological data, such as school, grade, and class. This data can be used to investigate transmission by proximity.

In the data collected, there were 188 cases of influenza across 9 schools. There are over 2,000 students across these schools who were eligible for infection.

Data were split up by strain - H3N2 and influenza B. Before proceding with estimation of transmission and epidemiological characteristics, we examine the number of cases by strain and school. While these aggregated cases are not adjusted by size of the school, it is of note that the four schools with the largest number of cases are predominantly influenza B infections.
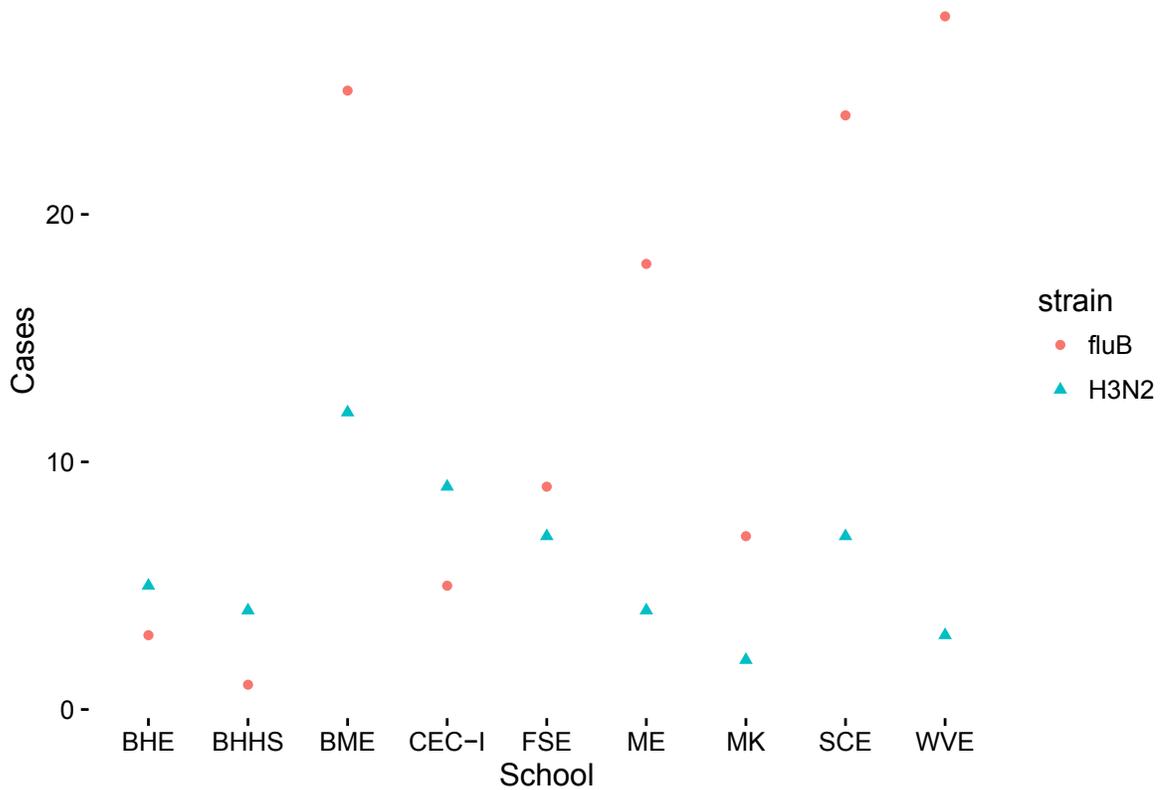
Figure 7: Influenza Cases by School and Strain

Methods

Reproductive Number

We investigate the transmission dynamics of these two strains of influenza in schools by calculating the reproductive number - the average number of secondary cases caused by an infection. The reproductive number is calculated for each school by each strain. We estimate the reproductive number by the Wallinga-Teunis method (Wallinga 2004). Implentation of this procedure was performed by the `EpiEstim` package in R (Cori et al. 2013).

Wallinga-Teunis estimation of the reproductive number requires an *a priori* specification of the serial interval - the time between transmissions. The serial interval is typically calculated from symptom onset of a primary case to symptom onset of a secondary case. Based on a prior study (Cowling et al. 2009), we parameterized the serial interval with a mean of 3.6 days and standard deviation of 1.6 days.

18

In this implementation of the Wallinga-Teunis method, the serial interval is described via gamma distribution. Instead of calculating an overall reproductive number, the Wallinga-Teunis method estimates reproductive numbers over time. The reproductive numbers are estimated over a moving window, which we specified as **seven days**. 95% confidence intervals for these time-varying reproductive numbers are determined by Monte Carlo simulation.

Probability of Tranmission Matrices

To examine transmission patterns between infected students, we constructed a matrix of probabilistic transmission. Using a similar parameterization of serial interval as the reproductive number estimation, we calculated pairwise probability of transmission. This probability was calculated using the dates of symptom onset. We calculated probability of transmission between a pair of infected students using a log-normal distribution with mean of **log(3.6)** and standard deviation of **log(1.6)**. Note that the matrix of pairwise transmission probabilities is **not** symmetric - if student A is infected before student B, there is a nonzero probability that student A infected student B, but a zero probability that infection occurred in the opposite direction.

After calculating these pairwise probabilities of transmission, matrices of pairwise probability of transmission by strain were restricted in three different ways

1) Transmissions can only occur within a class (zero probability of transmission occurring between classes, grades, or schools)
2) Transmissions can occur between and within classes but only within grade (zero probability of transmission occurring between grades or schools)
3) Transmissions can occur between grades but only within the same school (zero probability of transmission occurring between schools)

Phylogenetic Trees

To further explore models of transmission between infected children, we make use of the genomic sequences of isolated influenza viruses. Phylogenetic trees are a commonly used

tool in molecular epidemiology to create possible transmission trees (Holmes et al. 1995). Based on the created phylogenetic trees, we can assess which infected students are more closely related on a transmission chain.

Phylogenetic trees are fit using the R package `phangorn` (Schliep 2011). Multiple substitution models were used including Jukes-Cantor (Jukes and Cantor 1969) and Generalized Time Reversible (Tavaré 1986). Model selection was performed using Akaike Information Criterion (AIC). For both strains of influenza, the Generalized Time Reversible substitution model was selected.

Results

Wallinga-Teunis
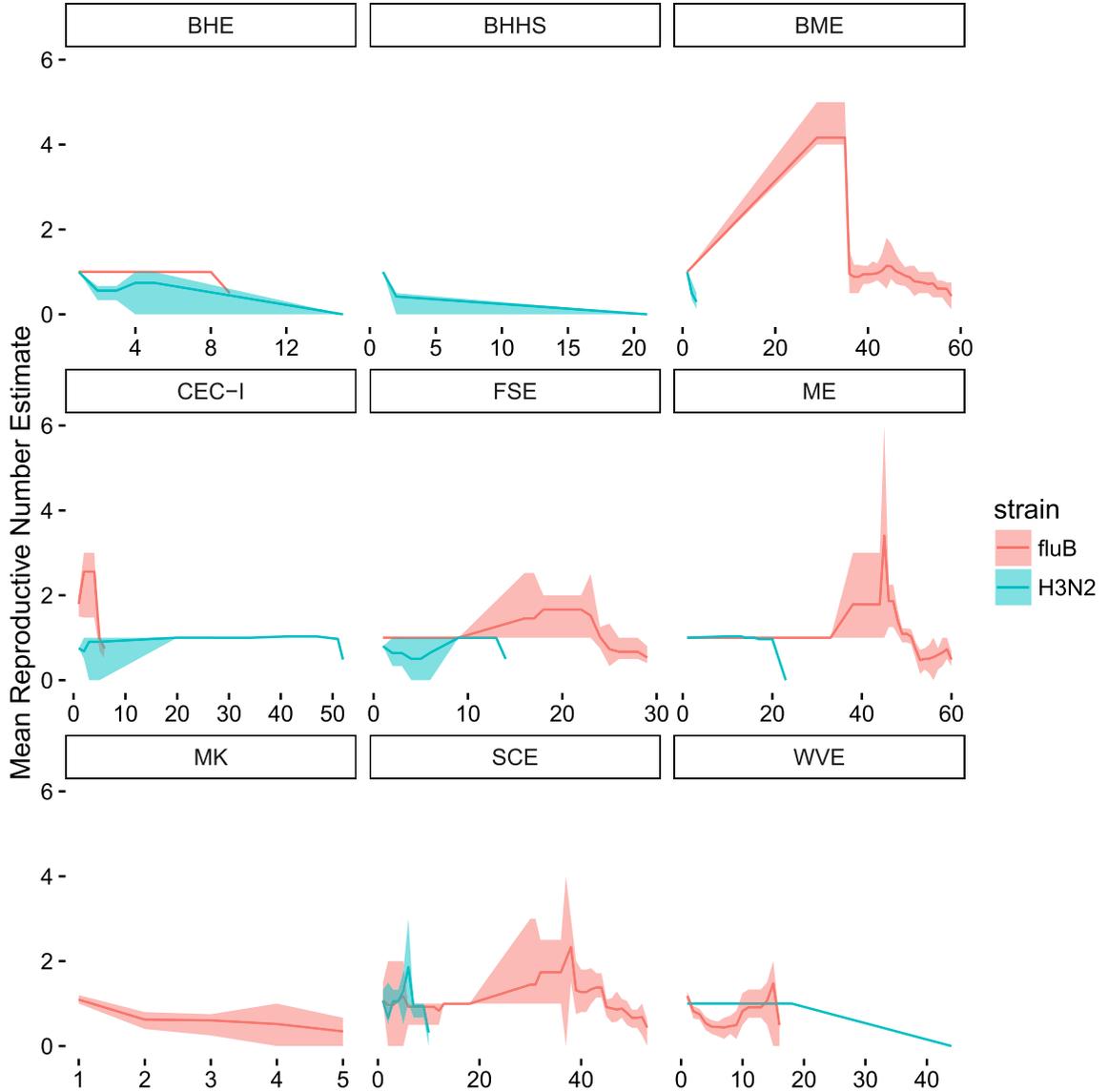
Figure 8: Case Reproduction Number of Flu

Note here that x axis in Figure 2 corresponds to the first day of the window. Additionally, the plot of the reproductive number estimate is left aligned by the day of the introduction of a given strain in a school.

While there is a lot of variabilitity in the estimates over time of the reproductive number, the influenza B strain generally has a higher reproductive number.

Probability of Transmission Matrices

Below, we depict the three transmission models per strain as heatmaps. The plots depict the probability of transmission from a student on the x axis to a student on the y axis. The probabilities are represented as colors ranging from "white" (no probability of transmission) and "dark blue" (high probability of transmission). Students are ordered by school, grade, and class. Thus, for a given student, the student is adjacent to students in the same class, nearest to students in the same grade, and within a connected set of students in the same school.

For the three models proposed, it can be seen how much of transmission is driven by classes, grades, and schools. Based on the degree of increasing "heat" in the plots, this depicts the relative amount of transmission driven by class, grade, or school according to this model.
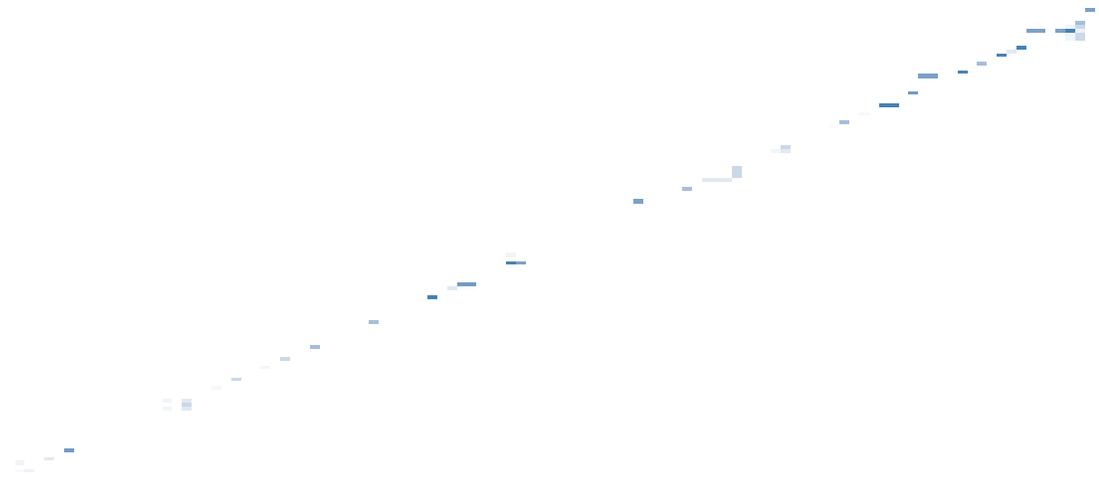
Figure 9: Flu B Transmission by Class
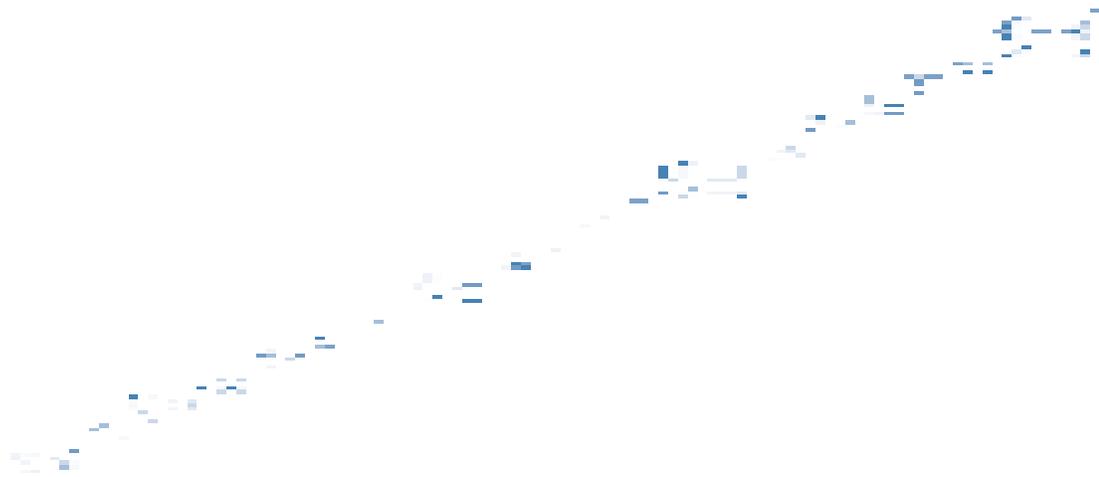
Figure 10: Flu B Transmission by Grade

Figure 11: Flu B Transmission by School

This model suggests that very little transmission is driven within classes, and instead most transmission occurs between classes and grades, as seen by the drastically increasing "heat." The four schools with more infections have a larger portion infections driven between grades than the smaller schools. However, since there are so few cases, this potential model is especially sensitive to the choice of serial interval.
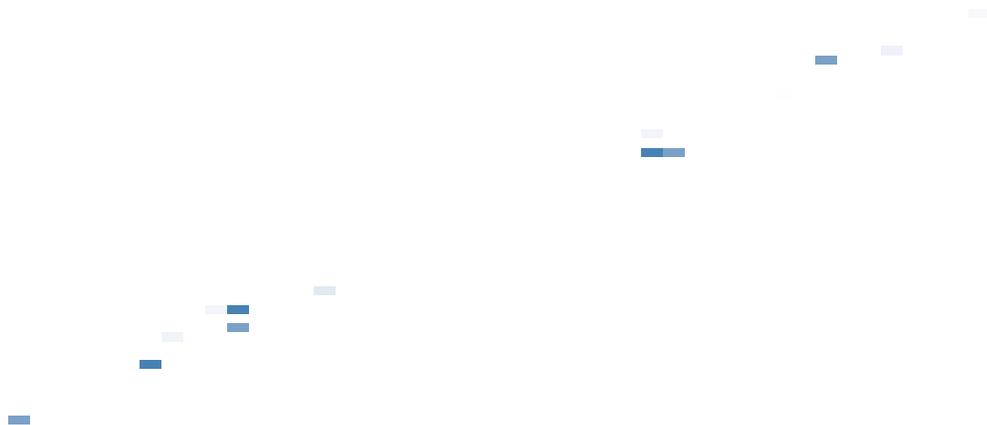
Figure 12: H3N2 Transmission by Class
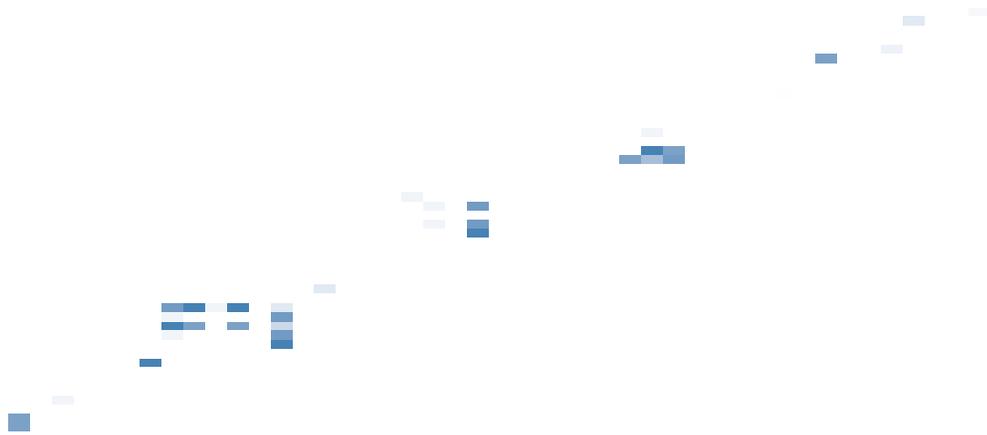
Figure 13: H3N2 Transmission by Grade

Figure 14: H3N2 Transmission by School

In this model of H3N2 transmission, there is very little within class transmission. Instead, transmission is driven largely by within grade/between class contact. Unlike the transmission models demonstrated above for influenza B, little transmission appears to occur within school/between grades.

Phylogenetic Trees

Here we depict the phylogenetic trees as fit by `phangorn` according to the Jukes-Cantor and Generalized Time Reversible substitution models. In these rooted trees, horizontal lines, but not vertical lines, separating students indicates the relative amount of genetic distance between the isolated influenza viruses.

To assess relationship of students, identifications were constructed. These identifications were constructed as follows

- "s" followed by a number indicates the school number that the child was a member of

- "g" followed by "K" or a number indicates the grade that the child was in

- "r" followed by a number indicates the room number (within the school and grade) that the child belonged to

- "id" followed by a number is a unique id constructed within the school for each student. This number has no meaning other than to separate two students in the same school, grade, and room

Students that are grouped together in the phylogeny (little to no genetic distance between the viruses) are referred to as clades.

To create the trees for the H3N2 virus, there were 49 students. For the influenza B phylogenetic tree, there were 131 students.
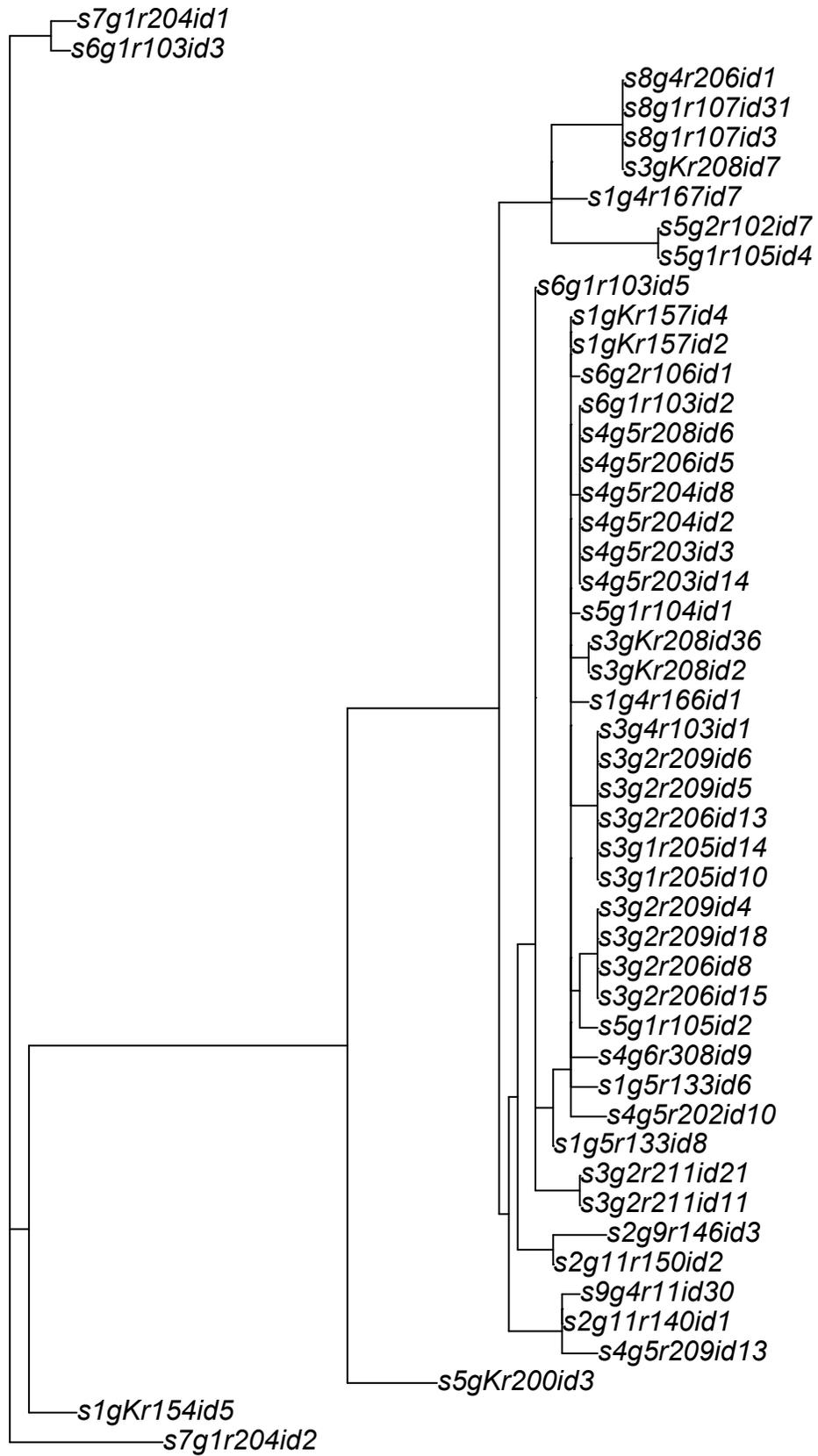
Figure 15: Jukes Cantor H3N2 Phylogenetic Tree
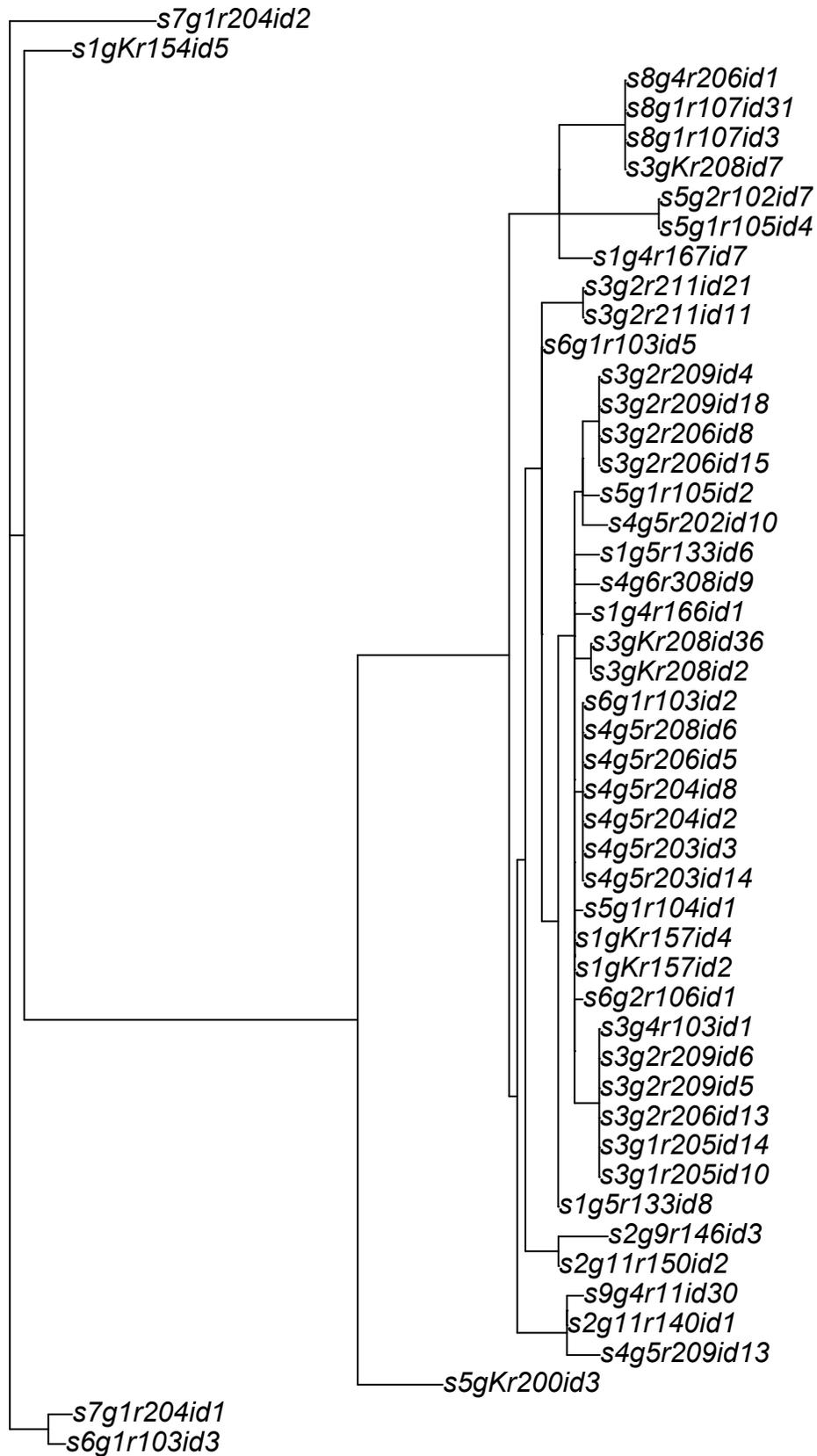
Figure 16: Generalized Time Reversible H3N2 Phylogenetic Tree

For both substition models depicted, children in the same schools are nearly perfectly discriminated into clades. The exceptions - such as "s5gKr200id3" - possibly indicate an exogenous (out of school) introduction of influenza.

Both models support the claim made earlier that transmission has only occurred somewhat within class, and instead is mostly explained by transmission between classes and grades.

Figure 17: Jukes Cantor Flu B Phylogenetic Tree

Figure 18: Generalized Time Reversible Flu B Phylogenetic Tree

In the phylogenies shown above for influenza B, we again see that most infections are relatively closely related. There is only one infection that appears to be an exogenous introduction - from school 5 - in both models. Here the transmission seems to have occurred to a less extent within classes and grades, and instead the influenza B infections are transmitted within the school as a whole.

Discussion

One of the strengths of this study is the collection of genomic sequences of viruses infecting students with known contact patterns. In addition to the symptom onset information, we are able to use the genomic data to construct likely transmission trees. These analyses have shown that proximity drives transmission. Regardless of the model used, the data suggest that close proximity to other infected students is a risk factor for respiratory infection.

References

Akyildiz, I. F., W. Su, Y. Sankarasubramaniam, and E. Cayirci. 2002. "Wireless Sensor Networks: A Survey." *Computer Networks* 38: 393–422.

Cauchemez, S., A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. M. Ferguson, D. Swerdlow, S. V. Sodha, M. E. Moll, and F. J. et al. Angulo. 2011. "Role of Social Networks in Shaping Disease Transmission During a Community Outbreak of 2009 H1N1 Pandemic Influenza." *Proceedings of the National Academy of Sciences* 108 (7): 2825–30. doi:10.1073/pnas.1008895108.

Cauchemez, Simon, Neil M Ferguson, Claude Wachtel, Anders Tegnell, Guillaume Saour, Ben Duncan, and Angus Nicoll. 2009. "Closure of Schools During an Influenza Pandemic." *The Lancet Infectious Diseases* 9 (8): 473–81. doi:10.1016/s1473-3099(09)70176-8.

Cauchemez, Simon, Alain-Jacques Valleron, Pierre-Yves Boëlle, Antoine Flahault, and Neil M. Ferguson. 2008. "Estimating the Impact of School Closure on Influenza Transmission

from Sentinel Data." *Nature* 452 (7188): 750–54. doi:10.1038/nature06732.

Cdc.gov. 2011. *The Flu: A Guide for Parents.* 1st ed. http://www.cdc.gov/flu/pdf/freeresources/updated/a_flu_guide_for_parents.pdf.

———. 2015a. "Early Estimates of Seasonal Influenza Vaccine Effectiveness - United States, January 2015." http://www.cdc.gov/mmwr/preview/mmwrhtml/mm6401a4.htm.

———. 2015b. "Guidance for School Administrators to Help Reduce the Spread of Seasonal Influenza in K-12 Schools| Seasonal Influenza (Flu) | CDC." http://www.cdc.gov/flu/school/guidance.htm.

———. 2015c. "Impact of Seasonal Influenza-Related School Closures on Families — Southeastern Kentucky, February 2008." http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5850a2.htm.

Chao, Dennis L., Elizabeth M. Halloran, and Ira M. Longini Jr. 2010. "School Opening Dates Predict Pandemic Influenza a(H1N1) Outbreaks in the United States." *The Journal of Infectious Diseases* 202 (6): 877–80. doi:10.1086/655810.

Cori, A., N. M. Ferguson, C. Fraser, and S. Cauchemez. 2013. "A New Framework and Software to Estimate Time-Varying Reproduction Numbers During Epidemics." *American Journal of Epidemiology* 178 (9): 1505–12. doi:10.1093/aje/kwt133.

Cowling, Benjamin J., Vicky J. Fang, Steven Riley, J S. Malik Peiris, and Gabriel M. Leung. 2009. "Estimation of the Serial Interval of Influenza." *Epidemiology* 20 (3): 344–47. doi:10.1097/ede.0b013e31819d1092.

Dalton, CB, DN Durrheim, and MA Conroy. 2008. "Likely Impact of School and Childcare Closures on Public Health Workforce During an Influenza Pandemic: A Survey." *Commun Dis Intell Q Rep* 32 (2): 261–2.

Earn, David J.D. 2012. "Effects of School Closure on Incidence of Pandemic Influenza in Alberta, Canada." *Annals of Internal Medicine* 156 (3): 173. doi:10.7326/0003-4819-156-3-201202070-00005.

Education, Pennsylvania Department of. 2016. "'URBAN/RURAL' CLASSIFICATION

oF SCHOOLS aND LOCAL EDUCATION AGENCIES (LEAS)." *Education.pa.gov.* http://www.education.pa.gov/Data-and-Statistics/Pages/School-Locale.aspx.

Esposito, Susanna, Luigi Cantarutti, Claudio Giuseppe Molteni, Cristina Daleno, Alessia Scala, Claudia Tagliabue, Claudio Pelucchi, Carlo Giaquinto, and Nicola Principi. 2011. "Clinical Manifestations and Socio-Economic Impact of Influenza Among Healthy Children in the Community." *Journal of Infection* 62 (5): 379–87. doi:10.1016/j.jinf.2011.02.015.

Ferguson, Neil M., Derek A. T. Cummings, Christophe Fraser, James C. Ca, Philip C. Cooley, and Donald S. Burke. 2006. "Strategies for Mitigating an Influenza Pandemic." *Nature* 442 (7101): 448–52. doi:10.1038/nature04795.

Glezen, W. P. 1996. "Emerging Infections: Pandemic Influenza." *Epidemiologic Reviews* 18 (1): 64–76. doi:10.1093/oxfordjournals.epirev.a017917.

Heymann, Anthony, Gabriel Chodick, Brian Reichman, Ehud Kokia, and Joseph Laufer. 2004. "INFLUENCE oF SCHOOL CLOSURE oN tHE INCIDENCE oF VIRAL RESPIRATORY DISEASES AMONG CHILDREN aND oN HEALTH CARE UTILIZATION." *The Pediatric Infectious Disease Journal* 23 (7): 675–77. doi:10.1097/01.inf.0000128778.54105.06.

Holmes, Eddie C., Sean Nee, Andrew Rambaut, Geoff P. Garnett, and Paul H. Harvey. 1995. "Revealing the History of Infectious Disease Epidemics through Phylogenetic Trees." *Philosophical Transactions: Biological Sciences* 349 (1327). The Royal Society. doi:10.2307/56121.

Jackson, C., E. Vynnycky, J. Hawker, B. Olowokure, and P. Mangtani. 2013. "School Closures and Influenza: Systematic Review of Epidemiological Studies." *BMJ Open* 3 (2): e002149–e002149. doi:10.1136/bmjopen-2012-002149.

Johnson, April J., Zack S. Moore, Paul J. Edelson, Lynda Kinnane, Megan Davies, David K. Shay, Amanda Balish, Meg McCarron, Lenee Blanton, and Lyn et al. Finelli. 2008. "Household Responses to School Closure Resulting from Outbreak of Influenza B, North Carolina." *Emerg. Infect. Dis.* 14 (7). doi:10.3201/eid1407.080096.

Jukes, T. H., and C. R. Cantor. 1969. *Evolution of Protein Molecules.* Edited by H. N. Munro. Academy Press.

Laibowitz, Mathew, Alex (Sandy) Pentland, and et al. 2006. "A Sensor Network for Social

Dynamics." In *IN PROCEEDINGS oF tHE 5TH INTERNATIONAL CONFERENCE oN INFORMATION PROCESSING iN SENSOR NETWORKS*, 483–91.

Lazer, D., A. Pentland, L. Adamic, S. Aral, A.-L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, and M. et al. Gutmann. 2009. "SOCIAL SCIENCE: Computational Social Science." *Science* 323 (5915): 721–23. doi:10.1126/science.1167742.

Lee, Bruce Y., Shawn T. Brown, Philip Cooley, Maggie A. Potter, William D. Wheaton, Ronald E. Voorhees, Samuel Stebbins, John J. Grefenstette, Shanta M. Zimmer, and Richard K. et al. Zimmerman. 2010. "Simulating School Closure Strategies to Mitigate an Influenza Epidemic." *Journal of Public Health Management and Practice* 16 (3): 252–61. doi:10.1097/phh.0b013e3181ce594e.

Lempel, Howard, Joshua M Epstein, and Ross A Hammond. 2009. "Economic Cost and Health Care Workforce Effects of School Closures in the U.S." *PLoS Curr* 1: RRN1051. doi:10.1371/currents.rrn1051.

Markel, H., A. Stern, and M. Cetron. 2008. "Nonpharmaceutical Interventions Implemented by US Cities During the 1918-1919 Influenza Pandemic." *International Journal of Infectious Diseases* 12: e432. doi:10.1016/j.ijid.2008.05.1263.

Mossong, Joël, Niel Hens, Mark Jit, Philippe Beutels, Kari Auranen, Rafael Mikolajczyk, Marco Massari, Stefania Salmaso, Gianpaolo Scalia Tomba, and Jacco et al. Wallinga. 2008. "Social Contacts and Mixing Patterns Relevant to the Spread of Infectious Diseases." *Plos Med* 5 (3): e74. doi:10.1371/journal.pmed.0050074.

Read, J. M, K. T.D Eames, and W. J. Edmunds. 2008. "Dynamic Social Networks and the Implications for the Spread of Infectious Disease." *Journal of The Royal Society Interface* 5 (26): 1001–7. doi:10.1098/rsif.2008.0013.

READ, J. M., W. J. EDMUNDS, S. RILEY, J. LESSLER, and D. A. T. CUMMINGS. 2012. "Close Encounters of the Infectious Kind: Methods to Measure Social Mixing Behaviour." *Epidemiol. Infect.* 140 (12): 2117–30. doi:10.1017/s0950268812000842.

Sadique, Md Z, Elisabeth J Adams, and William J Edmunds. 2008. "Estimating the Costs of School Closure for Mitigating an Influenza Pandemic." *BMC Public Health* 8 (1): 135.

doi:10.1186/1471-2458-8-135.

Salathe, M., M. Kazandjieva, J. W. Lee, P. Levis, M. W. Feldman, and J. H. Jones. 2010. "A High-Resolution Human Contact Network for Infectious Disease Transmission." *Proceedings of the National Academy of Sciences* 107 (51): 22020–5. doi:10.1073/pnas.1009094108.

Schliep, K.P. 2011. "Phangorn: Phylogenetic Analysis in R." *Bioinformatics* 27 (4): 592–93.

Stebbins, Samuel, Derek A. T. Cummings, James H. Stark, Chuck Vukotich, Kiren Mitruka, William Thompson, Charles Rinaldo, Loren Roth, Michael Wagner, and Stephen R. et al. Wisniewski. 2011. "Reduction in the Incidence of Influenza a but Not Influenza B Associated with Use of Hand Sanitizer and Cough Hygiene in Schools." *The Pediatric Infectious Disease Journal* 30 (11): 921–26. doi:10.1097/inf.0b013e3182218656.

Tavaré, S. 1986. "Some Probabilistic and Statistical Problems in the Analysis of DNA Sequences." In *American Mathematical Society: Lectures on Mathematics in the Life Sciences*, 17:57–86. Paperback; Amer Mathematical Society. http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20/&path=ASIN/0821811673.

Waber, Benjamin N., Ellen Pollock, Manuel Cebrián, Riley Crane, Leon Danon, and Alex Pentland. 2010. "Modeling Corporate Epidemiology." *CoRR* abs/1008.3402. http://arxiv.org/abs/1008.3402.

Wallinga, J. 2004. "Different Epidemic Curves for Severe Acute Respiratory Syndrome Reveal Similar Impacts of Control Measures." *American Journal of Epidemiology* 160 (6): 509–16. doi:10.1093/aje/kwh255.

Wallinga, J., P. Teunis, and M. Kretzschmar. 2006. "Using Data on Social Contacts to Estimate Age-Specific Transmission Parameters for Respiratory-Spread Infectious Agents." *American Journal of Epidemiology* 164 (10): 936–44. doi:10.1093/aje/kwj317.

WHO. 2016. *Limiting Spread.* 1st ed. http://www.who.int/influenza/resources/research/research_agenda_influenza_stream_2_limiting_spread.pdf.

Zhou, H., W. W. Thompson, C. G. Viboud, C. M. Ringholz, P.-Y. Cheng, C. Steiner, G. R. Abedi, L. J. Anderson, L. Brammer, and D. K. Shay. 2012. "Hospitalizations Associated

with Influenza and Respiratory Syncytial Virus in the United States, 1993-2008." *Clinical Infectious Diseases* 54 (10): 1427–36. doi:10.1093/cid/cis211.

2016. *Cdc.gov*. http://www.cdc.gov/flu/school/guidance.htm.

CV

jacob**carey**

statistician

about

5633 Columbia Rd
Apt 304
Columbia, MD 21044

☎ +1 (434) 203 9317
✉ jacobcvt12@gmail.com
⌂ jacobcvt12
in jacobcvt12

programming

R, JAGS, Stan,
C, C++,
Python, SQL,
SAS, Stata, VBA

tools

git, Vim, Bash
LaTeX

environments

Debian/Ubuntu
OS X

skills

Statistical computing
Bayesian statistics
Multilevel modeling
Machine learning
Regression analysis
Time series analysis
Data visualization
Web scraping

education

2014–Now   **Master** of Science in **Epidemiology**   Johns Hopkins University
           May 2016 Expected Graduation

2008–2012 **Bachelor** of Science in **Mathematics**   Virginia Tech
           Graduated with Honor

experience

2015–Present **Johns Hopkins University Division of Biostatistics and Bioinformatics**   Baltimore, Maryland
Research Assistant
– Develop software to perform statistical analysis of genomic data
– Use developed software to make inference on copy number status and disease association in large study

2014–Present **Johns Hopkins University Department of Epidemiology**   Baltimore, Maryland
Research Assistant
– Analysis and writeup of epidemiological studies

2014   **Marriott International**   Bethesda, Maryland
Manager, eCommerce Analysis
– Assisted with new Marriott user-centric sort order algorithm
– Developed time series model to determine effectiveness of hotel promotions.
– Calculated weekly values of search engine keywords in order to best allocate online marketing funds.
– Performed web scraping of external websites to provide business insights.
– Worked with datasets greater than one billion rows.
– Automated reports using R & ggplot2 and ETL using Python scripts.

2012–2014 **Deloitte**   Arlington, Virginia
Data Analyst
– Managed a client-facing team of three consultants in the analysis of datasets greater than one million rows. Led regular meetings and presented results and data visualizations to key clients and internal stakeholders.
– Developed web application to be used by federal agencies to optimize resource allocation. Used Python to perform server-side calculations.
– Assisted federal agency in meeting requirements of government act using sample selection and estimation of disbursed grants and transactions.

software

CNPBayes   https://github.com/scristia/CNPBayes
Role: Author
– Bayesian mixture model for estimating copy number polymorphisms