COMPUTATIONAL EPIDEMIOLOGY - ANALYZING EXPOSURE RISK: A DETERMINISTIC, AGENT-BASED APPROACH

Martin Joseph O'Neill II

Thesis Prepared for the Degree of

MASTER OF SCIENCE

UNIVERSITY OF NORTH TEXAS

August 2009

APPROVED:

Armin R. Mikler, Major Professor
Yan Huang, Committee Member
Bruce Hunter, Committee Member
Joseph R. Oppong, Committee Member
Krishna Kavi, Chair of the Department of Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of Engineering
Michael Monticino, Dean of the Robert B. Toulouse School of Graduate Studies

O'Neill II, Martin Joseph. Computational Epidemiology - Analyzing Exposure Risk: A Deterministic, Agent-Based Approach. Master of Science (Computer Science), August 2009, 56 pp., 3 tables, 32 illustrations, references, 18 titles.

Many infectious diseases are spread through interactions between susceptible and infectious individuals. Keeping track of where each exposure to the disease took place, when it took place, and which individuals were involved in the exposure can give public health officials important information that they may use to formulate their interventions. Further, knowing which individuals in the population are at the highest risk of becoming infected with the disease may prove to be a useful tool for public health officials trying to curtail the spread of the disease. Epidemiological models are needed to allow epidemiologists to study the population dynamics of the transmission of infectious agents and the potential impact of infectious disease control programs. While many agent-based computational epidemiological models exist in the literature, they focus on the spread of disease rather than exposure risk. These models are designed to simulate very large populations, representing individuals as agents, and using random experiments and probabilities in an attempt to more realistically guide the course of the modeled disease outbreak. The work presented in this thesis focuses on tracking exposure risk to chickenpox in an elementary school setting. This setting is chosen due to the high level of detailed information realistically available to school administrators regarding individuals' schedules and movements. Using an agent-based approach, contacts between individuals are tracked and analyzed with respect to both individuals and locations. The results are then analyzed using a combination of tools from computer science and geographic information science.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1. Introduction

Many infectious diseases are spread through interactions between susceptible individuals and infectious individuals. Knowing which individuals in the population are at the highest risk of becoming infected with a disease may prove to be a useful tool for public health officials trying to curtail the spread of a disease. Keeping track of where each exposure to the disease took place, when it took place, and which individuals were involved in the exposure can give public health officials important information that they may use to formulate interventions. While defining what is necessary to have a contact that could result in exposure to the disease may seem straightforward, it involves assumptions that must be changed depending on the environment, the population, and what disease is being modeled. Additionally, determining an individual's movements in the population can be extremely daunting, but being able to determine the set of contacts between individuals in the population is dependent on being able to determine each individual's movements in the population.

An example of a population and environment that is well-suited for this approach is children in an elementary school. At the elementary school level, course schedules dictate the majority of movements of individuals. These schedules are well-documented by school administrators. For exceptions to the course schedules, a log of hall passes may even be kept by teachers within the school. Combining the course schedules with the logs of hall passes paints a fairly accurate picture of the movements of individuals in the population.

What constitutes a contact and may lead to disease transmission must then be determined. Next, assumptions must be formulated concerning what is necessary to have a contact between two individuals that (if one was susceptible and the other infectious) would lead to an exposure. These assumptions depend on the disease in question. For example, human immunodeficiency virus (HIV) can be transmitted from individual to individual through sexual contact or sharing needles, but not through the air. Measles can be transmitted through the air or through contact with fluid secreted from an infected individual's nose or throat. Mononucleosis is transmitted through infected saliva or blood, and contacts such as kissing or sharing a drink may lead to transmission. Therefore, the mode of transmission of the disease being modeled must be used when determining what is necessary to have a contact between two individuals that could lead to an exposure to the disease.

The disease used as an example in this research is chickenpox. Chickenpox can be spread through the air or through direct contact. Hence, for a susceptible individual to be exposed to the disease by an infectious individual, they only need occupy the same room at the same time. In an elementary school, most rooms are sufficiently small that individuals in each room can be considered uniformly mixed due to the airborne nature of the disease. While a chickenpox vaccine is available, chickenpox remains a common childhood illness and cause for concern.

1.2. Chickenpox

Chickenpox (also known as varicella) is known as a childhood disease (6). It is characterized by slight fever, mild constitutional symptoms, and a rash that develops vesicles (fluid filled sacs (17)) on the skin for 3-4 days (7). Outbreaks of chickenpox are not uncommon in schools (7). Chickenpox is airborne and enters susceptible individuals through the respiratory tract and conjunctiva (5). At least 90% of the population in temperate climates will have had chickenpox by age 15 (7). This number rises to 95% when considering young adults in the same population (7). Chickenpox is highly contagious (2), and after household exposure, susceptible individuals' risk of infection is about 80%-90% (7)(5).

While chickenpox is known as a mild childhood disease, it can be contracted by and be much more severe in adults than children (7)(6). Fever may be much higher and constitutional symptoms may be much worse (7). Additionally, the case-fatality rate for adults is

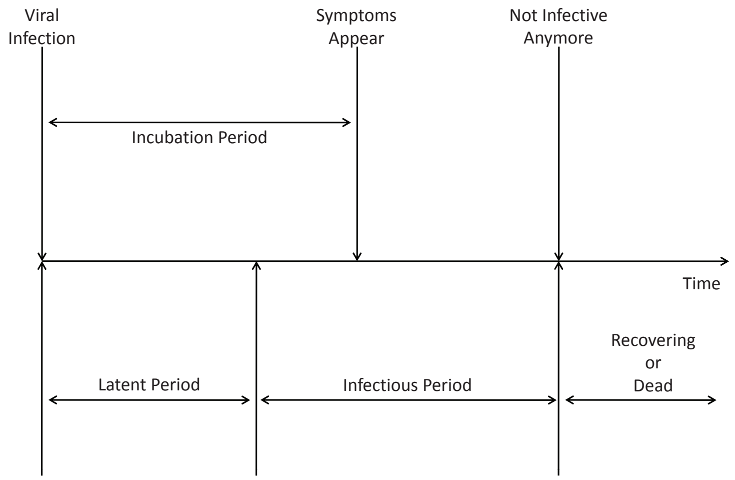


FIGURE 1.1. Example Infection Timeline

much higher with 1 in 5,000 adults with chickenpox dying from chickenpox, and only 1 in 100,000 children with chickenpox dying from chickenpox (7). Therefore, while it may be a very common childhood disease, susceptible adults should take precautions to not become infected.

Figure 1.1 depicts the periods an infectious individual must pass through during a disease such as chickenpox. The *latent period* is the period of time between initial viral infection and the beginning of the infectious period. The *infectious period* is the period of time in which an infected individual is infectious and may spread the disease to others. The *incubation period* is the period of time between initial viral infection and the appearance of symptoms. It generally overlaps both the latent period and part of the infectious period.

An understanding of the infection timeline shown in Figure 1.1 is needed to understand the difficulty of preventing the spread of chickenpox within a population. The latent period of chickenpox varies in the literature from 9-16 days (or more commonly 12-15 days) after initial infection (13)(7) to 6-15 days (or 8-10 days on average) after initial infection (15). However, the incubation period lasts usually 1-2 days (and as many as 5 days) longer than the latent period (5). Hence, infectious individuals may spread chickenpox to others for a period as long as five days before their own symptoms begin. During this period, it is likely that these infectious individuals are unaware they are infected with the disease, therefore they continue with their normal daily routines. Thus, tools to analyze risk of becoming infected with chickenpox are needed.

Having contracted chickenpox once usually confers long immunity (7)(3). Due to the potentially severe nature of the disease when contracted by adults and the immunity conferred after having the disease as a child, many parents have held chickenpox parties, where susceptible children were intentionally exposed to a child infectious with chickenpox (12). Parents' hopes have been that their own children would contract the disease, have only a mild case of chickenpox, and then gain an immunity to the disease. However, in a small number of cases, children infected with chickenpox have developed a more serious case of the disease. "It is not possible to predict who will have a mild case of chickenpox and who will have a serious or even deadly case of disease."(3) Furthermore, chickenpox may persist in the body as a latent infection long after the initial illness. This infection may reactivate, causing the individual to have herpes zoster (also known as shingles). Shingles is more common in older individuals. Severe pain is a common symptom for individuals infected with herpes zoster (7)(5), therefore it is considered a more severe illness.

There is now an alternative to the chickenpox party. In 1995, the Varivax® live attenuated varicella virus vaccine (Merck & Co., Inc., Whitehouse Station, NJ, USA, http://www.merck.com) was licensed for use in the United States (7)(5). "This vaccine had a cumulative efficacy estimated at 70%-90% in preventing chickenpox in children followed for up to 6 years. Postlicensure effectiveness estimates ranged from about 85%-90% for prevention of all disease and 100% for prevention of moderate or severe disease. If an immunized person does get chickenpox, it is usually a very mild case with fewer lesions (usually less than 50, which are frequently not vesicular), mild or no fever and shorter duration of illness." (7) Additionally, if this vaccine is administered to an individual within 3 days (and possibly as long as 5 days) after exposure to chickenpox, in 70% to 100% of cases, it is effective in lightening the severity of the disease, and, in some cases, may actually prevent illness due to chickenpox entirely (7)(5). This vaccine may lead to the development of zoster later in life. However, the rate at which this happens in immunized individuals appears to be lower as compared to individuals who had been infected with the natural disease (7)(5). For all of these reasons, vaccination is now preferred over chickenpox parties as a method for developing immunity to chickenpox.

1.3. Overview of Thesis

This research focuses on tracking and analyzing exposure risk to chickenpox in an elementary school setting. While the work presented here uses a synthetic data set as an example, if actual data was used, the techniques presented in this thesis would facilitate an analysis of exposure risk to chickenpox. Data used during this analysis includes spatial, spatio-temporal, and non-spatial data. Tools were designed and built to aid in entering schedule data for students and tracking contacts that occur between students throughout their daily movements. ESRI ArcInfo® software (Environmental Systems Research Institute, Inc., Redlands, CA, USA, http://www.esri.com) is used in the analysis of contact data created by the custom-built tools.

In Chapter 2, a description of previous and related work is provided, thus establishing the context into which this work should be placed. In Chapter 3, a discussion concerning the modeling of exposure risk is provided, replete with descriptions of how data necessary to facilitate this analysis may be collected. Chapter 4 begins with an overview of the tools and techniques used in this research. A description of the synthetic data set is then included, and a detailed analysis of exposure risk to chickenpox using this data set is provided. In Chapter 5, ideas and suggestions for future work are discussed and the thesis is concluded.

The contribution of this work is in the application of techniques previously used for large, agent-based Monte Carlo simulations where little detail is known regarding the subjects to be modeled, to smaller, deterministic paradigms where significantly more is known about the subjects whose interaction dynamics are to be analyzed. Furthermore, the use of tools, techniques, and data from the domain of geographic information science (GIS) are used in the analysis of contact data. This work results in a good starting point for the study of smaller-scale populations and disease spread dynamics using the application of GIS to concepts within computational epidemiology.

CHAPTER 2

PREVIOUS WORK

The field of *epidemiology* is concerned with “describing and explaining disease occurrence in a population” and “developing, prioritizing, and evaluating public health programs.”(16) While the disease progression inside the human body and disease transmission between individuals are well studied, much less is known regarding disease manifestation and spread dynamics in populations. Epidemiologists study the occurrence and spread of diseases in populations (13), however they may only do so retrospectively. Once an outbreak has been detected, it would be unethical to allow the outbreak to continue for the purpose of gathering data. Therefore, epidemiologists’ efforts are primarily directed towards the mitigation of the outbreak, thereby eliminating the source of new data necessary to better understand the spread of the disease in the population.

Epidemiological models may be used to study “the population dynamics of the transmission of infectious agents and the potential impact of infectious disease control programs.”(15) These models allow epidemiologists to formulate interventions and study the effects on the spread of the disease, which obviously would not be possible in reality due to ethical concerns. Additionally, epidemiological models may be used in biological disaster planning, or in education of new epidemiologists and researchers. However, it is not possible to model reality with 100% accuracy. Assumptions concerning reality must be made and clearly stated in any epidemiological model, allowing epidemiologists and other researchers to better understand the relationship between reality and the model outcomes.

2.1. Mathematical Models

Most mathematical models are derived from a model referred to as the *SIR model.* The SIR model is named after the states an individual who acquires a disease may pass through.

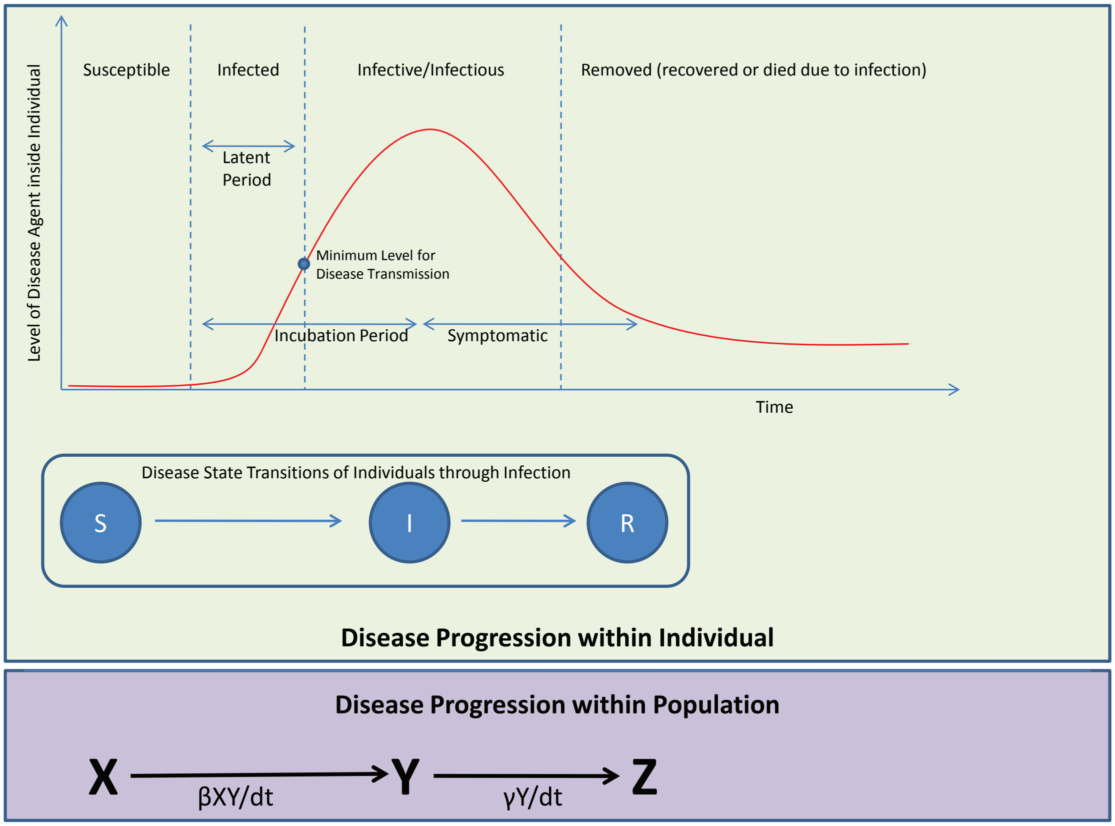


FIGURE 2.1. Level of Disease Agent inside an Individual as Related to the State Transitions from S, to I, to R, and the Population Densities X, Y, and Z of the SIR Model

A person who may acquire a certain disease is said to be susceptible to that disease. Once this person acquires the disease, this person is said to be infective. The person either gets better or dies from the disease. In either case, person is then said to be removed. For a certain disease, a person who gets better may become susceptible to that disease again, but for many others, long immunity is conferred, making this person removed from the population who may once again become infected.

The SIR model relies on a set of differential equations, tracking population densities X, Y, and Z of the number of susceptibles, infectives, and removals (respectively), such that

$$X + Y + Z = N \tag{1}$$

where N is the total population in the model with respect to the area in which this population exists. Population densities are used since one of the underlying assumptions of the SIR Model is that the population is closed. Another assumption is that the population is both well-mixed and homogeneous. Therefore, in a disease where transmission occurs through contacts between susceptible and infective individuals, it is assumed that all susceptible individuals are equally likely to come into contact with all Infective individuals during any length of time dt. The set of all contacts occurring in dt is then XY. An additional parameter β is used to limit the set of all contacts to those that actually did occur and further, to those that actually resulted in disease transmission. Therefore, the change in X over dt can be represented as

$$\frac{dX}{dt} = -\beta XY \tag{2}$$

since βXY individuals transition from X to Y in dt, and the population is assumed to be closed and infective individuals can only transition to the removed population Z, so X is never replenished. Another parameter γ represents the proportion of Y that transitions to Z over dt. Thus, the change in Y over dt can be represented as

$$\frac{dY}{dt} = \beta XY - \gamma Y \tag{3}$$

and the change in Z over dt can be represented as

$$\frac{dZ}{dt} = \gamma Y \tag{4}$$

Figure 2.1 shows the level of disease agent needed inside of an individual with relation to the progression of the disease and the individual's current disease state. In the bottom portion of this diagram, population densities X, Y, and Z are shown for the entire population and are related to the state diagram for individuals in that population. The parameters β and γ are also shown as determining the flow from X to Y to Z over dt (15).

When the SIR model is first initialized, initial population densities for X, Y, and Z must be assigned. If the value of Y grows beyond what it was initialized as, this is said to be an *epidemic*. This would occur if the change in Y over dt is greater than zero. Therefore, the requirement for an outbreak becoming an epidemic can be expressed by the following threshold condition:

$$X > \frac{\gamma}{\beta} \tag{5}$$

This allows researchers to adjust the parameters β and γ and the initial values of X, Y, and Z to model different interventions and observe their effectiveness.

Many variations on the SIR model exist, but the one perhaps most pertinent to this research is the SEIR model. The SEIR model adds an additional state, and therefore an additional, separate population density to be tracked, to the SIR model. This state is the exposed state. A susceptible individual becomes exposed due to contact with an infectious individual. However, this does not necessarily mean that this individual will become infected. Instead, a transition back to the susceptible state may be made instead of progressing on to the infectious state. This works to overcome two shortcomings of the SIR model. This exposed state gives an individual a state to reside in during the latent period depicted in 2.1. It also allows for a level of uncertainty regarding whether this individual is actually still in the process of acquiring infection at all, thus relaxing the assumptions of the SIR model.

2.2. Agent-Based Models

Much previous work has been done using agent-based models in computational epidemiology. However, these models investigate the spread of the disease rather than the risk of individuals in a population. In these models, individuals are represented as agents, and probabilities are computed in an attempt to more realistically guide the course of the many random experiments that are conducted to determine state transitions of each agent due to its exposures to disease. This methodology results in Monte Carlo simulations in which many trials must be conducted. This can become computationally expensive and results in outcomes that are likely far less specific than the large amounts of actual data used as parameters in such models.

The problem with Agent-Based Models lies in their needed specificity versus what is actually known about the movements of individuals in populations. Therefore, approximations of individuals' movements within populations must be created to use as input. In (8), (9), and (10),a separate model named TRANSIMS is described and used as a large-scale simulation of traffic systems. Output from this model are then used as input representing individuals' movements in an agent-based computational epidemiology model named EpiSims. However, the spatial granularity of the data resulting from TRANSIMS, and thus also used in these epidemiological simulations, is unrealistically low, placing roughly four locations in the area of each city block and then assuming each of these locations is well-mixed. Some of these locations "represent a collection of office buildings holding hundreds or thousands of workers."(8) Additionally, the TRANSIMS data assumes a "24-h approximate periodicity of people's contacts."(9) In certain specific smaller-scale simulations, this may be a reasonable assumption. However, when modeling entire populations of large cities, this becomes much more unrealistic. While a step in the right direction, any advantage of the specificity of representing individuals in the population as agents is likely lost by the incredibly low spatial granularity of the locations assumed to be well-mixed.

EpiSimdemics(4) is a more recent model that builds upon the concepts of the EpiSims model, but is modified to scale to simulate populations of 100 million individuals. Instead of using input data from TRANSIMS, EpiSimdemics uses a variety of more detailed input data obtained from a variety of differ

However, EpiSimdemics is built using the same ideas as its predecessors. In the 100 million-large population, each individual's movements, contacts, and disease state are tracked. These movements are determined before the simulation begins, though exceptions are computed to model the modified behavior of those who have become infected. The simulation progresses in *phases* of 24 hours. At the end of each phase, each susceptible individual's set of contacts during that phase are used to determine that individual's probability of having been infected during this phase. A random experiment is then conducted to determine whether each susceptible individual has caught the disease. Before the next phase begins, modified movements are determined for each of the infected individuals.

While EpiSimdemics is a major improvement over its predecessors, it maintains the same major drawback. Even though it accepts extremely detailed information concerning the movements and attributes of individuals in the simulated population, it still relies upon approximated input. While this approximated input is derived from a variety of actual data, it is highly unlikely that each individual's actual movements can be approximated using this data with any reasonable degree of accuracy. Therefore, by attempting to simulate such a large population at such a fine level using insufficient data, the computational costs of this technique become high enough to necessitate the use of High-Powered Computing, while the data resulting from the simulation should be viewed with a large degree of uncertainty.

The work presented in this thesis builds upon the techniques used in such models as EpiSims and EpiSimdemics, applying these techniques to a population in a setting where much more in known about the environment, individuals, and their movements. Assumptions from these models regarding contacts' dependency on co-location of individuals are used. These assumptions become more realistic when the spatial granularity is set at the room level rather than the quarter-of-a-city-block level. However, since significantly more is known about the subjects being modeled, the stochastic nature of these previous models is discarded in favor of a deterministic approach. This results in a blanket of exposure risk over the population rather than a simulated disease outbreak.

CHAPTER 3

MODELING EXPOSURE RISK

3.1. Obtaining Individual Schedules

Students in an elementary school generally follow a specific schedule. Often, students in a particular grade are assigned one of several different home rooms, and all students in a particular home room stay together for the majority of each school day. These home room schedules can account for the majority of movements of students in the school, but there are still additional movements not accounted for by these home room schedules alone.

It is the responsibility of the teachers to keep track of where their students are at any given time. For example, a student who would like to visit the restroom must first ask permission from his or her teacher. If this teacher grants the student permission to go to the restroom, the teacher then knows where the student is going and has a reasonable expectation concerning how long it will take the student to return. Many schools issue a hall pass specifying such information as the student's name, home room, and destination. If each teacher kept a log of all hall passes he or she issued (including student name, destination, departure time, and return time), this could be used in combination with the school's home room schedules to account for most of the students' movements in the school.

Schedules in elementary schools often repeat several times per week. In some schools, there may be variations on a single schedule that each repeat two or three times per week, and in other schools, the same basic schedule may repeat five times per week. Furthermore, many students may visit the restroom (or visit some other part of the school) at about the same time every day, five times per week. If found to be accurate in a particular school, this observation may be useful. If hall pass information is needed, but missing, for a certain day, another day's hall pass information may prove to be a reasonable substitution.

However, not all students attend the school everyday. For this reason, attendance information must be considered when determining movements in an elementary school. If a specific student is not present during a day in question, then this student should not be considered when modeling risks of exposure for that day. Therefore, this student must be temporarily removed from the appropriate home room student roster, and any movements this student makes during a normal school day must be temporarily suspended until the student returns to school.

If schedule information, hall pass logs, and attendance information are integrated together, this may paint a reasonably accurate picture of movements within the school. These movements may then be used to determine contacts between students in the case of an infectious disease outbreak. School administrators or public health officials, armed with movement information from the school, may be able to use this information to formulate location-based, schedule-based, or other interventions.

3.2. Contacts

A *contact* is an interaction among individuals that makes transmission of a certain disease possible. As described in the Introduction, different diseases are transmitted in different ways. Certain assumptions must be made regarding what constitutes a contact between two individuals such that, if one was susceptible and the other infectious, this contact would then be an exposure to the disease. In the case of chickenpox, these contacts can be airborne. Further, due to the highly contagious nature of this disease, as much as 80%-90% of contacts result in an exposure. Classrooms in elementary schools are usually small. Even large schools with hundreds of students in a particular grade usually break each grade apart into a set of much smaller home rooms. The small size of the rooms children must occupy for long periods of time, combined with the highly contagious nature of chickenpox, make the assumption that all individuals sharing a room together at the same time are well-mixed reasonable. In other words, a contact is assumed to exist between any pair of individuals occupying the same room at the same time.

A contact that lasts 10 seconds may have a smaller chance of leading to disease transmission than an otherwise equal contact that lasts two hours. Therefore, not all contacts should be considered equal. Longer contacts likely are more effective at transmitting the disease than are shorter contacts. Hence, information about a contact should not only include which pair of individuals were involved in that contact, it should also include the starting time and ending time of that contact, which may then be used to compute the time elapsed during that contact.

It is likely that each different room in the school has unique dimensions. A room's dimensions determine the volume of air the room may hold. Rooms containing larger volumes of air allow airborne pathogens to be more thinly dispersed, thus potentially lowering the quantity of airborne pathogen contained in each unit volume of air. Lower quantities of pathogen per unit volume of air may lessen the strength of contacts between individuals in the room. Therefore, contacts in larger rooms should not be considered equal to contacts in smaller rooms. To account for this, properties of the room in which a contact occurred should also be included in information describing a contact.

A single contact lasting two minutes may not necessarily be equal to a series of 30 second contacts spread out over an entire day (4). Therefore, it may not be realistic to aggregate these contacts by simply adding up the duration of all contacts between two individuals. While aggregating contacts in this way may be useful for certain types of analysis, it would also limit the amount and usefulness of the information resulting from the simulation. It is important that public health officials and school administrators be able to easily analyze contacts before such aggregation has occurred.

3.3. Case Study

As an example, we are considering the following scenario: The principal of an elementary school is informed by his staff that one of the school's students, Johnny Smith, is at home with chickenpox.

(i) What could this principal do to ensure that chickenpox does not begin to spread through the entire school?

(ii) How could this principal know who is at risk of having been infected with the disease?

Other students in Johnny's home room are likely at risk, but what about students in other home rooms? However, home rooms in elementary schools are rarely, if ever, held in complete isolation from each other. This means that many other students outside of Johnny's home room may also be at risk, but how can the principal know which students are likely at the greatest risk? To attempt to answer this question,we should first examine students' behaviors inside an elementary school.

In elementary schools, students are usually grouped into home rooms. Students in a particular home room share, for the most part, the same schedule. They all go to lunch together; they all go to recess together; and they all go to other classes such as music or P.E. together. School administrators have records of these schedules, and these schedules often repeat five times a week. Can the principal then use these records in his risk analysis?

While these records represent the overall movements of students in the elementary school, they are not complete. It is not unusual for a student to leave his home room (without the rest of the students in his home room) to visit the restroom, the water fountain, or to help the teacher run errands. Even though this is not unusual, teachers are required to keep track of their students' whereabouts at all times. Therefore, with some simple record-keeping, the principal may then be able to "fill in the gaps" of the students' overall schedules.

To accomplish this record-keeping, each teacher may be instructed to keep a list of every time one of his students left the rest of the students in his classroom. This list would include the student's name, the tie the student left, the student's destination, and the time the student returned. Having all of this information, could the principal then determine each student's risk of catching the disease? It is very difficult to make sense of such a large amount of schedule data. Because of this, it is still unlikely that the principal will be able to determine each student's risk of catching the disease.

It is the goal of this research to place all of the available schedule information into a context where it can be more effectively analyzed in terms of time (when), location (where), and individuals involved (who). This is done using an agent-based computer simulation to model and record contacts between individuals in the school. Contacts are important because they are necessary for the transmission of chickenpox. However, only a subset of all contacts, those contacts between a susceptible individual and an infected individual, may actually result in transmission.

The first person to acquire a disease during a disease outbreak is referred to as the "primary case." In our case study, if Johnny is the first individual at the school during this outbreak to have have caught chickenpox, then he is the index case. When the index case is known, we can be concerned about whom the index case had contacted while infectious.

Johnny likely became contagious and began spreading chickenpox one or two days before the onset of his symptoms. It can take an average of 14-16 days (but possibly as few as 10 or as many as 21 days) for a person who has been infected with chickenpox to become contagious, and then it may take another one to two days for this same person to become symptomatic. It is therefore likely that Johnny may have spread chickenpox to many of his classmates who are now infected, but are not yet contagious or symptomatic. If school administrators could get an idea of which individuals in the school were most likely to have come into contact with Johnny, then they may be able to take actions to contain or hinder the further spread of the chickenpox outbreak.

The records mentioned earlier lack proper context to determine each individual's contact to Johnny over the one to two days prior to the onset of his symptoms. Therefore, computational tools must be designed and constructed to track (based on the records) the movements of individuals in the school during the one to two days prior to the onset of

Johnny's symptoms. The result would then be a list of all contacts (that may have lead to transmission of chickenpox) between all students.

However, the amount of data resulting from the simulation is likely to be so large that additional tools are needed in order to analyze it. This is where geographic information system (GIS) tools can be very useful. ESRI ArcInfo® software (Environmental Systems Research Institute, Inc., Redlands, CA, USA, http://www.esri.com) can be used to analyze the data using a variety of methods. Additionally, if a floor plan of the school building is available, data from the computational simulation can be overlaid on top of the floor plan using a variety of methods.

CHAPTER 4

ANALYSIS OF EXPOSURE RISK

4.1. Contact Simulator

The contact simulator consists of the three parts shown in Figure 4.1: the Schedule Builder, the Contact Tracker, and the Contact Summarizer. After the three parts of the simulator are used in the order listed, data can then be analyzed using ESRI ArcGIS. While there is no smallest time unit implemented in any part of the simulator or analysis methodology, one second will be used as a convenient minimum time unit. The remainder of this chapter describes the three parts of the contact simulator.

4.1.1. The Schedule Builder

The Schedule Builder was written in Perl and provides a menu-driven interface that facilitates the building of the schedule, based on the actual data that may be realistically

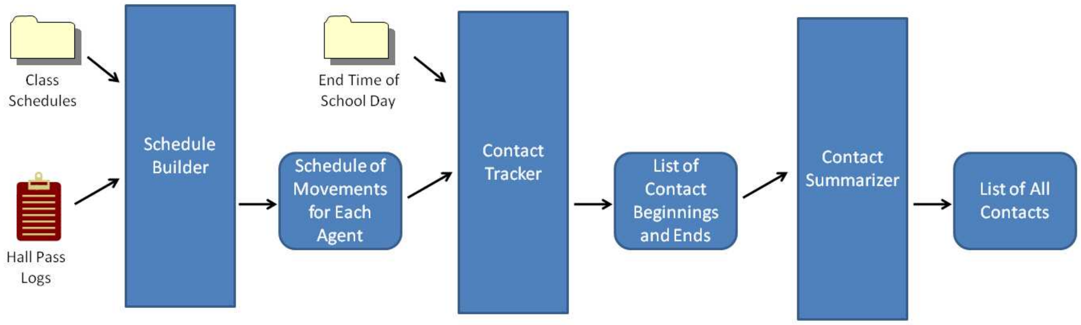
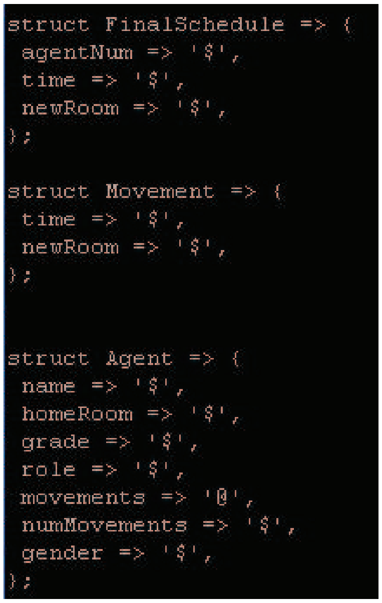


Figure 4.1. How the Different Parts of the Simulator Work Together



```
struct FinalSchedule => {
 agentNum => '$',
 time => '$',
 newRoom => '$',
};

struct Movement => {
 time => '$',
 newRoom => '$',
};


struct Agent => {
 name => '$',
 homeRoom => '$',
 grade => '$',
 role => '$',
 movements => '@',
 numMovements => '$',
 gender => '$',
};
```

Figure 4.2. Perl Structures Used in Schedule Builder to Maintain Student Information

collected by the school. It is designed specifically for use with modeling elementary schools. There are two types of information maintained by the Schedule Builder: how many students have been entered, and what movements have been entered for each student. The movements entered for each student are maintained by the following Perl structures shown in Figure 4.2.

As described earlier, elementary schools usually assign each student to a home room. Also, all students in a particular home room may move from room to room together, throughout a normal school day. Therefore, in order to avoid having to enter the same basic schedule for each student in a particular home room, students are created by the creation of entire home rooms of students. The user may specify how many students are in a particular home room being created. In the unusual case that a student does not belong to any home room, a unique home room containing only this one student may be added.

Once a home room has been created, movements may be added for all students in that home room. No location is automatically added when the home room is created, so the earliest movement added should be at the time school starts in the morning and in the room in which this home room starts school. Movements do not have to be added in chronological order. Once all of the movements of each home room of students have been entered, movements of individual students may be added. The movements do not have to be added in chronological order, either.

Data concerning movements for entire home rooms of students may be obtained from class schedule information kept my school administrators. Movements data missing from the class schedule information may be obtained from the hall pass logs described earlier in the Obtaining Individual Schedules section. Once these two sets of data are combined, they may paint a reasonably accurate picture of students' movements in the school. Once all home room and individual movements have been added, the Schedule Builder creates a schedule file for input into the Contact Tracker.

4.1.2. The Contact Tracker

The Contact Tracker required the use of complex data structures using pointers and was written in C++. It accepts as the students' schedules the output produced by the Schedule Builder. It also accepts a parameter for the time of the end of the day. It uses these data to track contacts as they are made throughout a simulated school day.

Individual students are represented by agents. At any time during the simulation, each student is in a room. Each room is assumed to be sufficiently small that all agents occupying a particular room at the same time are considered to be uniformly mixed. This assumption means that a contact capable of spreading the disease occurs any time two agents occupy the same room at the same time.

Due to this assumption, at any point during the simulation, agents exist in a series of cliques of a disconnected graph as shown in Figure 4.3. All agents in a particular room at a particular time form a clique. This clique is disconnected from any other part of the graph. The weight of edges between any two agents in a clique is a function on the properties of the room they both occupy and the duration of their current contact. The graph may then be viewed outside of the context of the environment in which the contacts occurred as shown in Figure 4.4.

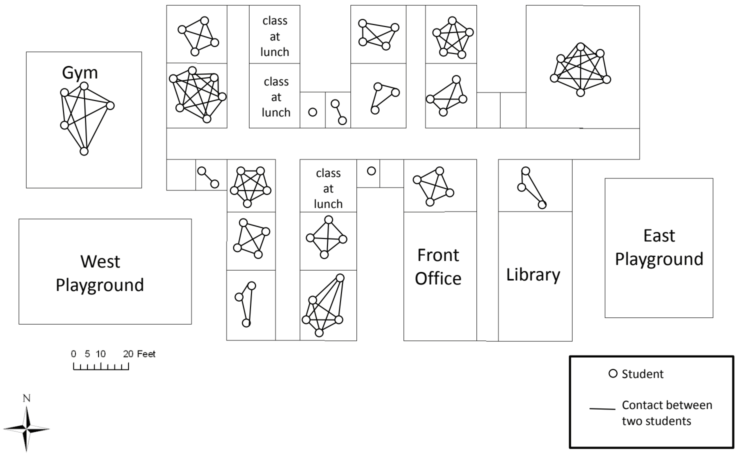


FIGURE 4.3. Example of Cliques Formed by Agents at Any Point during the Simulation

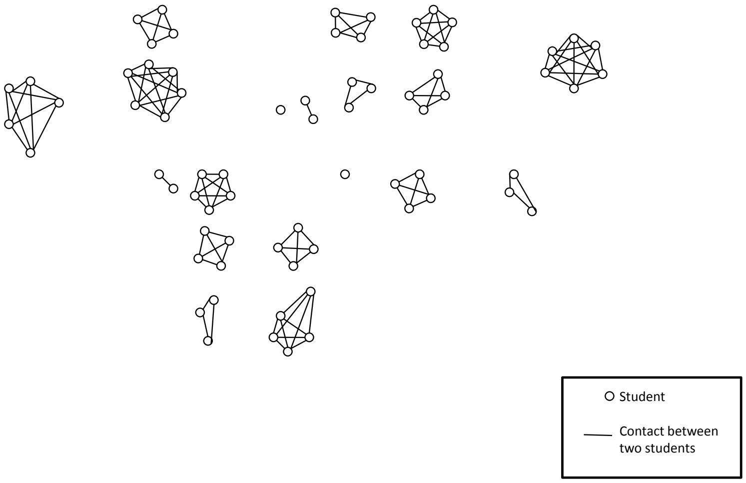


FIGURE 4.4. Graph Resulting from Simulated Groups of Agents Occupying Rooms Together

The Contact Tracker simulation is event based, so simulated time instantaneously jumps from one scheduled event to the next. Every time an agent enters a new room, a contact is begun. The Contact Tracker then records the room in which the contact took place, the two agents involved, and the time the contact begun. Likewise, when an agent leaves a room, all contacts with other agents in this room are ended. The room in which the contact took place, the two agents involved, and the time the contact ended are recorded by the Contact Tracker. All recorded contact beginnings and endings are then output to a log file. Contact beginnings are differentiated from contact endings in the contact log file by using a Boolean value associated with each entry in the log file.

Once all events have been processed, the Contact Tracker uses the ending time of the school day to determine the end of the last set of contacts. At the time when the school day ends, each agent is sent to a separate location, not tracked by the the Contact Tracker, effectively ending all contacts by sending each agent to a unique home.

4.1.3. The Contact Summarizer

The Contact Summarizer required a great deal of string manipulation and was written in Perl to take advantage of using regular expressions. It reads as input the contact log file produced by the Contact Tracker. It then finds corresponding entries for the beginning and end of each contact and combines them. It outputs a list of all contacts with the following data for each contact: the first agent, the second agent, the contact start time, the contact end time, and the room in which the contact took place.

4.2. Analysis of Risk Using Contacts

Once all contacts have been found, the problem then lies in analyzing these contacts to assess risk of becoming infected. This risk can be analyzed according to either location of contacts or by each individual's contacts.

4.2.1. Analysis of Individuals' Risk

Analyzing the risk levels of individuals in the model can be useful for determining what path a disease may take to spread from individual to individual. This allows for the discovery of behavior patterns that may tend to put certain individuals at a higher risk of contracting the disease. Different methods of aggregating contacts for assessing and analyzing individuals' risk levels can be used. All individuals represented in the model fall into one of the following three sets at any given time:

A_{SE} - the set of Susceptible and Exposed individuals

A_I - the set of Infectious individuals

A_R - the set of individuals who, for some reason, are not Infectious and who may not contract the disease. This could be due to successful vaccination or long immunity conferred from having previously gone through the course of the disease. Therefore, the total population of agents A represented in the model is $A = A_{SE} + A_I + A_R$.

4.2.1.1. Aggregating Individual Contacts by Duration

It is likely that a susceptible individual will have multiple contacts with infectious individuals. In this case, how can this susceptible individual's risk of becoming infected be determined using the known contacts? A variety of methods for aggregating an individual's contacts can be used. The simplest method of aggregating these contacts involves adding each individual's total contact time with an infectious individual. For example, if a susceptible individual had three contacts of duration t_1, t_2, and t_3 with infectious individuals, then the total duration T of these contacts may be represented as the sum of the three durations, $T = t_1 + t_2 + t_3$.

Using this method, each individual's risk level is proportional to the total duration of his or her total contact time. Total contact time for all individuals can be calculated and then used to sort all individuals into descending order, thus placing individuals in descending order according to their risk levels. However, this method fails to take into account attributes of the locations in which each of an individual's contacts occurred.

4.2.1.2. Aggregating Individual Contacts by Duration and Attributes of Location

All individuals in this model who occupy a particular location at the same time are considered to be uniformly mixed. However, contacts between individuals in certain locations may be stronger than contacts between individuals in other locations. For example, a contact between two individuals in a small room is likely to be stronger than a contact of equal duration in a much larger room. Therefore, attributes of the location in which a contact occurred should be taken into account and used in conjunction with the contact duration in that room. While these attributes may include many different factors, the volume of each room is used as a room-specific parameter in this research.

If a susceptible individual had three contacts of durations t_1, t_2, and t_3 that occurred in rooms of volumes v_1, v_2, and v_3, respectively, then the total weight of these contacts could be represented as $T = \frac{t_1}{v_1} + \frac{t_2}{v_2} + \frac{t_3}{v_3}$. The list of individuals can be sorted in descending order by their total contact weights, and this list would also be an ordering of individuals' risk levels.

4.2.1.3. Using Individuals' Contacts to Determine Probabilities of Disease Transmission

In probabilistic simulations such as EpiSimdemics(4) or the Global Stochastic Field Simulator(14), determining the probability an exposed individual will become infected with chickenpox given a single contact of known duration with an infectious individual in a room of known volume may be a difficult problem. This determination depends on factors such as the virulence of the disease and the immune system and response of the susceptible individual. Therefore, an approximation of this probability would likely be $p = \frac{I*V*t}{v}$, where I represents the susceptible individual's immune response, V represents the virulence of the disease, t is the duration of the contact, and v is the volume of the room in which the contact occurred.

While I and V may be difficult to determine, t and v are known through the results of this model. V is a property of the disease and may therefore be assumed to remain constant outside of abnormal atmospheric changes or shifts in the disease itself. I may vary from person to person, but, for the purposes of this analysis, can be assumed to remain sufficiently constant for each individual through the duration of this model. Hence, an individual i's probability of becoming infected can be analyzed by recognizing that $p_i \propto \frac{t}{v}$.

However, it is likely that an exposed individual will have more than just a single contact with infectious individuals, necessitating a method of aggregating the contacts that do occur. The method of aggregating contacts by simply adding them together is insufficient when dealing with probabilities. If two contacts X and Y occurred (less than the sum of the lengths of the latent period, infectious period, and any period of immunity apart), then it is insufficient to simply represent the total probability of becoming infected as $p(X) + p(Y)$.

This could result in a probability greater than 1, and all probabilities p_i must be constrained to the range $0 \leq p_i \leq 1$.

To aggregate the probabilities of these two contacts X and Y leading to infection, the following observation must be made: An individual may only be infected as a result of contact Y if that individual was not already infected as a result of contact X. Therefore, once an individual is infected once, any immediately following contacts can no longer infect this individual and may be disregarded. While something is known regarding the probabilities of infection associated with each contact, whether a contact actually does lead to infection remains unknown. Therefore, in the case of the two contacts X and Y, the probabilities of these two contacts must be aggregated as follows: $p(X) + p(Y \mid !X)$. This is the probability that at least one of the two contacts X or Y will lead to infection with the disease.

All individuals $a_i \in A_{SE}$ start out by having a probability of zero of becoming infected with the disease. In the case that a contact occurs between two individuals a_i and a_j such that $((a_i \in A_{SE}) \oplus (a_j \in A_{SE})) \wedge ((a_i \in A_I) \oplus (a_j \in A_I))$ is True, then the probability of infection of the individual a_x (the individual in (a_i, a_j) that is in set A_{SE}) is denoted as a_{x_p} and must be updated by $a_{x_p} = \frac{v}{t} * (1 - a_{x_p}) + a_{x_p}$. However, the goal of this research is to use available, highly specific, deterministic information, thus eliminating the immediate need for such probabilities.

4.2.2. Analysis of Risk by Location

Risk may also be analyzed by location. Each contact that occurred has a location associated with it where it occurred. Analyzing risk by location may reveal spatial, or even behavioral, patterns within the modeled environment in which individuals may be at higher risk of contracting the disease. This analysis may be with respect to those individuals known to be infectious, or it may be more generalized for cases where infectious individuals may not necessarily be known. Tools from Geographic Information Science (GIS) may be employed to easily conduct this analysis and visualize the resulting data spatially.

4.2.2.1. Analyzing Risk by Locations when Infectious Individuals Are Known

When the set of infectious individuals in the population is known, contacts that occur between these individuals and the set of susceptible individuals are extracted from the set of all contacts and examined according to contact location. The contacts for each location may be aggregated by finding the total duration of contacts in a location. One or more attributes of each location (such as a room's volume) may be taken into account in order to better compare the levels of exposure (and therefore risk) that occurred in each of the modeled locations.

Locations with higher total contact levels are locations where individuals are likely at higher risk of contracting the disease. These locations should be examined more closely. In environments such as elementary schools where administrators have control over schedules, which, in turn, may affect the movements of individuals within their schools, public health officials and school administrators may wish to intervene by strategically modifying schedules in order to reduce the number of contacts that occur in locations of higher risk. Other interventions may also be formulated with respect to the disease, the population, and what is known about the contacts in each room.

4.2.2.2. Analyzing Risk by Locations when Infectious Individuals Are Unknown

Public health officials and school administrators may wish to plan for disease outbreaks in a school before the outbreak actually occurs. In this case, the set of infectious individuals in the population may not be known. Disease transmission occurs through contacts between Susceptible and Infectious individuals. These contacts are a subset of all contacts that occur in the population during the model. Therefore, analyzing the modeled locations using the contacts that occur at these locations may result in interesting findings regarding how to lessen the total number (or total strength) of the contacts during the course of the model.

If many contacts occur in a certain room, is there anything that can be done to lessen the number of contacts in this room? For example, if many contacts are observed to occur in a certain bathroom, then could schedules be rearranged to lessen the number of contacts in this bathroom? Could other interventions ranging from the installation of ultraviolet lights(18) to increased ventilation be undertaken? Could an extra bathroom be installed in an attempt to lessen the number of contacts that occur in this room? These are just some of the questions that may be asked as a result of analyzing risk by locations when infectious individuals are unknown.

4.2.2.3. Using Geographic Information Science Tools to Facilitate Analysis and Visualization of Data

While the analysis methods discussed above may be easily implemented as an additional module in the simulator, using the database tools in the ESRI ArcInfo® software (Environmental Systems Research Institute, Inc., Redlands, CA, USA, http://www.esri.com) allows for more flexibility in analyzing exposure risk with regard to either individual students or to the different locations in the school. Further, existing GIS expertise in public health is exploited, thus making public health officials more comfortable and more likely to embrace and use this computational tool. The list of contacts output from the simulator can be imported as a database table in ArcInfo. This data can then be manipulated, normalized, and summarized using the database tools in ArcInfo. The resulting analyzed data may then be graphed, exported, or visualized spatially.

An ESRI shapefile(1) of a floor plan of the simulated building may be used (or easily created if one is unavailable) to visualize the analyzed data spatially. The database tools in ArcInfo software allow joins and relates between tables of data and spatial features, such as the polygons that represent individual rooms in the building floor plan. Finally, a variety of symbology and/or categorization techniques may be used to visualize the rooms in which the most contacts occurred.

4.2.3. Synthetic Population, Schedules, and Environment

A synthetic population of students with corresponding schedules was created along with a synthetic school building. Each student was assigned an integer student ID number in the range [0,274]. The gender of the students was split roughly 50/50, using even and

Table 4.1. Home Room Student Assignments and Grades

Home Room	*Grade*	*Range of Student ID's*
4	1	0 - 24
5	2	75 - 99
6	1	25 - 49
7	2	100 - 124
8	1	50 - 74
9	5	300 - 324
10	5	325 - 349
11	5	350 - 374
12	4	225 - 249
13	4	250 - 274
14	4	275 - 299
15	3	150 - 174
16	2	125 - 149
17	3	175 - 199
18	3	200 - 224

odd student ID numbers to designate male and female students, respectively. As shown in Table 4.1, each student was assigned to one of fifteen different home rooms, and each home room was assigned 25 students. Individual home rooms were assigned to a specific grade 1-5. The home rooms were distributed among the grades by assigning three home rooms to each grade.

A detailed schedule for the entire school was created including an activity schedule rotation using the three activities of music education, physical education, and library education. These activities correspond to specific locations within the school. Physical education occurs in a gymnasium, music education occurs in a designated music room, and library education

Table 4.2. Lunch Rotation Schedule Assuming One Serving Line

Home Room	*Grade*	*To Lunch*	*To Recess*	*To Restroom*	*To Home Room*
4	1	11:30 A.M.	12:00 P.M.	12:20 P.M.	12:30 P.M.
6	1	11:35 A.M.	12:05 P.M.	12:25 P.M.	12:35 P.M.
8	1	11:40 A.M.	12:10 P.M.	12:30 P.M.	12:40 P.M.
5	2	11:45 A.M.	12:15 P.M.	12:35 P.M.	12:45 P.M.
7	2	11:50 A.M.	12:20 P.M.	12:40 P.M.	12:50 P.M.
16	2	11:55 A.M.	12:25 P.M.	12:45 P.M.	12:55 P.M.
15	3	12:00 P.M.	12:30 P.M.	12:50 P.M.	1:00 P.M.
17	3	12:05 P.M.	12:35 P.M.	12:55 P.M.	1:05 P.M.
18	3	12:10 P.M.	12:40 P.M.	1:00 P.M.	1:10 P.M.
12	4	12:15 P.M.	12:45 P.M.	1:05 P.M.	1:15 P.M.
13	4	12:20 P.M.	12:50 P.M.	1:10 P.M.	1:20 P.M.
14	4	12:25 P.M.	12:55 P.M.	1:15 P.M.	1:25 P.M.
9	5	12:30 P.M.	1:00 P.M.	1:20 P.M.	1:30 P.M.
10	5	12:35 P.M.	1:05 P.M.	1:25 P.M.	1:35 P.M.
11	5	12:40 P.M.	1:10 P.M.	1:30 P.M.	1:40 P.M.

occurs in the library. Following each home room's activity period, that home room has a ten minute restroom break.

The two different lunch schedule rotations shown in Tables 4.2 and 4.3 were also created. The schedule in Table 4.2 was created assuming the cafeteria had only a single serving line, and, therefore, only one home room of students were sent to the cafeteria at a time. The lunch schedule rotation shown in Table 4.3 was created assuming that two serving lines existed in the cafeteria, allowing two home rooms of students to be sent to the cafeteria and served simultaneously.

Both lunch rotations included a 30 minute lunch, a 20 minute recess, and a ten minute restroom break. Home rooms were sent to lunch generally in five minute intervals (one or

Table 4.3. Lunch Rotation Schedule Assuming Two Serving Lines

Home Room	*Grade*	*To Lunch*	*To Recess*	*To Restroom*	*To Home Room*
4	1	11:30 A.M.	12:00 P.M.	12:20 P.M.	12:30 P.M.
6	1	11:30 A.M.	12:00 P.M.	12:20 P.M.	12:30 P.M.
8	1	11:35 A.M.	12:05 P.M.	12:25 P.M.	12:35 P.M.
5	2	11:35 A.M.	12:05 P.M.	12:25 P.M.	12:35 P.M.
7	2	11:40 A.M.	12:10 P.M.	12:30 P.M.	12:40 P.M.
16	2	11:40 A.M.	12:10 P.M.	12:30 P.M.	12:40 P.M.
15	3	11:45 A.M.	12:15 P.M.	12:35 P.M.	12:45 P.M.
17	3	11:45 A.M.	12:15 P.M.	12:35 P.M.	12:45 P.M.
18	3	11:50 A.M.	12:20 P.M.	12:40 P.M.	12:50 P.M.
12	4	11:50 A.M.	12:20 P.M.	12:40 P.M.	12:50 P.M.
13	4	11:55 A.M.	12:25 P.M.	12:45 P.M.	12:55 P.M.
14	4	11:55 A.M.	12:25 P.M.	12:45 P.M.	12:55 P.M.
9	5	12:30 P.M.	1:00 P.M.	1:20 P.M.	1:30 P.M.
10	5	12:30 P.M.	1:00 P.M.	1:20 P.M.	1:30 P.M.
11	5	12:25 P.M.	12:55 P.M.	1:15 P.M.	1:25 P.M.

two at a time depending on the lunch rotation schedule) starting at 11:30 A.M. until the last home room was sent to lunch. During recess, students were sent to either the East Playground or West Playground. The schedule alternated between the two playgrounds in order to minimize inter-grade student mixing during recess.

Four pairs of gender-specific restrooms were located in the school. During each home room's restroom break, each student from this home room visited the restroom appropriate to his or her gender in a specific pair of restrooms. Students were assumed to be in the restroom for the entire ten minute duration of each restroom break.

In this school, no two rooms have the same area or volume. In order to more accurately simulate a real-life scenario, the map shown in Figure 4.5 was actually drawn using the North

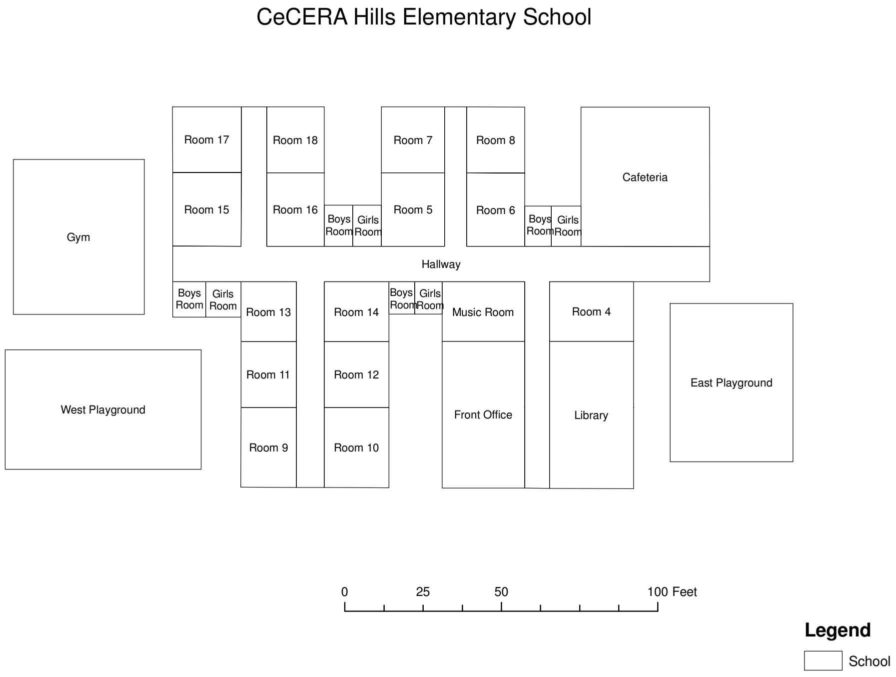


Figure 4.5. Map of School Used with Synthetic Data

American Datum of 1983, State Plane Coordinate System - North Central Texas projection. In a real-world use of these techniques, spatial GIS data regarding the layout of the school may be available as an ESRI shapefile(1). In this case, this pre-existing data could be used (even if some modifications or additions to the specific coordinates of the building are needed) in the same way the shapefile of the example school is used in this analysis. Also, due to the fact that the layout of the school was already in a projected coordinate system, the area of each room could be easily found using the Calculate Geometry features in the ArcInfo software.

Instead of using the area of each room in this analysis, the volume was used. This allows researchers to differentiate between the potential openness of different rooms using each room's height. For example, the gymnasium is likely to have a much higher ceiling than a normal classroom. An outdoor playground can be assigned an extremely large value for the height to represent its openness. For example, the height for the two playgrounds used in this sample school were set at 10,000 feet. Having the height and area of each room in the school, the volume can easily be found using the Field Calculator tools in the ArcInfo software.

While the Schedule Builder, Contact Tracker, and Contact Summarizer modules were capable of tracking students through the halls to their destinations, students were assumed to instantaneously move from source to destination. With more detailed spatial, nonspatial, and schedule data, these assumptions could be relaxed, allowing simulated students to move through different hall segments in order to get to their destinations. Also, while the smallest unit of time tracked in this example was the minute, the modules are flexible and may track either larger or smaller units of time, ranging from seconds to hours or days.

In addition to analyzing the contacts by individual, they were also analyzed by location. Using ArcInfo software to manipulate the contact data, different aspects of the contacts were summarized and joined to a map of the school for visualization. This analysis was performed both in the case of a known infectious student, and in the case where no infectious student was known.

The set of all contacts was imported into the ArcInfo software. Tools within ArcInfo were used to determine the total duration of each contact, extract contacts using SQL queries, and summarize the contacts. Finally, the table of contacts and data table for the polygon school features were joined in order to further analyze and visualize the data.

4.2.4. Analysis with Known Infectious Individual

To facilitate this analysis, a first grade, female student with ID number 1 in home room 4 is assumed to have been infectious during the day being analyzed. This student is assumed to be infectious, but not yet symptomatic. Therefore, while this student is spreading the infection to her classmates, she is proceeding with her normal daily schedule and routines. In the following sections analyzing contacts by locations, only the locations visited by the infectious student, highlighted in Figure 4.6, are included in the figures.

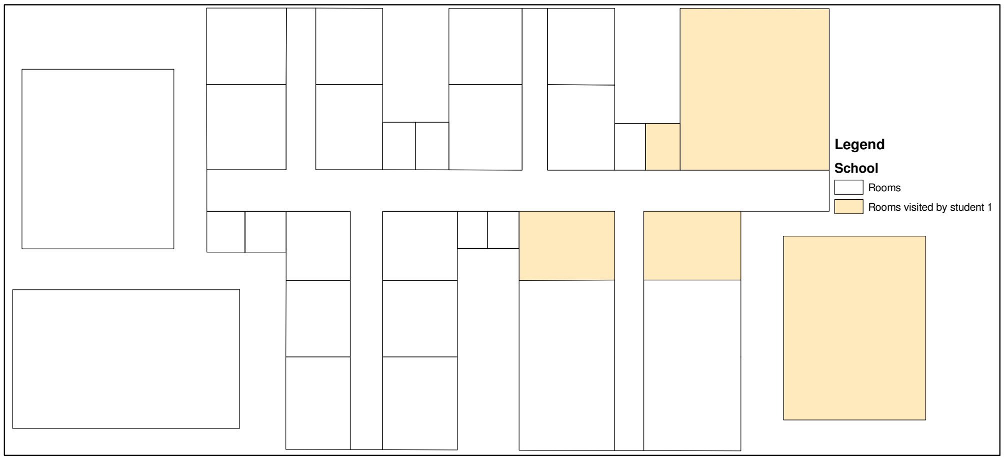


FIGURE 4.6. Locations (Highlighted) Visited by the Infectious Student

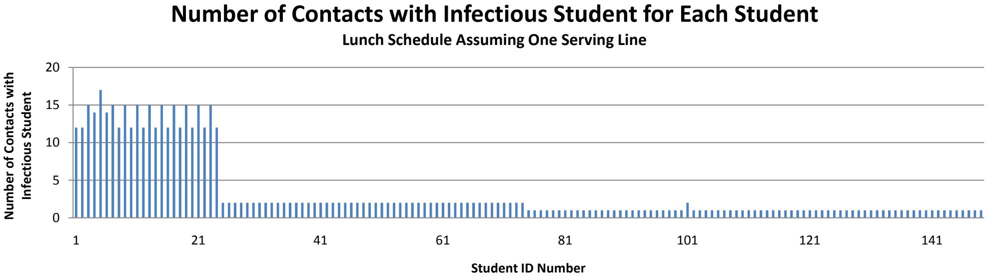


FIGURE 4.7. Number of Contacts with Infectious Student - Using Lunch Schedule Assuming One Serving Line

4.2.4.1. Analysis by Individual of Number of Contacts with Infectious Individual

The contacts involving the infectious student were extracted from the set of all contacts. Using the tools in the ArcInfo software, this table of contacts was then summarized to show the total number of contacts each student in the population shared with the infectious student. Finally, the resulting summarized records were sorted in descending order by the total number of contacts.

Students in the same home room as the infectious student had the highest number of contacts with the infectious student. Female students in this home room had more contacts with the infectious student than male students due to the gender-specific restrooms used

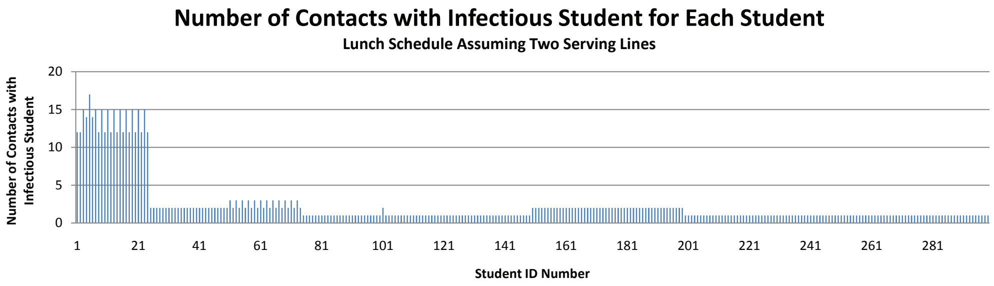


FIGURE 4.8. Number of Contacts with Infectious Student - Using Lunch Schedule Assuming Two Serving Lines

during restroom breaks. Additionally, one female student in this home room had a higher number of contacts due to visiting the restroom while using a hall pass.

During the schedule assuming only one serving line, due to the method of rotation of the lunch schedule, students of different grades were kept largely separate throughout their daily schedules. As can be seen in Figure 4.7, this resulted in minimal contacts of students in different grades from the infectious student. In fact, no contacts occurred between the infectious student and students in third, fourth, or fifth grade. With a single exception, only one contact occurred between each second grade student and the infectious student. The exception involved a student who was visiting the restroom using a hall pass while the infectious student was visiting this same restroom.

The lack of inter-grade mixing was a direct result of the rotation of the lunch schedule by grade, combined with the alternating use of the two playgrounds for recess. When the lunch schedule was rearranged such that all home rooms of a certain grade were no longer grouped together to go to lunch, then a different pattern emerged regarding contacts and students' grade levels. When all other aspects of the students' schedules remained the same, the patterns resulting from the lack of inter-grade mixing broke down. Instead, it became obvious that the differing contact levels among the student population were affected far more by home room assignment than by grade level.

During the schedule assuming two serving lines, Figure 4.8 shows that additional contacts occurred, but they were still largely affected by home rooms. The female students in the

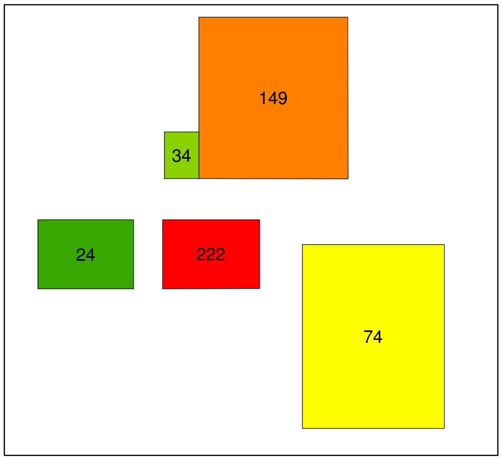


Figure 4.9. Number of Contacts in Each Room - Using Lunch Schedule Assuming One Serving Line

same home room as the infectious student still had higher numbers of contacts than did their male classmates. Additionally, the single female student who shared time in a restroom with the infectious individual still has a higher number of contacts than the rest of her home room classmates.

4.2.4.2. Analysis by Location of Number of Contacts with Infectious Individual

As shown in Figure 4.9, using the schedule assuming one serving line, the highest number of contacts occurred in the infectious student's home room. This number is elevated because the infectious student visited the restroom using a hall pass, effectively ending any contacts in her home room when she left the room, and starting new contacts when she returned. While this elevates the total number of contacts, it does not necessarily increase total contact time. On the contrary, it may serve to decrease total contact time.

However, as shown in Figure 4.10, when the schedule assuming two serving lines was examined, the numbers of contacts in both the cafeteria and the playground are elevated. In the case of the cafeteria, they are nearly doubled. This is due to more students being in the cafeteria at any time. More specifically, the infectious student has contacts with every student in the school with the exception of the three fifth grade home rooms. The number of contacts that occur in the playground are not as elevated as those in the cafeteria due to a

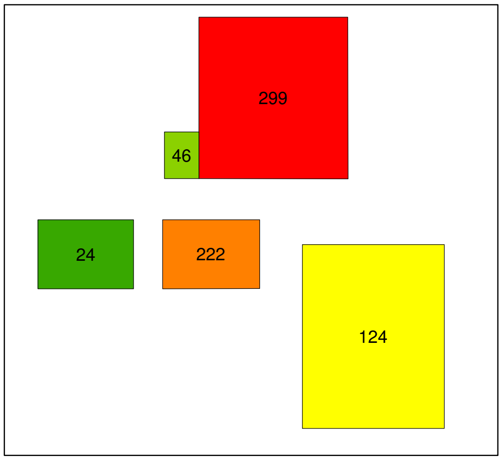


Figure 4.10. Number of Contacts in Each Room - Using Lunch Schedule Assuming Two Serving Lines

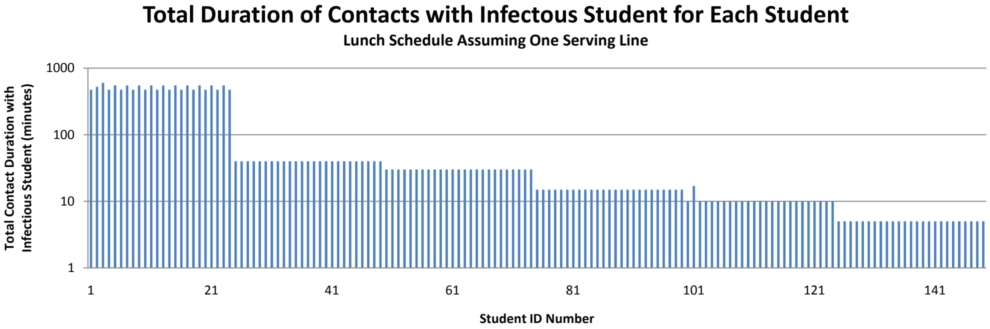


Figure 4.11. Duration of Contacts with Infectious Student - Using Lunch Schedule Assuming One Serving Line

continued rotation between the East and West playgrounds. Finally, since nothing but lunch rotations are different between the two schedules, the number of contacts in the infectious student's home room and in the Music Room are unaffected.

4.2.4.3. Analysis by Individual of Total Duration of Contacts with Infectious Individual

Tracking the duration of contacts, instead of just their total number, is likely a better measure of exposure risk. Using each of the lunch rotation schedules, the total duration of contacts for each student with the infectious student was computed and can be seen in Figures 4.7 and 4.8. The results are similar to the total number of contacts with the infectious student. Each student's contact level still seems to be largely based upon his or her home

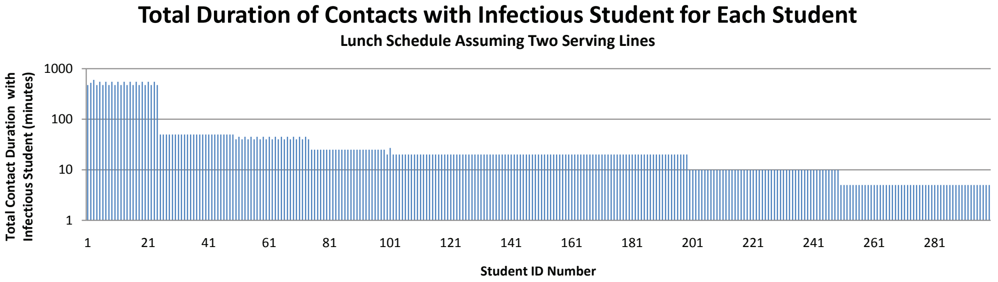


FIGURE 4.12. Duration of Contacts with Infectious Student - Using Lunch Schedule Assuming Two Serving Lines

room. However, many differences between the students' contact levels can be observed when examining the total contact durations; differences unseen when only examining the total number of contacts. Further, the relationship between the different students' contact levels can be more precisely observed. For example, the total contact duration of the female student who shared time in a restroom with the infectious student is nearly indistinguishable from that of her classmates. This is due to the fact that the two students were only co-located in the restroom for a few minutes.

4.2.4.4. Analysis by Location of Total Duration of Contacts with Infectious Individual

Examining Figures 4.13 and 4.14, the largest total duration of contacts occurred in the infectious student's home room. This is expected considering most of the infectious student's time during the school day is spent in this home room with 24 other students. Additionally, the total contact duration in both the infectious student's home room and in the Music Room are unchanged across the two different schedules. This is because only the lunch rotation changed from one schedule to the next. The total time spent in locations outside of the lunch rotation were not affected.

In the schedule assuming two serving lines, the total contact duration in the cafeteria is nearly doubled as compared to the other schedule. In this schedule, more students are in the cafeteria at any given time. Additionally, these students are co-located in the cafeteria with the infectious student for longer periods of time.

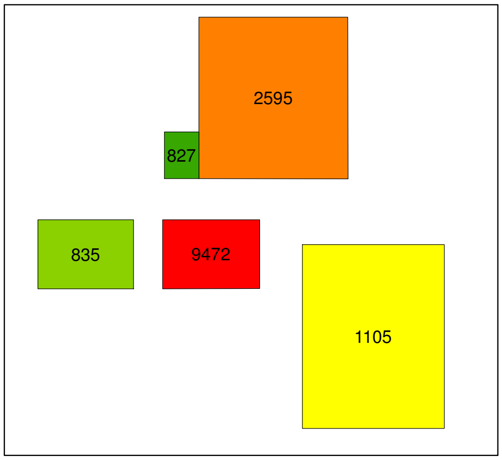


FIGURE 4.13. Map of Locations where Infectious Agent Had Contacts - Using Lunch Schedule Assuming One Serving Line

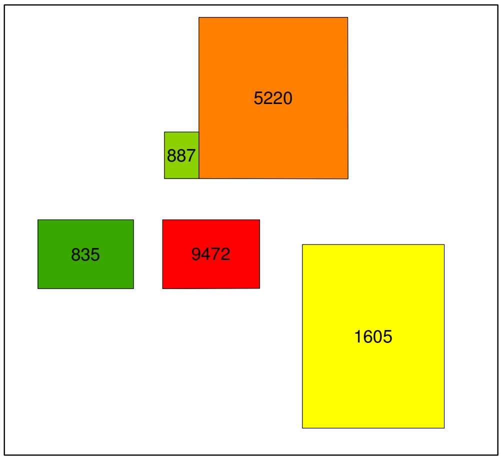


FIGURE 4.14. Map of Locations where Infectious Agent Had Contacts - Using Lunch Schedule Assuming Two Serving Lines

The average duration of contacts with the infectious individual are almost identical across both schedules. In the cafeteria, it is slightly elevated in the schedule assuming two serving lines. This is due to an extra five minute interval spent with 25 students of Home Room 6.

4.2.4.5. Analysis by Individual of Total Contact Density with Infectious Individual

Contacts were also tracked using a combination of the duration of each contact and the volume of the room in which each contact occurred. This gave more weight to contacts that occurred in smaller rooms than contacts of the same duration that occurred in larger rooms.

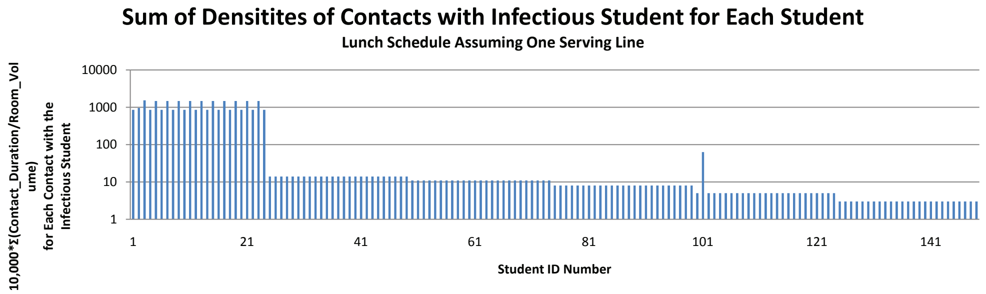


FIGURE 4.15. Sum of Density of Contacts with Infectious Student - Using Lunch Schedule Assuming One Serving Line

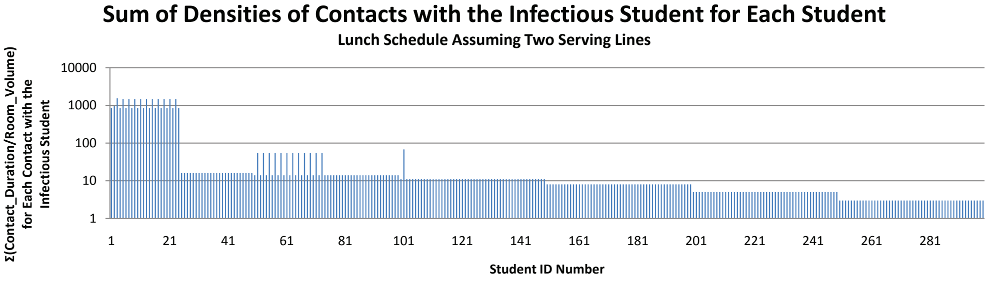


FIGURE 4.16. Sum of Density of Contacts with Infectious Student - Using Lunch Schedule Assuming Two Serving Lines

As shown in Figure 4.15, this resulted in a very noticeable difference between those students in the same home room as the infectious student and those who were not. The difference is due not only to the higher total contact duration of students in the same home room, but also to the significantly smaller volume of the classrooms as compared to the cafeteria or the playgrounds. The student who shared time in a restroom with the infectious student can also be seen on this graph more easily than on the total contact duration graph. This is due to the significantly smaller volume of the restroom in which this contact occurred.

The small contribution from contacts with the infectious student in the cafeteria can be seen in Figure 4.16. In this scenario, two home rooms of students were sent to lunch at the same time. The students with ID numbers 50 - 74 in home room 8 were sent to the cafeteria at the same time as those in home room 4 (which included the infectious student). The very low contact density values for students in home room 8, compared with corresponding values for students in home room 4, show the relatively lower weight given to contacts in the

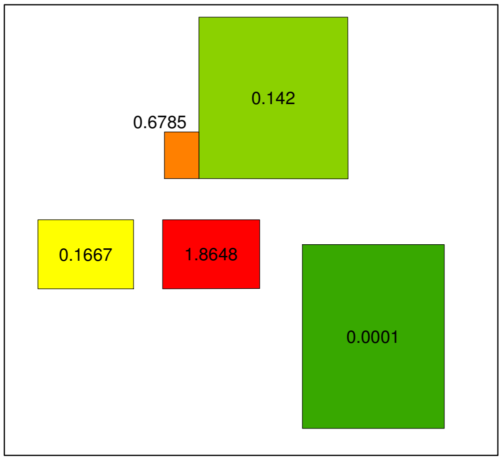


FIGURE 4.17. Locations with Corresponding Contact Densities with Infectious Student - Using Lunch Schedule Assuming One Serving Line

cafeteria. This is due to the cafeteria's larger volume. Further, the significance of contacts in the significantly smaller volume restroom can also be seen on this graph.

4.2.4.6. Analysis by Location of Total Contact Density with Infectious Individual

As shown in Figures 4.17 and 4.18, the total density of contacts is identical over both schedules in the playground, in the infectious student's home room, and in the Music Room. The total density appears identical in the playground for both schedules due to the height of 10,000 feet used when computing the volumes of each location. This increased volume made contacts that occurred in the playground nearly insignificant when analyzing using densities. The identical densities in the other two rooms result from the only difference between the two schedules being the lunch rotation.

The total contact density is nearly doubled in the schedule assuming two serving lines. This is due to the same factors that increased the total contact duration for this room. However, even with increased number of students in the cafeteria at once, the total contact density observed there is far below that of the infectious student's home room. It is also below the total contact density observed in the rest room However, in the case of the restroom, the elevated density is likely due to smaller volume than more frequent or longer length contacts.

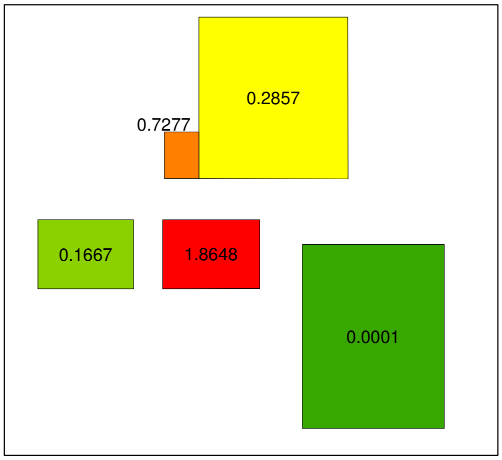


Figure 4.18. Locations with Corresponding Contact Densities with Infectious Student - Using Lunch Schedule Assuming Two Serving Lines

4.2.5. Analysis when Infectious Individual Is Unknown

Analyzing contacts when infectious individuals are unknown can provide useful information to school administrators and public health officials. The two schedules using different lunch rotations were used to generate these contacts. The schedule assuming one serving line generated a total of 186,198 contacts, and the schedule assuming two serving lines generated a total of 218,426 contacts.

The entire set of contacts for the school was also analyzed according to location. The data were manipulated using the ArcInfo software and then joined to a map of the school for visualization purposes. In the following analysis, only locations where contacts occurred are represented in the figures.

4.2.5.1. Analysis by Individual of Number of Contacts

The total number of contacts are greatly affected by the lunch rotation. In the schedule assuming one serving line, Figure 4.19 shows that the number of contacts tends to be higher for students in home rooms 16, 15, 17, 18, and 12. This is due to their positions in the course of the lunch rotation. Home rooms of students are sent to the cafeteria one at a time in five minute intervals. Each home room has the same length lunch time of 30 minutes. Therefore, both the first five and last five home rooms to be sent to lunch have fewer number

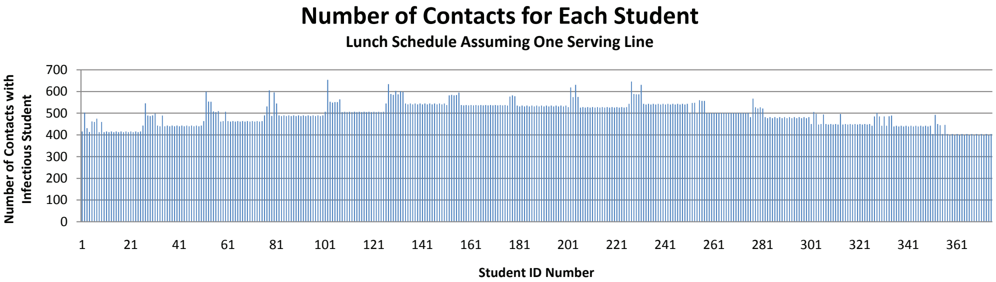


Figure 4.19. Number of Contacts for Each Student - Using Lunch Schedule Assuming One Serving Line

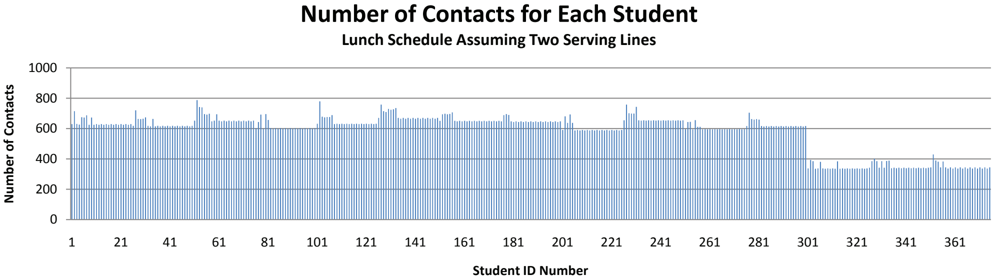


Figure 4.20. Number of Contacts for Each Student - Using Lunch Schedule Assuming Two Serving Lines

of contacts in the cafeteria. In fact, students in home rooms within the first or last five home rooms in this rotation tend to have lower contacts the closer they are to the beginning or end, respectively, of the rotation.

As shown in Figure 4.20, most students total number of contacts are significantly higher than in the schedule assuming only a single serving line. A greater number of students occupy the cafeteria during most of the lunch rotation in the schedule assuming two serving lines causing a significantly larger number of contacts. However, the last three home rooms to be sent to lunch are sent only after the cafeteria has been cleared of other students. Therefore, there are fewer students in the cafeteria at the same time when the last three home rooms are at lunch and recess. This results in significantly reduced numbers of contacts for students in these three home rooms.

Spikes in number of contacts at seemingly regular intervals are due to students who obtained hall passes and used the restroom outside of the normal schedule. For the sake of

FIGURE 4.21. Number of Contacts by Location - Using Lunch Schedule Assuming One Serving Line

simplicity, when the set of hall pass logs were created, students in each home room with the lowest ID numbers were generally the ones who obtained and used hall passes to use the restroom. Additionally, each home room contains 25 students. Therefore, the spikes seen in the graphs for both lunch rotations are spaced approximately 25 records apart, marking the beginning of each home room's assigned students.

4.2.5.2. Analysis by Location of Number of Contacts

The largest number of contacts in each schedule occurred in the cafeteria, as shown in Figures 4.21 and 4.21. This is due to the large number of students in the cafeteria at the same time. Even though students share much more time in the cafeteria in the schedule assuming two serving lines, the total number of contacts is only elevated by 14%.

The number of contacts across the home rooms are fairly consistent due to the similarity of the home rooms in composition and the similarity of the two different schedules. Total number of contacts are only significantly increased in locations affected by the difference between the two lunch rotations. Therefore, the total number of contacts in all of the bathrooms and both of the playgrounds are higher in the schedule assuming two serving lines.

FIGURE 4.22. Number of Contacts by Location - Using Lunch Schedule Assuming Two Serving Lines

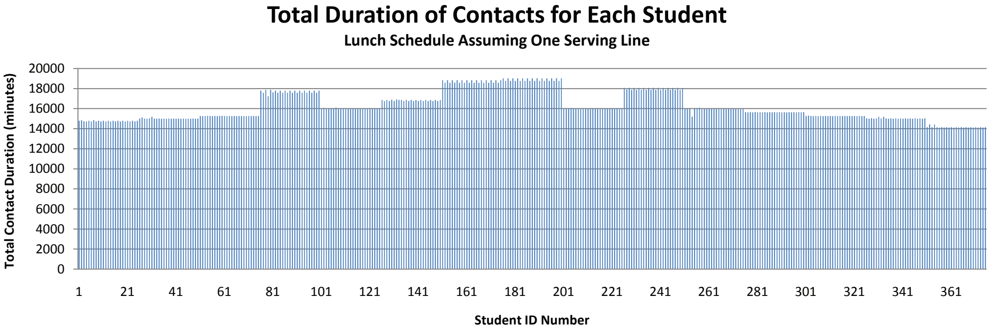


FIGURE 4.23. Total Duration of Contacts for Each Student - Using Lunch Schedule Assuming One Serving Line

4.2.5.3. Analysis by Individual of Total Contact Duration

The total duration of each student's contacts are not as sensitive to events such as the use of hall passes to visit the restroom. Instead, they are far more sensitive to changes in the lunch schedule. While the graphs of total contact durations for the each lunch schedule shown in Figures 4.23 and 4.24 may have similar shapes, the total contact durations are much higher in the lunch schedule assuming two serving lines. This is due to the higher number of students occupying the cafeteria at the same time.

The effect of isolating the last three home rooms sent to lunch while assuming two serving lines can be seen on the total duration graph. The total duration of contacts for both lunch

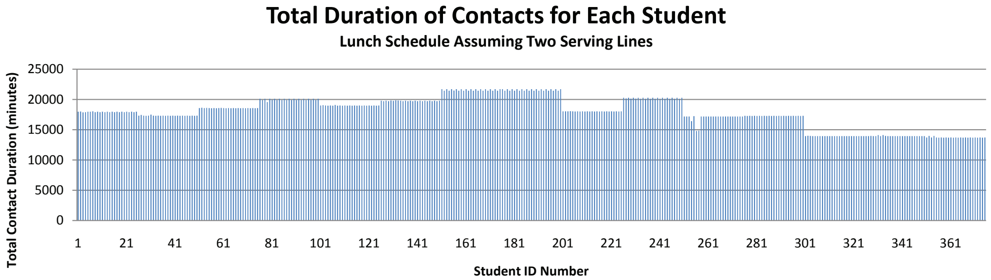


FIGURE 4.24. Total Duration of Contacts for Each Student - Using Lunch Schedule Assuming Two Serving Lines

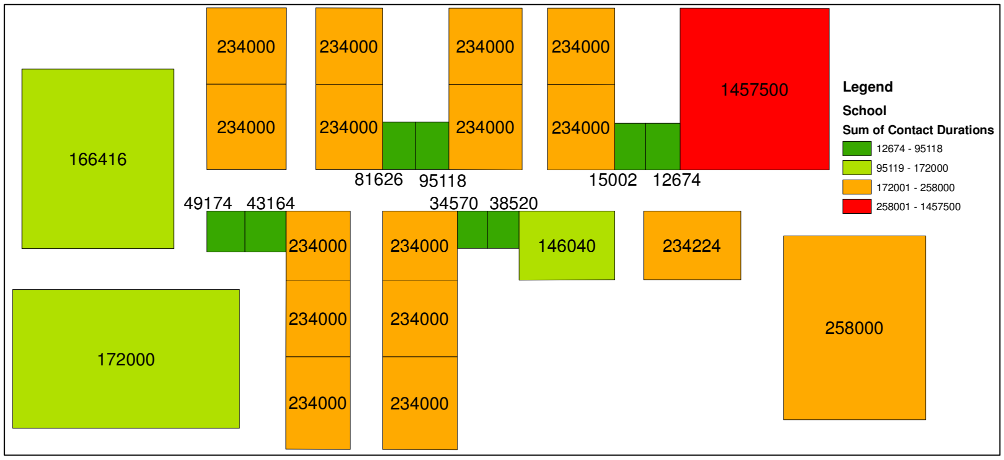


FIGURE 4.25. Total Duration of Contacts by Location - Using Lunch Schedule Assuming One Serving Line

schedules are very similar for these last three home rooms. This is due to the similar number of students in the cafeteria during the lunch period for these three home rooms in either lunch schedule.

4.2.5.4. Analysis by Location of Total Contact Duration

The sum of contact durations in each room is unchanged across both schedules as shown in Figures 4.25 and 4.26. Additionally, this sum is identical across every home room in the school except for room 4. The extra 224 minutes in Home Room 4 were likely caused by differences in hall passes used by students in each home room. Each room contains 25 students. While 224 minutes sounds like a lot of time, only ten additional minutes spent with each student would more than make up for this difference.

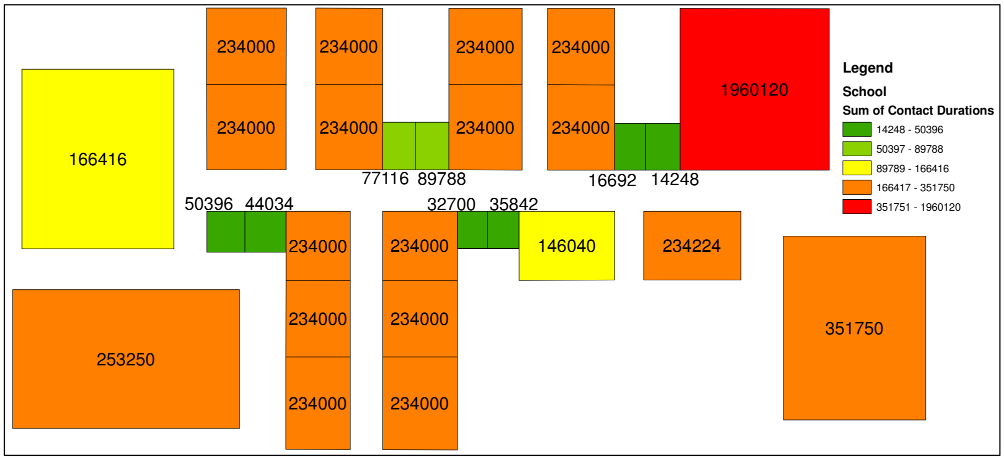


FIGURE 4.26. Total Duration of Contacts by Location - Using Lunch Schedule Assuming Two Serving Lines

The total contact duration in the cafeteria increased by over 35% when the number of serving lines was increased from one to two. This result is different from the total number of contacts because each contact in the cafeteria is now generally longer. Total contact durations also increased in other locations affected by the lunch schedule such as playgrounds and restrooms.

The average contact duration stayed consistent between the two schedules across the home rooms. However, all locations affected by the lunch rotation have more significant differences. While average contact duration increased in the cafeteria and playgrounds with the addition of a second serving line, it significantly decreased in the restrooms.

The increased average contact duration in the cafeteria is due to similar numbers of contacts being made by each student, but longer time shared in the same room for the schedule assuming two serving lines. Increases in the playgrounds are very small due to the rotation between the East and West playgrounds that exists in both schedules. Decreased average contact durations in the restrooms are likely due to additional contacts during scheduled ten-minute restroom breaks decreasing the average.

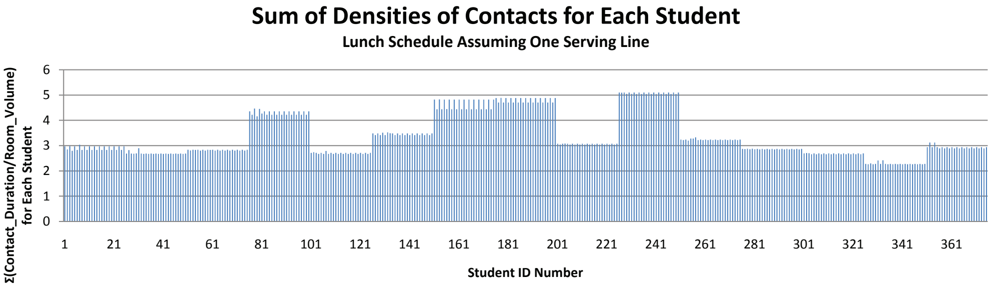


FIGURE 4.27. Sum of Density of Contacts for Each Student - Using Lunch Schedule Assuming One Serving Line

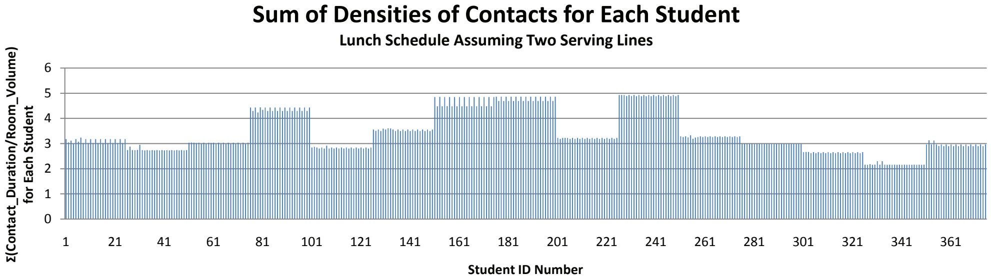


FIGURE 4.28. Sum of Density of Contacts for Each Student - Using Lunch Schedule Assuming Two Serving Lines

4.2.5.5. Analysis by Individual of Sum of Contact Densities

The total contact density for the lunch schedule assuming one serving line is greatly affected by the lunch schedule. Patterns of variations in the different sums of contact density by home room shown in Figure 4.27 are obvious. However, these patterns also reflect the volume of each home room, since no two rooms in the school have the same volume. Much of the contact time occurs in the home rooms, themselves. Therefore, students in home rooms which are smaller in volume are likely to experience higher sums of contact densities.

Additionally, in many home rooms, a difference between the sum of contact densities can be observed between male and female students. Again, this is due to the differences in volumes of the restrooms. There are four pairs of restrooms in the school. Each restroom has a different volume. Larger differences in volume between corresponding boys' and girls' restrooms in each pair yield larger differences between male and female students' sum of contact densities.

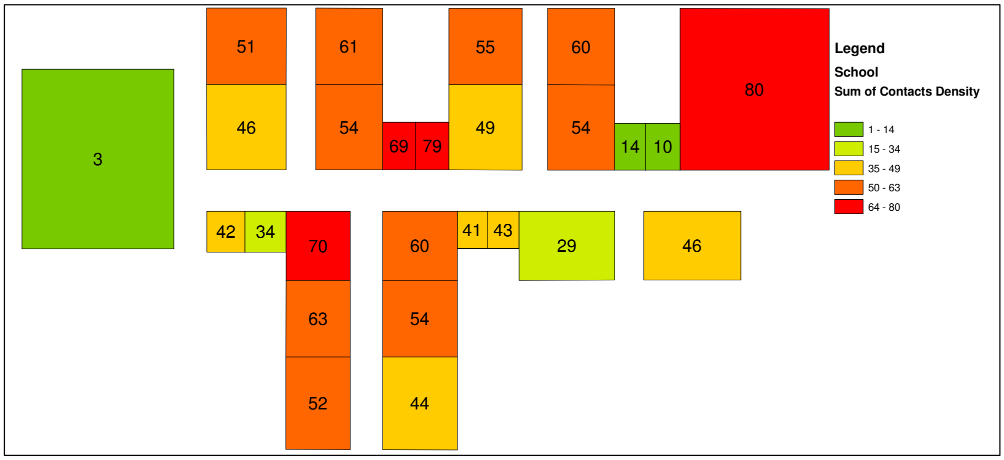


FIGURE 4.29. Sum of Denisty of Contacts by Location - Using Lunch Schedule Assuming One Serving Line

4.2.5.6. Analysis by Location of Sum of Contact Densities

In Figures 4.29 and 4.30 for both schedules, the highest sum of contact densities occurs in the cafeteria. This is due to the large number of students who occupy the cafeteria at the same time. Even though the volume of the cafeteria is much larger than the volume of all home rooms, classrooms, and restrooms in the school, it has the highest sum of contact densities because it is the location in which most of the mixing of students occurs.

The sum of the contact densities in the Gymnasium is very low due to the relatively low number of students in it at any given time and the large volume of the room. The sum of the contact densities in the restrooms increase with the addition of the second serving line in the cafeteria. This is due to their role in the lunch schedule rotation.

Each of the home rooms have identical sums of contact densities in both schedules. This is due to the only difference between the schedules being the lunch schedule rotation. Therefore, the class schedules remained largely unaffected.

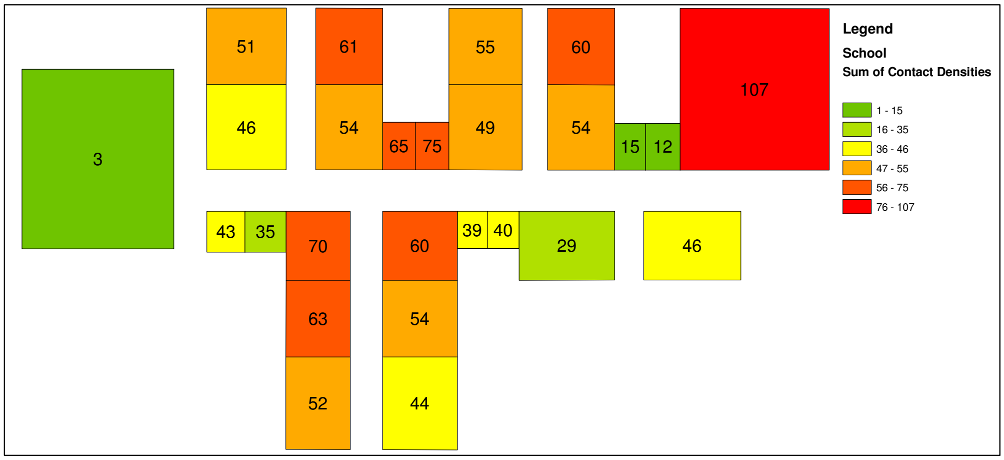


Figure 4.30. Sum of Density of Contacts by Location - Using Lunch Schedule Assuming Two Serving Lines

CHAPTER 5

SUMMARY AND FUTURE WORK

5.1. Summary

In this thesis, a deterministic, agent-based approach was used to track contacts within an elementary school setting. While these synthetic data and analyses only examined a single day, in settings such as elementary schools, schedules often repeat every day of the week. If different schedule and hall pass log data are available for weeks or months, the same tools and techniques may be used. The contacts being tracked are used to analyze exposure risk to chickenpox both among the student population, and in the environment. Many techniques used in this work had already been applied to large (on the order of 100 million) populations where little was actually known about the subjects being modeled. However, the techniques described in this thesis focused on the application to relatively smaller populations where much is known about the individuals, their movements, and the environment in which they are located.

The analyses of the synthetic dataset yielded results that were not always in line with what the author had predicted *a priori*. The hundreds of thousands of contacts tracked by the tools and then analyzed using an assortment of techniques yielded interesting results suggesting strong links between relative exposure risk, schedule, and environmental factors. If real data were used, the analysis may prove valuable to public health officials and school administrators wishing to contain an infectious disease such as chickenpox.

The analyses with a known infectious individual was interesting from a variety of perspectives. These techniques could be used retrospectively to determine, in the case of a diagnosis of chickenpox within the student body of an elementary school, who else is at risk of already having contracted the disease. Students determined to be at higher risk may then be isolated from the rest of the student population through adjustments in their schedules or in which rooms they visit on a daily basis. Proper use of these techniques may keep a potential outbreak contained to a relatively lower number of students. These analyses also resulted in the identifying of locations in the school at which students where at higher risk of exposure to the disease. Furthermore, these locations, combined with temporal data concerning the infectious individual's movements, elucidated patterns where students may be at higher levels of exposure risk. What may be even more important were the locations identified as being relatively much lower risk, where the infectious student had not occupied at all.

These same analysis techniques may be used from a different perspective. Instead of assuming the agent (the student with ID number 1) whose contacts are being tracked as infectious, the agent could be considered susceptible. In this case, the number, location, and other attributes of this agent's contacts could be analyzed to determine its generalized airborne exposure risk. This may be useful when developing schedules for immune-suppressed children in public or private elementary schools. Higher-risk types of contacts, whether they be high risk due to contact location, duration, or other dynamics, may be identified by public health officials, and then the tools and techniques presented in this thesis may be used to customize a schedule minimizing any identified types of higher-risk contacts.

The analysis without a known infectious individual (which resulted in numbers of contacts of the order of 100,000) allowed for a more detailed analysis of interactions among the entire student population. Areas where the most contacts would be expected to take place were identified using a variety of specific techniques. The interplay between the spatial locations and the schedule of the students was elucidated through the analysis of contacts resulting from the use of two strategically-different schedules. While adjustments in the schedule could serve to minimize contact durations between students, modifications or augmentations to the school building could also be used both to minimize contact durations and contact densities. These factors may be analyzed by public health officials in an effort to prevent outbreaks of many different diseases from occurring withing the school.

Using the synthetic data, the cafeteria was found to be one of the most significant locations in the school affecting exposure risk. The risk during lunch could be minimized using a variety of techniques. Separate cafeteria rooms may be used to reduce unnecessary mixing between different grades of students. Ultraviolet lights, disinfectants, or better ventilation within the cafeteria, as implemented by public health officials, may also reduce actual exposure risk in the cafeteria. Finally, simply rearranging the lunch rotation schedule to further separate grades (or other groups) of students may be the easiest and least expensive approach to minimizing exposure risk during lunch in the cafeteria.

5.2. Future Work

The techniques presented in this thesis may be extended in a variety of ways. Different methods of tracking exposures, aggregating exposures, and relating exposures to the location in which they occurred may all be explored. While the focus in the literature tends to be towards building simulations of thousands, millions, or hundreds of millions of people, more detailed analysis and modeling of smaller, more specific populations is still yet to be explored.

Additional work can be done in modifying the dynamics of individual exposures as tracked using these tools. Properties of the individual rooms in the school may be added to implement such features as opening windows or increasing the volume of air exchanged per time unit. Further, location within a room may also be used to influence the weight of edges in the graph resulting from students co-located in rooms inside the school. The tools and techniques may also be modified to track exposure risk to other infectious diseases which spread through fomite transmission.

In many epidemiological simulations, probability of disease transmission is often computed, and then a random experiment is conducted in order to force the results to a state of either *transmission successful* or *transmission unsuccessful.* However, the outcome of each random experiment closer to the beginning of the simulation will likely change the course, and therefore the outcome, of the entire simulation. This is somewhat similar to opening Schrodinger's box to determine the status of his cat.(11) While the box remains unopened, the cat may be thought to functionally exist in both states *alive* and *dead* simultaneously. It is only by opening the box and observing the cat that the result of the random experiment is forced. Instead, the simulated universe may be allowed to branch, thus passing probabilities of second-hand infections based on probabilities computed at the time of each contact. Agents in such a simulation may exist in both the *Infectious* and *Susceptible* states simultaneously. At the end of the simulation, total probability of infection may be computed for each individual, effectively placing a blanket of risk over the entire population. While, for even relatively small populations and shorter lengths of simulated time, this would be very expensive both computationally and in the amount of data stored, the results may prove to be the best estimates of risk of acquiring the disease. The results may then be used to judge the validity and efficacy of other, less expensive, approximation techniques.

BIBLIOGRAPHY

[1] Environmental Systems Research Inc., "ESRI Shapefile Technical Description," *Environmental Systems Research Inc.* 1998. [Online]. Available: http://www.esri.com/library/whitepapers/pdfs/shapefile.pdf. [Accessed August 3, 2009].

[2] Division of Healthcare Quality Promotion, National Center for Preparedness, Detection, and Control of Infectious Diseases, "Overview of Varicella (Chickenpox) in Healthcare Settings," *Division of Healthcare Quality Promotion, National Center for Preparedness, Detection, and Control of Infectious Diseases*, 2005. [Online]. Available: http://www.cdc.gov/ncidod/dhqp/id_chickenpox.html. [Accessed August 3, 2009].

[3] National Center for Immunization and Respiratory Diseases, "Varicella Disease Questions & Answers," *National Center for Immunization and Respiratory Diseases*, 2007. [Online]. Available: http://www.cdc.gov/vaccines/vpd-vac/varicella/dis-faqs-gen.htm. [Accessed August 3, 2009].

[4] C. L. Barrett, K. R. Bisset, S. G. Eubank, X. Feng, and M. V. Marathe, "Episimdemics: An Efficient Algorithm for Simulating the Spread of Infectious Disease over Large Realistic Social Networks," presented at 2008 ACM IEEE Conference on Supercomputing, Austin, TX, USA, November 2008.

[5] Centers for Disease Control and Prevention, *Epidemiology and Prevention of Vaccine-Preventable Diseases.* Washington, DC: Public Health Foundation, 2008.

[6] Centers for Disease Control and Prevention, "Chickenpox Vaccine: What You Need to Know," *Centers for Diseases Control and Prevention*, 2008. [Online]. Available: http://www.cdc.gov/vaccines/pubs/vis/downloads/vis-varicella.pdf. [Accessed August 3, 2009].

[7] J. Chin, *Control of Communicable Diseases Manual.* 17th Edition, American Public Health Association, 2000.

[8] S. Eubank, "Scalable, Efficient Epidemiological Simulation," presented at 2002 ACM Symposium on Applied Computing, Madrid, Spain, 2002.

[9] S. Eubank, H. Guclu, M. V. Marathe, A. Srinivasan, Z. Toroczkai, and N. Wang, "Modeling Disease Outbreaks in Realistic Urban Social Networks," *Nature*, vol. 429, no. 6988, pp. 180–184, 2004.

[10] S. Eubank, V. S. A. Kumar, M. V. Marathe, A. Srinivasan, and N. Wang, "Structural and Algorithmic Aspects of Massive Social Networks," presented at 2004 ACM-SIAM Symposium on Discrete Algorithms, Philadelphia, PA, USA, 2004.

[11] J. R. Gribbin, *In Search of Schrodinger's Cat: Quantum Physics and Reality.* Toronto, Canada: Bantam Books, 1984.

[12] S. Hambleton and A. Arvin, "Chickenpox Party or Varicella Vaccine?" *Hot Topics in Infection and Immunity in Children*, vol. 568, pp. 11–24, 2005.

[13] R. M. Merrill and T. C. Timmerick, *Introduction to Epidemiology*, 4th Edition. Jones and Bartlett Publishers, 2006.

[14] A. R. Mikler, S. Venkatachalam, and K. Abbas, "Modeling Infectious Diseases Using Global Stochastic Cellular Automata," *Journal of Biological Systems*, vol. 13, no. 4, pp. 421–439, 2005.

[15] K. E. Nelson and C. M. Williams, *Infectious Disease Epidemiology: Theory and Practice*, 2nd Edition. Jones and Bartlett Publishers, 2007.

[16] W. A. Oleckno, *Epidemiology: Concepts and Methods.* Waveland Press, Inc., 2008.

[17] J. Rice, *Medical Terminology with Human Anatomy.* Pearson Prentice Hall, 2004.

[18] M. Weiss, P. D. Weiss, D. E. Weiss, and J. B. Weiss, "Disrupting the Transmission of Influenza A: Face Masks and Ultraviolet Light as Control Measures," *American Journal of Public Health*, vol. 97, pp. S32–S37, 2007.