ECONOMICS 620

Allegheny College
Meadville, Pennsylvania
16335

The Cost of Community: Do Neighborhoods Affect Disease-Causing Socialization?



Mackenzie Bluedorn
April 13, 2015

THE COST OF COMMUNITY: DO NEIGHBORHOODS AFFECT DISEASE-CAUSING SOCIALIZATION?

by

Mackenzie Bluedorn

Submitted to The Department of Economics

Project Advisor: Russell Ormiston

Second Reader: Asuman Baskan

Date: April 13, 2015

I hereby recognize and pledge to fulfill my responsibilities as defined in the Honor Code and to maintain the integrity of both myself and the College as a whole.

Mackenzie Bluedorn

Acknowledgements

First and foremost, I want to extend my eternal gratitude to my parents, for providing me with unconditional support. You have unfailingly encouraged me to pursue my passions and challenge myself, both academically and otherwise. I also want to thank the Allegheny College Economics Department, where I discovered my interests and found my home here at Allegheny. Specifically, I am extremely grateful to my first reader, Russell Ormiston, for having unending patience and insight throughout this entire process; despite every obstacle and setback, you remained encouraging and optimistic. And to my second reader and advisor, Asuman Baskan, I want to say thank you for sparking my interest in economics my very first semester here at Allegheny; it was your class that engaged me and ultimately convinced me to become an economics major. Thank you guiding me the past four years.

Table of Contents

Page
List of Tables iii
List of Figures iv
Abstract v

Chapter One: Introduction 1

Chapter Two: Literature Review 4
2.1 Introduction to Previous Literature 4
2.2 Influenza and Neighborhood Choice 4
2.3 Epidemic Models and Social Networking 8
2.4 Application of Disease Transmission and Spatial and Social Networks 10

Chapter Three: Theory 16
3.1 Introduction to Theory 16
3.2 Neighborhood Choice 16
3.3 Social Network Analysis 23
3.4 Conclusion 26

Chapter Four: Empirical Results 27
4.1 Pennsylvania County level Data 27
4.1.1 Description of Data Sources 27
4.1.2 Summary of Selected Variables 28
4.1.3 Empirical Model 31
4.1.4 Results 32
4.2 HCUP Nationwide Inpatient Sample 37
4.2.1 Description of Data Sources 37
4.2.2 Summary of Selected Variables 38
4.2.3 Empirical Model 41
4.2.4 Results 42

Chapter Five: Conclusion 46

Bibliography 51

List of Tables

Page
Table 4.1. Summary of County Data 30

Table 4.2. Regression Results, County Level 33

Table 4.3. Summary of HCUP Data 40

Table 4.4. Regressions Results, HCUP 42

List of Figures

Page
Figure 3.1. Segregated Neighborhood 18

Figure 3.2. Integrated Neighborhood.....20

Figure 3.3. Mixed Neighborhood.....21

Figure 3.4. Social Network Matrices 24

Abstract

This project explores a possible relationship between location and flu transmission using socialization as a vehicle. Under the supposition that individuals will interact more frequently with others who share a similar desirable quality, such as race or income, those people who live closely in homogenous areas will have increased daily socialization due to this "like me" preference. As influenza is a communicable disease, increased socialization will also mean higher flu rates. Thus, a variety and control and experimental social variables became the subject of analysis for this project. Ultimately, the results were nebulous and overall inconclusive, though they did provide some potentially interesting insight on the linkages between locational homogeneity, socialization, and flu infection.

Chapter One: Introduction

Socialization is a fundamental aspect of human living. Human beings are naturally attracted to other humans because they crave social contact. Man was not created to be a solitary creature, and, as such, people automatically gravitate toward those with whom they can connect. People are drawn to others who share similar beliefs, ideas, or lifestyles. One's neighbors, friends, and coworkers become a family away from home and ultimately construct a major part of an individual's social network scheme. These social networks establish the bases of community and civilization. They connect people through all manner of interaction, from personal to professional.

Unfortunately, there are some negative consequences associated with socialization. One of the major disadvantages of social communication is the spread of disease; airborne illnesses spread rapidly through interpersonal contact. Influenza, or more colloquially simply the flu, is an infamous airborne communicable disease every year. Despite vaccination, new strains arise annually and plague the population. The more social interactions an individual involves himself or herself in, the greater the spread of the disease. An infected individual can potentially infect anyone and everyone he or she comes into contact with through proximity and shared surfaces. As such, disease transmission and socialization are fundamentally connected.

However, there is another major topic that must be considered in this relationship. Location, specifically at the neighborhood level, may have important implications for disease transmission. Many different factors influence location decisions, ranging from availability of public services to neighbors' traits to monetary feasibility. People subsequently divide themselves into different areas based on the preferences of individual

households. Every household may have its own prioritized characteristic; for instance, one home may dominantly value a good education system while another values social cohesion in a community.

This project will synthesize these major topics, socialization, disease transmission, and location decisions, into a single study. By building off the supposed relationship between the three subjects, this project seeks to answer how neighborhood composition affects the sociability of households, which in turn may have consequences for flu transmission. In particular, an exploration of urban economic theory, namely neighborhood choice theory, will provide the foundational support for linking location with social interaction and its consequences. While there are many factors that influence how households choose neighborhoods, this project will focus solely on social variables. While locating, some households will prioritize community and social connections with neighbors. It is imperative to note that this project focuses on social desirability. For other aspects of location decision, such as choosing a residence based on public services, homogeneity versus heterogeneity may be of no consequence.

Moving into this project, there are several basic assumptions necessary to establish before proceeding any further. First and foremost, the assumption of a default "like me" preference is important to make. In other words, when assessing location decisions based on social preferences, people naturally desire to surround themselves with others who share a similar valued quality. For example, a household that greatly prioritizes religion may express a preference for living with others of the same religion. Alternatively, another college-educated household may show a preference for neighbors of the same education bracket.

Additionally, there also exists the underlying assumption that preferences do, in fact, affect socialization. Essentially, this project will test if people are socializing more when living according to their assumed neighbor preference; if this is true, then homogenous communities, which are predicted to be the most desirable, should have a greater prevalence of influenza. Alternatively, if a household is in a heterogeneous neighborhood, surrounded by neighbors who do not share a prioritized quality, that household may withdraw from the community and not create as many social connections within the community.

Chapter Two: Literature Review

2.1 Introduction to Previous Literature

By pulling together previous studies on both disease transmission and the implications of neighborhood choice, this section creates a cohesive overview of the topic to be explored in this project. First and foremost, this section will discuss the fundamentals of both influenza incidence and locational decisions. This project merges these topics; location may prove to have certain implications on socialization, which implies more or less potential for disease-spreading interactions. Finally, previous literature provides a thorough review of theory on this subject as well as empirical application.

2.2 Influenza and Neighborhood Choice

In order to efficiently approach the proposed relationship between neighborhood choice and disease-transmitting interactions, it is important to first break down and individually review the topics of interest. While disease and illness come in infinite diversities, this study focuses solely on communicable diseases. Specifically, influenza transfers within a population based on interactions with infected individuals, as established in most literature regarding disease transmission. By utilizing the known communicability of the flu, researchers can then build epidemic models designed around interaction (Bansal, Pourbohloul, Hupert, Grenfell, and Meyers, 2010). As with any other communicable disease, the likelihood of contracting the flu stems from the number and intensity of one's interactions with an infected individual. Therefore, groups with

more contacts experience greater susceptibility to flu infection. Bansal et al. (2010) used this knowledge of communicable disease to address the overarching question of how interactions and disease transmission differ demographically. By focusing specifically on age blocks, their study addressed how disease-causing interactions change across these groups. School-age children typically have the most contagious interactions with each other, therefore having the highest flu incidence in an initial pandemic. Influenza is a partially immunizing disease, meaning that following the initial contraction of a certain strain, an individual will retain a partial, temporary immunity to that strain. Thus, while children are the most susceptible during an outbreak of a new flu strain, they will subsequently be the most protected during a reoccurrence of that strain.

This study remains highly relevant; it establishes not only the infectious nature of the flu and the necessity of interaction for infection, but also explores how transmissibility can change with regard to demographics. Furthermore, especially during later research on this project, it raises the consideration about how the presence of children in households may promote greater disease transmission in a neighborhood. The importance of demographics and socioeconomic status on flu transmission will come to light after further study on neighborhood choice.

With the aspects of the flu related to this project now defined, a review of neighborhood choice is next. Naturally, many studies on neighborhood choice focus on divisions by race. The influx of immigrants into the United States provides a way to assess how different races, namely minorities, make "residential choices" (Clark and Ledwith, 2007). Clark and Ledwith's (2007) article addresses both homogenous and mixed neighborhoods. The authors wanted to analyze the influences of socioeconomic

status on how households move; ultimately, does higher income lead Hispanics to stay in same-race neighborhoods or push them to enter mostly-white neighborhoods, and similarly, will whites continue to move only within white neighborhoods?

To begin addressing this question, the authors divided neighborhoods into three groups: minority, mixed, and majority. Minority neighborhoods indicated that the race of the moving household accounted for less than 40% of that neighborhood's composition while majority neighborhoods were comprised of more than 60% of the moving household's own race; mixed neighborhoods fell between these percentages. Studies show that income plays a pivotal role in determining neighborhood choice. On average, white movers who earn above-median incomes remain within majority white neighborhoods while those earning below-median incomes end up in minority white or mixed neighborhoods, thus expressing a general trend of white households with white preference (Clark and Ledwith, 2007). More so even than income, education proves to be a divisive factor. While some discrepancies between races exist when looking at decisions by income, higher education overwhelmingly correlates with moving into predominantly white neighborhoods for both whites and Hispanics. It is important to note the assumed correlation between predominantly white neighborhoods and high status living. Clark and Ledwith (2007) indicate this may be because "the current structure of the urban mosaic reflects historical patterns in which white households had disproportionate access to higher status neighborhoods" and this pattern may be subject to change in the future.

Clark and Ledwith's (2007) study provides evidence for neighborhood choice theory, which will be discussed at length in the following chapter, and how this theory

applies differently across races. Furthermore, this study highlights how individuals express preferences for location based on the makeup of potential neighbors as well as how these partialities balance against socioeconomic means to obtain such preferences. Essentially, although preferences play a major role in location decisions, they are managed by affordability. Thus there exists an inherent connection between spatial residence and social preference. As social connectivity is of special importance to this project, it will be interesting to explore if residents constrained by income to a neighborhood contradictory to their preferences exhibit limited interaction within the community. Additionally, Clark and Ledwith's (2007) research on neighborhood choice raises some unique questions to consider throughout the rest of this project; for example, how are interactions affected due to local homogeneity in one factor, such as income or education, but heterogeneity in another, such as race?

Previous literature also raises many important considerations for examining the relationship between sociability and location. While looking at neighborhood choice and its role in this project, it is also important to define a difference between social interaction and spatiality. As mentioned above, there is a correlation between where people locate and with whom they socialize. Emch, Root, Giebultowicz, Ali, Perez-Heydrich, and Yunus (2012) suggested "social clustering might be due to spatial clustering; that is, individuals who are socially connected are more likely to live close to one another…" (p. 1010), raising concerns about causality. Will groups tend to cluster because of previously established social ties or will social ties arise due to spatial clustering? As this project develops, the issue of causality will be of interest when exploring neighborhood choice and, eventually, its influence on disease spread.

Together, disease transmission and the interactions generated by location decisions form the foundation of this project. To build from this foundation, a theoretical approach in the following paper begins to synthesize these two topics.

2.3 Epidemic Models and Social Networking

To effectively review theories on this subject, Keeling and Eames (2005) separately addressed disease epidemic models and networks before melding the two. Basic infection models include subjects grouped into three categories: susceptible, infected, and recovered. Based on these three groups, two models are derived. The first, Susceptible-Infected-Recovered (SIR), accounts for diseases that grant lifelong immunity after they have been contracted. The second, Susceptible-Infected-Susceptible (SIS), pertains to diseases that are common for repeat infections, such as sexually transmitted diseases. However, most studies use these models in a simulation of random mixing. In a random mixing scenario, populations may be divided into subgroups, such as by age, but it is still assumed that people within those groups mix with complete randomness. This model, unfortunately, does not account for the reality that interactions are typically fixed within certain social networks and not random, which Keeling and Eames (2005) sought to remedy. Turning to these social networks, the vernacular across the literature differs greatly, though the theory remains the same. These networks are often used to deduce the nature of interactions or even the social prominence of an individual. When beginning to pair social networks against disease transmission, many difficulties can occur. For example, exactly how much or how little contact is actually needed to pass on flu infection? Furthermore, every disease has individual qualifications for infection,

meaning that models must be disease-specific and are not directly applicable across multiple diseases (Keeling and Eames, 2005).

Three major techniques have been devised in order to gather information on networks with respect to disease. The first, infection tracing, utilizes a backward approach. Each infected subject is traced back to the individual from whom they contracted the disease as well as connected with others to whom they then passed on infection. This method gives a simplistic overview of network connections, but still provides some understanding to disease transmission. Second, contact tracing attempts to track all of a subject's potentially disease-causing interactions, creating an intricate matrix of network activity. Unfortunately, this method is especially susceptible to subject error and reliability. Third, diary-based studies are much less time-consuming and work-intensive for researchers because they consist of subjects logging and tracking their own interactions as they occur. However, this technique once again suffers from a subject's capability and may therefore not be completely reliable or uniform across subjects. Data collection for disease transmission along social networks is fraught with difficulties and is constantly subject to limited resources. Due to this nature, computer simulated models are popular for this research (Keeling and Eames, 2005).

Finally, once data collection has been conducted, a variety of network models can be used. Each accounts for a different schematic of interactions. Notably, all of these models account for reduced spread of infection with each occurrence of infection; as one individual infects a contact, said individual then has a smaller pool of potential neighbors to contaminate. There are many different models to account for different desirable characteristics of networks. For instance, random network models disregard subjects'

proximity and location in addition to assuming homogenous contact patterns across the entire population. In contrast, lattice models function on interactions rigidly defined by spatial locations. Small-world networks fall on the spectrum between the previous two models. They focus on clustering and rapid infection throughout localized areas. These models represent three of the most popular models, though other popular choices exist, including spatial networks, scale-free networks, and exponential random graph models (Keeling and Eames, 2005).

Keeling and Eames (2005) used theories to delve into the heart of this project: "the research in graph theory and social sciences generally considers an understanding of the network itself to be the ultimate goal; in contrast, epidemiological interest is focused on the spread of the disease, in which case the network forms a constraining background to the transmission dynamics." Their discussion on these theories married the two subjects of interest, flu infection and neighborhood choice. The combination of these theories allows for empirical study, which this project as a whole will pursue.

2.4 Application of Disease Transmission and Spatial and Social Networks

Previous literature provides a surplus of material from which to create a conceptual approach. Yet other studies offer insight on application of these topics through empirical work. Published literature specifically regarding influenza is limited; however, studies conducted on other personally and environmentally communicable diseases still offer substantial information on defining the relationship between locality and infection. When beginning to analyze disease transmission, some studies explore either spatial or social networks. Emch et al. (2012), however, took the unique approach

of surveying the combined effects of social ties and spatial clustering on disease transmission. The authors used kinship as the primary connection in this study as they examine the spread of cholera and shigellosis. Familial relationships provide a promising study group because common customs within a family include frequent visiting and the sharing of meals (Emch et al., 2012). These interactions are then the perfect ground for spreading communicable disease. Limiting analysis solely to familial socialization also removes the complexities and difficulties of generating complete school and work socialization models. As mentioned previously by Keeling and Eames (2005), due to the variability of illness and disease-causing interactions, every study must make adjustments specific to the disease of interest. For example, in this study, spatial networks were also considered because cholera and shigellosis infection can be environment-related, showing the flexibility of empirical work to model various disease types. Basing their study in Bangladesh, the authors built social models for kinship across *baris*, closely grouped households related through male lineage, and spatial models for neighborhoods of varying sizes. They then analyzed cholera and shigellosis incidence across a span of 20 years, from 1983 until 2003. They conducted two tests, one that examined spatial and social effects separately and one that tested the two variables jointly.

They found through a combined social effects-spatial error model that spatial clustering returned a significant result on cases of the two diseases for every year in the 20-year time period. Spatial clustering has a profound effect on the spread of both cholera and shigellosis, as evidenced further by significant results over several neighborhood scales of different sizes. Social networks, however, only returned significant results for five and six years for cholera and shigellosis, respectively. This

result seems to imply that social networks only have minimal impact, and that influence may only exist in certain, specific situations (Emch et al., 2012). The separate testing of the variables also mirrored this same pattern of spatial significance and limited social significance.

Unlike cholera and shigellosis, influenza is an airborne illness and may have a different relationship with spatial and social networks. As discussed with neighborhood choice, there may be a tendency for similar people to group together into homogeneous clusters, either demographically or socioeconomically. Thus, demographics and socioeconomic status can have major implications for disease transmission if individuals move together or interact with each other based on these variables. To investigate further into the connection between influenza and socioeconomic status, Tam, Yousey-Hindes, and Hadley (2014) compared occurrence of flu hospitalizations in a Connecticut county against the socioeconomic status and sex of the patients. Other studies have explored the prevalence of disease across demographic groups, but few have collected information on the correlation with socioeconomic status, even though health can be greatly subject to an individual's environment and neighborhood (Tam et al., 2014). With this information, the authors hoped to draw conclusions between neighborhood socioeconomic status and hospitalization among adults as well as test if females showed a higher incidence of flu-related hospitalizations than males.

Using the Public Health Disparities Geocoding Project, Tam et al. (2014) geocoded patients over four influenza seasons to connect them with neighborhood Census Bureau data on socioeconomic factors. As expected, their results showed increasing chance of hospitalization with an increase in age. However, even after

controlling for age discrepancies, they still found higher hospitalization incidence for women. Of more relevance to this project, they found several connections between hospitalization cases and socioeconomic status. Their study shows a positive correlation between hospitalization and percent of people living below the poverty line, number of people lacking a high school diploma, and household size as well as a negative relationship with income. Ultimately, "a significant trend of increasing incidence of influenza-associated hospitalizations was found with decreasing SES status" (Tam et al., 2014). While the magnitude of this trend differed through race/ethnicity, the pattern itself remained consistent. The authors then provided several important considerations for why this trend might exist. As lower socioeconomic status is frequently associated with greater household crowding, people in these households have more exposure with other infected household members as well as greater interaction with children, who are more likely to spread influenza. Individuals of a lower socioeconomic status are also less likely to have insurance or receive other medical care, perhaps meaning less immunization against the flu (Tam et al., 2014).

While all of the previous studies have evaluated the effects of social networks and personal characteristics on disease transmission, a key part of economic research involves exploring all avenues of causation. That is, in reverse, how might influenza affect socialization and what new implications might this alternative causation have? In most disease transmission literature, interaction patterns are held constant. However, in reality, these patterns may fluctuate greatly; for example, individuals have varying contact schedules based on weekdays versus weekends, or on a holiday (Van Kerckhove,

Hens, Edmunds, & Eames, 2013). To build from this premise, Van Kerckhove et al. (2013) investigated how social patterns change as a result of illness.

In approaching this study, Van Kerckhove et al. (2013) surveyed a large group of people living in England who exhibited flu-like symptoms during the 2009 pandemic. The authors used a diary-based study by having the people of this group keep records of their interactions while sick and again weeks later after they had recuperated from the illness. As they assessed the data, comparing interaction records from ill individuals with those of healthy participants, the authors considered both the number of contacts occurring as well as how those contacts were distributed across age groups. As expected, the authors witnessed participants withdrawing from social interaction while ill with the flu, creating a significant impact on social networks due to the presence of illness. They then used these findings to analyze the implications of symptomatic versus asymptomatic flu contraction. If people begin avoiding social contact due to illness, it naturally follows that those people expressing symptoms will refrain from interactions and thus not spread as much disease; however, "symptomatic persons are more infectious per contact but make fewer contacts than asymptomatic persons" (Van Kerckhove et al., 2013). Although symptomatic people might lessen contact with others, what limited contact they do have appears to be more infectious than the unconstrained interactions of asymptomatic participants. They ultimately found that this heightened infectiousness weighs more greatly, and symptomatic people, despite pulling back from socialization, are predominantly responsible for disease transmission by being nearly three times as infectious as asymptomatic participants (Van Kerckhove et al., 2013).

All of the literature explored in this chapter lends different perspectives to

analyzing the connections between location, socialization, and disease. Together, these previous studies create a working background on the significance of disease transmission and neighborhood choice, as well as how the two may interact with each other. An individual's socioeconomic status may allow or constrain that person from locating in a preferred neighborhood; depending on such preferences, the individual may then be more or less inclined to form social networks and interact with the surrounding population. These interactions, or lack thereof, will then directly affect the spread of communicable disease, for this project influenza, within that neighborhood.

Chapter Three: Theory

3.1 Introduction to Theory

This project seeks to answer one question: How is disease transmission, by way of socialization, impacted by location decisions? Unfortunately, there is no single, fundamental theory to perfectly address this question. However, a combination of multiple economic theories can be used to approach this topic and create hypotheses about potential outcomes. The spread of communicable disease and neighborhood choice are inherently linked; it is assumed that neighborhood choice will directly and significantly affect social interaction because those who locate within their preferred demographic or socioeconomic group will be more inclined to participate communally with neighbors.

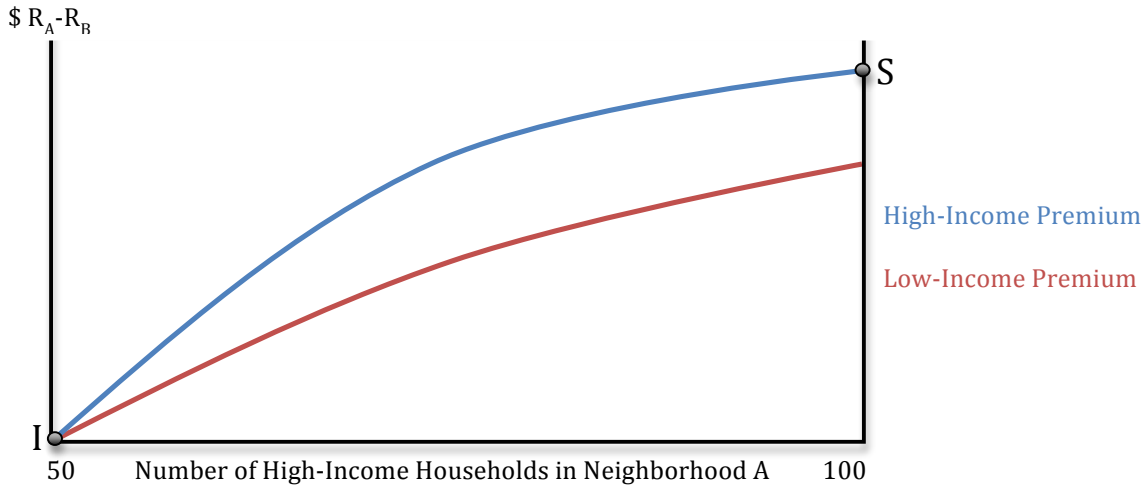
3.2 Neighborhood Choice

"Neighborhood choice" refers to the basic economic theory about how people make residential choices and what kind of neighborhoods they form. This theory analyzes residential decisions through several insights; for this project, however, the sorting of residents through preference for neighbor features, namely similarities versus differences in income, education, race, or perhaps religious affiliation, will be the primary focus. That is to say, based on a balance of neighbor characteristics and premiums, will residents congregate into homogenous or mixed neighborhoods? Although everyone wants desirable neighbors, only those people who are willing and, even more importantly, able to pay for those preferences will be able to reside in their optimal location. Neighborhood choice categorizes neighborhoods into three possibilities:

segregation, integration, and mixed. In modeling neighborhood choice, the premiums of households, a term that represents households' willingness to pay, will determine how these households separate into different communities. Theoretically, segregated neighborhoods occur when there is complete homogeneity regarding the indicator of interest, meaning that residents in these segregated neighborhoods all share the same valued quality, whether they are all of the same race, income bracket, or other desired characteristic. Conversely, integrated neighborhoods are an even blend, and mixed neighborhoods fall in between the two. To divide into these residential categories, neighbors may sort themselves based on a variety of indicators; for example, income may be a major determinant.

To build a simple model based on income preferences, there is the assumption that the most desirable neighbors are those with higher incomes. There are several other assumptions important to this model: there are only two groups being considered, high income and low income, and there are an equal number of members in both groups. This model will then assess how these two groups will divide themselves between two neighborhoods. For the purpose of this model, we can assume 100 households in each income group, for a total of 200, and two otherwise equivalent neighborhoods, A and B, that offer 100 residences. While both of these neighborhoods can be assumed to start at 50-50 integration, the households will ultimately resort themselves according to preference. Based on this information, any of the three neighborhood categories may result depending on the expression of willingness to pay. If high-income households express a willingness to pay a greater premium to live next to other high-income households, thus prioritizing homogenous living, the following division will result:

Figure 3.1. Segregated Neighborhood

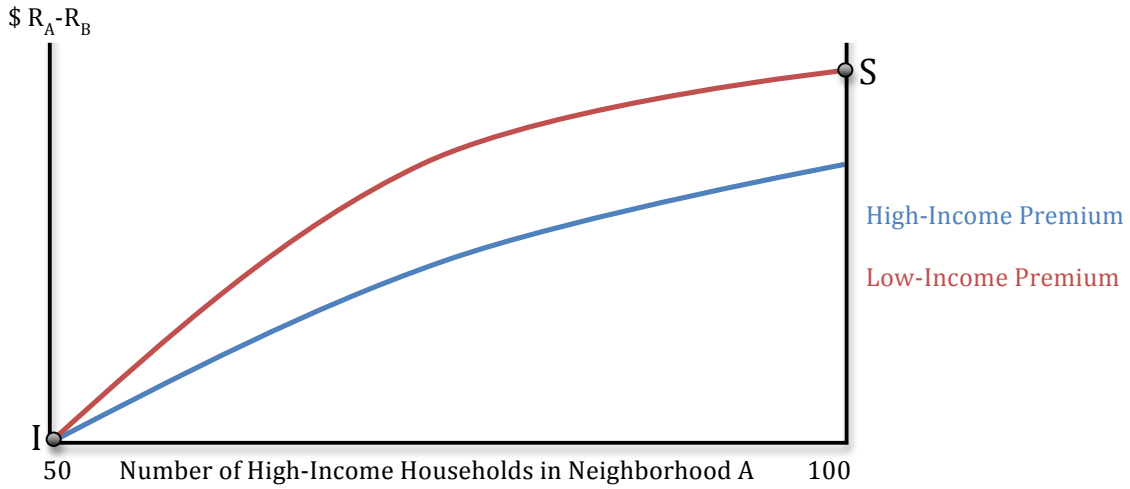


This figure examines the willingness-to-pay premiums for both the high-income and low-income households. The vertical axis represents the rent gap between Neighborhood A and Neighborhood B ($ $R_A - R_B$) while the horizontal axis measures the number of high-income households residing in the first neighborhood. In this scenario, all high-income households are willing to pay a greater premium for living with high-income neighbors. Thus, the natural equilibrium for the 200 households being considered will be segregation, marked as point S on this diagram. High-income households will continually outbid their low-income counterparts until Neighborhood A consist solely of 100 high-income residents. For every high-income household that moves into Neighborhood A, other high-income households will value this neighborhood even more, thus creating a self-reinforcing effect that results in segregation. The slope of the blue line in Figure 3.1 represents the increasing desire of high-income individuals to live in a high-income neighborhood, which substantially outdistances the red line representing

low-income willingness to pay. Naturally, if high-income is a desirable quality, households' valuation of the neighborhood will increase as the number of high-income families in the neighborhood increases. Importantly, however, the slopes of the lines increase on a decreasing scale. When there are few high-income households in a neighborhood, the addition of a single high-income resident can greatly increase the value of that area. However, if the vast majority of the neighborhood is already high-income, the addition of another high-income household is still desirable, but it does not drastically alter the evaluation of the neighborhood. By default, the low-income population will also be segregated into Neighborhood B. Both income groups express a preference in this situation to locate with neighbors of their same group.

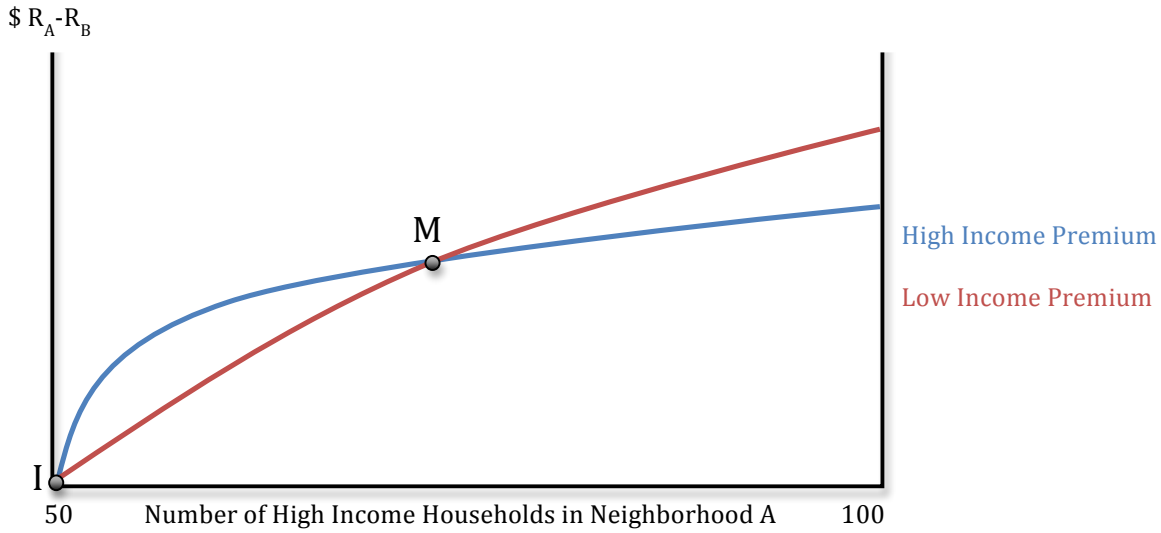
However, segregation will not always be the resulting equilibrium. In contrast to the first scenario, low-income households may actually be willing to pay a higher premium to live with high-income neighbors. Although this desire may seem illogical, there are actually many reasons such a scenario might occur. A low-income home might be willing to pay more to locate within a high-income neighborhood for access to the pubic services offered in that community. High-income neighborhoods often offer better education, job, or transportation opportunities. If low-income households are willing to pay a higher premium to live in Neighborhood A, which presumably offers these higher-end amenities due to the presence of high-income residents, than high-income households, the premiums would look as follows:

Figure 3.2. Integrated Neighborhood



Unlike in the first scenario, the low-income premium is now higher than the high-income premium. Low-income households will begin outbidding high-income households for space in Neighborhood A, leading to an integrated neighborhood where the split between the qualities of interest is perfectly even. Point I, therefore, becomes the stable equilibrium for these neighborhoods as a self-correcting effect occurs. If the neighborhood starts to move to the right along this point, by way of a new high-income household moving in, other low-income households will then place greater value on the neighborhood and pay more to move in, thus moving Neighborhood A back down to point I. Due to these effects, neighborhoods A and B will have the same makeup with both consisting of half high-income and half low-income residents and likely offer comparable public services.

Figure 3.3. Mixed Neighborhood



The third neighborhood composition occurs when the high-income and low-income premiums intersect. This intersection of premium curves may occur when households of either group do not all share the same mentality. That is to say, perhaps only a portion of high-income households actually value and are willing to pay more for locating in a predominantly high-income neighborhood. Additionally, this version of the model might also be applicable to explain a mixed-income neighborhood created via government subsidy and housing programs. In this instance, one group's premium does not always outbid the other. For example, the diagram shown above in Figure 3.3 might be created. Point M, a mixed neighborhood, will be the resulting stable equilibrium. The term "mixed" refers to a neighborhood that falls anywhere between complete segregation and perfect integration. To the left of this point, high-income households value Neighborhood A more and will pay a higher premium, but to the left of point M, low-income households will begin bidding highest for residences in this neighborhood.

The discussions above apply neighborhood choice theory to income preferences among neighbors. However, the theory is equally applicable for other desirable neighbor characteristics, including race and education. Overall, this area of neighborhood choice theory attempts to model how neighborhoods will divide themselves based on a desirable factor and residents' willingness to pay for that factor. Those households who greatly value a certain indicator will be willing to pay more to have that quality in their neighbors, so people will sort themselves based on neighbor preference. This model implies that all residents will end up in a desirable living situation based on preference. If this were true, all three neighborhood scenarios would be prime for close social networking because the rent premiums reflect households' desires to live in those neighborhoods. However, this theory does not necessarily account for the fact that a resident might be willing but unable to pay a certain premium.

The consideration of willingness against ability to pay for certain preferences is especially important when assessing indicators other than income, such as race, education, or even political preference. For example, although a resident may be willing to pay more for a home in a same-race neighborhood than another individual, he or she may simply be unable to pay that amount. Subsequently, this disgruntled resident may actually live in a neighborhood that does not conform to his or her preferences or premium. As such, this project will analyze how social networking differs amongst residents in homogenous (segregated) versus mixed (integrated or mixed) neighborhoods, based on assumptions about the most desirable neighborhood indicators. Most importantly, however, this project will assume the existence of a "like me" preference in all residents under consideration; in other words, for the duration of this project,

segregation based on a variety of indicators will be considered the default or desirable outcome. The theoretical model discussed above shows instances where a specific condition, in these cases a high-income neighborhood, was desirable to all households. It would be impractical, however, to conduct a project attempting to account for the subjective desires of each and every individual household. This assumption of an all-inclusive "like me" preference thus builds a connection between neighborhood makeup and inclination to socialize by using the foundation of neighborhood choice theory in a new way. Individuals living in a preferential setting will theoretically socialize more with each other, which then raises the consideration of communicable disease transmission. If certain neighborhoods, such as segregated communities, promote more social interaction, the negative aspects of socializing should also be more prevalent. Thus, by using neighborhood choice theory in conjunction with the "like me" assumption, specific neighborhoods can be identified as targets for high flu counts if the denizens are homogenous with respect to a valued social factor. In contrast, more diversified communities would be expected to have less flu transmission.

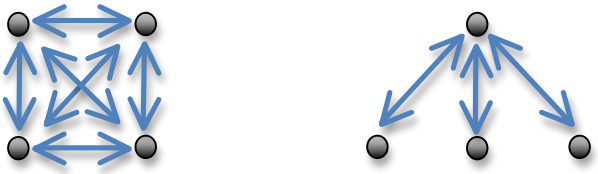
3.3 Social Network Analysis

While neighborhood choice explores how location may ultimately influence socialization, social network analysis, which is a subgroup of Graph Theory, helps better explain the correlation between socialization and lines of disease transmission. In the past, social networks have not been widely incorporated in economic studies, but they apply to a range of topics, from labor markets to corporate hierarchies. While this project will not perform an in-depth analysis of social network models, it is important to

establish the fundamentals of social network theory. These basics provide the underlying rationale for how neighborhood choice may impact social ties and, subsequently, communicable disease transmission.

Social network analysis utilizes graphs to map interactions and explores the "dyadic relation" shown therein (Bandyopadhyay, Rao, & Sinha, 2011). There are many statistical implications of network analysis in the economic realm, but the core of this project rests primarily on the basic network principle of building these matrices. Unlike most economic work, which analyzes the characteristics of a single individual such as socioeconomic or demographic indicators, social network analysis focuses on the relationship between individuals. The term "nodes" refers to these individuals of interest, while "edges" are the determined interactions connecting individuals. Using nodes to represent individuals in group interactions, diagrams can show both number and direction of such interactions. For example, one case may consist of four nodes that are all interconnected, meaning each of the four have the same ability to interact with all of the others (left on Figure 3.4); another case, however, may show a group of four where only a single node has the ability to interact with the others while the rest are cut off from each other (right on Figure 3.4).

Figure 3.4. Social Network Matrices



These two cases are only examples of the vast possibilities. Networking matrices can show total group interaction or hierarchies, symmetric or asymmetric interactions; edges

may be able to flow in all directions or only in one. These diagrams can become increasingly intricate and expanded depending on the group or community under analysis, and they can be used to represent any level of community: individuals, households, or even entire neighborhoods (Bandyopadhyay et al., 2011).

As explored in the previous literature review chapter, social networks and illness have a two-way relationship; each can affect the other. The interaction maps built in social network analysis study the flow of socialization, which also then follows lines of potential disease transmission. For example, the two simple instances shown in Figure 3.4 display how disease may be transmitted between interacting individuals. Infection and contact tracing, two methods of tracking flu transmission discussed in the previous chapter, can utilize these matrices to determine the possible pathways for disease spread. Thus, in multiple methods of disease research, social network analysis plays an important role. Whether a researcher attempts to trace a disease back to its origin or follow it forward through the origin's potentially contaminating interactions with others, he or she must understand what these avenues of interaction are. Looking back at Figure 3.4, the two matrices might experience different flu transmission paths. For example, the case on the left shows open interaction between all four individuals, meaning flu can spread easily to all four members because they are all openly and easily communicating with one another. In the instance on the right, however, disease spread would be stunted or lagged, as the disease could only spread to other individuals through contamination of the "top tier" individual. Using this basic level of network analysis can allow researchers to visualize all potential pathways of disease transmission. This networking theory helps

understand the intricacies of social interactions, which provides background and insight on both the spread of influenza and the importance of location in forming networks.

3.4 Conclusion

Urban economics, namely neighborhood choice theory, provides the theoretical background that inspired this project. Naturally, perfect application of theoretical models to reality does not exist. A completely segregated neighborhood is surely not a frequent enough occurrence in reality to truly test, but the theory can still be generalized and applied to this project. Ultimately, neighborhood choice theory speaks to the role and importance of preferences. The original theory applies to any quality that is deemed desirable across all groups, but for the purpose of this project, the application of this theory is through the use of an assumed "like me" preference. Under this assumption, neighborhood makeups can be compared where larger groups of similar people are presumably desirable, while more diversity assumedly generates less social interaction. Along with this urban model, the theory of social network analysis displays how socialization can be tracked. The matrices used in this theory represent all interactive connections within a group of individuals, which can also be interpreted as potential pathways for disease transmission. "Like me" preference would thusly make the conjecture that homogenous communities of people will have more connected matrices. These theories about the relationship between neighborhood makeups, or simply location, and disease-causing socialization will be tested on two separate levels in the following chapter.

Chapter Four: Empirical Analysis

4.1 Pennsylvania County level Analysis

4.1.1 Description of Data Sources

With coverage of previous research and a theoretical foundation now defined, this chapter delves into empirical analysis in an attempt to identify relationships between flu incidence and socialization determined by neighborhood. Ultimately, two separate datasets were collected and analyzed individually. The first created dataset includes influenza and social variables at the county level for the 67 counties of Pennsylvania, and the next four sections will discuss only the analysis of this dataset. By pulling together information from a variety of sources, data was mined for instances of flu to compare against a chosen selection of indicators believed to potentially influence socialization. In order to create the dataset, information was drawn from a number of different sources.

The compilation of this dataset primarily utilized government-run sources. First and foremost, the Pennsylvania Department of Health provided data on flu prevalence at the county level. The department accounted for all 67 counties of Pennsylvania and listed flu instances by major strain. For all variables, this project utilizes the most recent published data; flu cases are representative of the 2013/2014 flu season. Next, information for several variables on the areas, in this instance the 67 counties, was collected from the U.S. Census Bureau for 2013. The Census Bureau provided not only information on living proximity but also on identifiers such as age and income levels. These indicators are in percentage form, showing the distribution of each of the selected groups across the counties. Thus, the Census Bureau is responsible for a collection of control variables as well as variables likely linked to social interaction. Additionally,

another government source, the Pennsylvania Department of State, hosted listings for political affiliation, which have been included as potential determinants of socialization. Political data, although as recent as possible, slightly lags behind the previously described data and was last updated in 2012.

Finally, the last of the social variables came from the Association of Religious Data Archives. This source provided information on religious affiliation for the Pennsylvania counties. Notably, though, this data reflects numbers recorded in 2010, but should still be largely representative of current populations as well.

4.1.2 Summary of Selected Variables

The complete dataset is comprised of 27 variables, all of which can be viewed at the end of this section in Table 4.1. This data only accounts for flu cases officially diagnosed in hospitals, but as total numbers outside of hospitals are impossible to account for, this data provides the best possible proxy for prevalence of the flu. The data tracked the two primary strains of influenza, A (*FluA*) and B (*FluB*), as well as a third category, U (*FluU*), for other unidentified strains. These counts were then summed together for each county into a total number of flu cases (*FluTotal*).

Population data (*Pop*) helps to control for the typical spread of flu; the more people there are in any area, the greater likelihood of them transmitting communicable diseases to each other. Along with population, in an attempt to account for living closeness, the average persons per household (*HHPersons*) and the persons per square mile (*PPSM*) were also included in this data. These three variables lend themselves to determining living proximity. All communicable diseases, especially the flu, are affected

by how closely people are living with one another. For instance, higher numbers in all of these variables indicate closer living, which should correlate with higher counts of the flu.

Finally, a third grouping of variables was chosen to represent factors that may affect social patterns. Race, political, religious, and economic standing variables are included in this model in an attempt to reveal the "like me" preference mentioned in previous chapters. Three race variables account for the percentage of white (*White*), African American (*AfricanAmerican*), and Asian and other race (*AsianOther*) individuals in each county. Along with race, this dataset includes the percentages of people registered for either of the major political parties, Republican (*Repub*) and Democrat (*Dem*). The dataset also contains four religious variables, accounting for the percentages of people belonging to Evangelical Protestantism (*EvanProt*), Mainline Protestantism (*MainProt*), and Catholicism (*Catholic*). The fourth category of unclaimed (*Unclaim*) consists primarily of atheists and agnostics, but does not represent a perfect sample of no religiosity because it also includes individuals belonging to sects not prevalent enough in the state to be formally recognized as separate variables, at least as defined in this particular dataset. These variables were all originally listed as individual counts of people, but were then converted into percentages of the county population for easier use in this project. An age variable (*Under18*) was included to count the effect of minors by measuring the percentage of individuals under the age of 18. Children typically have the greatest amount of socialization with each other through schools and other social obligations, making them prime carriers of disease. Lastly, ten income variables were generated to display income distribution across the county. These variables show an

average income over the 2009-2013 time period. The variables indicate the percentage of the population falling into ten different income brackets, ranging from the lowest bracket earning less than $10,000 (*Income1*) to the highest earning group at greater than $200,000 (*Income10*).

Table 4.1. Summary of County Data

Variable	Mean	Minimum Value	Maximum Value
FluA	339.82	5	2267
FluB	70.17	1	653
FluU	10.94	1	120
FluTotal	414.49	6	2486
Pop	190653.7	4886	1553165
HHPersons	2.45	2.26	2.81
PPSM	467.33	12.8	11379.5
White	92.2	45.5	98.4
AfricanAmerican	4.78	0.4	44.2
AsianOther	3.02	1.2	10.3
Repub	28.77	8.54	39.49
Dem	26.98	14.19	55.12
EvanProt	9.58	2.39	23.62
MainProt	17.02	6.24	27.51
Catholic	20.57	1.07	71.4
Unclaim	51.2	11.42	71.63
Under18	20.34	8.8	24.3
Income1	6.49	3.1	14.7
Income2	5.96	2.5	9.9
Income3	12.26	6.1	18.5
Income4	11.66	6.6	15.9
Income5	15.19	9.8	20
Income6	19.7	15.2	23.5
Income7	12.35	7.5	16.8
Income8	10.62	3.5	19.2
Income9	3.18	1.1	10.7
Income10	2.56	0.2	13

4.1.3 Empirical Model

In order to test the proffered hypothesis of socialization's effect on disease transmission via location decisions at the county level, a model must be built around the dataset. This model will attempt to observe variation in flu incidence caused by the various variables explored above. Multiple regressions allow for the identification of the effects of numerous independent variables on the single dependent variable. The final model is as follows:

$$FluTotal_i = \beta_0 + \beta_1 Pop_i + \beta_2 HHpersons_i + \beta_3 PPSM_i + \beta_4 AfricanAmerican_i + \beta_5 AsianOther_i + \beta_6 Repub_i + \beta_7 Dem_i + \beta_7 EvanProt_i + \beta_9 MainProt_i + \beta_{10} Catholic_i + \beta_{11} Under18_i + \beta_{12} Income2_i + \beta_{13} Income3_i + \beta_{14} Income4_i + \beta_{15} Income5_i + \beta_{16} Income6_i + \beta_{17} Income7_i + \beta_{18} Income8_i + \beta_{19} Income9_i + \beta_{20} Income10_i + e_i$$

Here, the total flu count is the dependent variable, or the variable being tested for change. The "i" script attached to each of the variables signals observations at the county level. The 20 variables on the opposite side of the equation are independent variables, or those variables believed to have some influential power over the dependent variable. The independents are those living and social factors predicted to, at least partially, determine the spread of the flu. For each variable, β represents the coefficient, which reveals both the direction and magnitude of the independent variables' effect. β_0 stands in for the constant, or the expected level of the dependent variable should all independent variables be zero.

Notably, several of the variables from the original listing have been omitted during application of the model. Specifically, the variables accounting for percentage of white population, unclaimed religion, and the lowest income bracket will not be included in running the regression. The exclusion of these variables gives the remaining variables in each category a baseline against which to be compared.

4.1.4 Results

Now that a model is in place and both the dependent and independent variables have been identified, the regression can be estimated to search for any meaningful relationships between the independent variables and their effects on the dependent variable. Despite certain aforementioned data restrictions, using a confidence level of 95%, the regression returned some significant interesting results. The results of the regression on this county level dataset can be reviewed in Table 4.2 on the next page.

Table 4.2. Regression Results, County Level

Variable	Coefficient	Z-Statistic
Population	.002	8.38*
Average Household Persons	-262.277	-0.66
Persons Per Square Mile	-0.257	-5.18*
%African American	-0.868	-0.08
% Asian or Other	82.983	2.08*
% Republican	-2.873	-0.31
% Democrat	-1.759	-0.19
% Evangelical Protestant	-18.252	-2.06*
% Mainline Protestant	5.385	0.74
% Catholic	-1.827	-0.49
% Under 18	55.442	2.69*
% $10,000 - $14,999	43.624	0.54
% $15,000 - $24,999	-43.664	-0.88
% $25,000 - $34,999	0.826	0.02
% $35,000 - $49,999	-39.425	-0.95
% $50,000 - $74,999	-7.947	-0.20
% $75,000 - $99,999	-59.046	-1.03
% $100,000 - $149,999	-45.351	-1.17
% $150,000 - $199,999	100.057	1.08
% > $200,000	-121.556	-2.01*
R-squared 0.78		

*denotes statistical significance at a 95% confidence level

Of the 20 independent variables, the most obviously statistically significant variable is population. As expected, this variable displays a positive correlation with flu incidence. Higher populations mean higher flu counts. Naturally, a larger number of people mean not only a greater victim pool, but also a higher potential for disease-causing interactions. Similarly, two other variables attempted to account for grouping and living proximity. While the average persons per household variable was negligible, the persons per square mile indicator shows a significant but unexpected result, indicating that for each additional person per square mile, flu rates will actually decrease by 0.257

infections. While this result contradicts expectations, there could be several underlying explanations for the inverse relationships shown between this variable and flu rates. Namely, there could be a discrepancy between actual flu incidence and those counts recorded in this data, as the data can only account for individuals who were officially diagnosed in hospitals. As such, people living in high-proximity cities may recognize flu or cold-like symptoms and simply stay home until they recover. On the other hand, individuals in rural, low-proximity areas may be going to clinics for their symptoms and therefore account for an unrealistically high proportion of the flu counts in this data.

Next, the variable representing the number of people in the county under the age of 18 also shows a significantly positive relationship with flu transmission. As discussed through a former study in the literature review, school age children typically have the highest number of contacts with each other than any other age bracket. It therefore logically follows that this variable consisting of school-age children will show higher counts of the flu. Especially with their children, parents may be more likely to visit a hospital at the first sign of symptoms and have their children diagnosed with influenza, as opposed to adult populations who may be more likely to wait out or self-treat sickness. Additionally, even outside of the under 18 groups, the presence of a school-age child in a household could mean that child brings home the disease and spreads it to others outside of this variable, causing even higher counts of the flu.

Turning to the hypothesized indicators of social interaction, all three groups of race, religion, and income returned a single statistically significant result that may offer some insight into the implications of socially-charged groupings on flu rates. The percent of Asian and other races variable shows a significantly positive relationship with flu

transmission with a coefficient of 82.983; in other words, as this percentage of races increases, flu rates are also expected to rise by 82 infections per 1% increase. There could be several social reasons to explain this relationship. As this variable represents a comparison against the white majority, the result indicates that counties with a greater population of Asian and other races will have higher flu rates than the large populations of white individuals. This result could potentially imply that Asians and other minority races not otherwise accounted for in this analysis hold an overwhelming "like me" preference that is even greater than that of the white and African American populations. The stronger a group's "like me" preference, the greater effect homogeneity will have. These implications could possibly mean that if these minority races inherently express a great preference for socializing with others within their racial group, increased socialization and flu transmission due to homogeneity will substantially outweigh the observable effects of homogeneity on other racial groups. To substantiate the possibility of this explanation, consider the potential of a language barrier. It is possible that many Asian Americans either can only speak or simply prefer to speak their native languages; as such, they would have a more prominent desire to seek out others with whom they can converse in their preferred choice of language.

The significant income variable, which represents the highest income bracket of those people earning greater than $200,000 a year, produces the opposite result. This income variable expresses a negative relationship with flu transmission, showing that as a county's distribution of highest-income earners increases by 1%, flu infections will decrease by 121.556, in comparison to flu infections of the lowest income bracket. The highest income bracket consists of a small portion of the population. Assuming that these

people would then only socialize with their equally high-income compatriots, even if this group grows slightly, that influx will represent a number of people leaving the greater pool of socialization and only socializing within a small group. This change would mean less connections and, overall, less interaction to potentially spread the flu. While there would be some increased interaction within the top income bracket, this increase would be overpowered by the decrease in transmission pathways from the larger pool outside of this bracket. As such, the negative relationship between these high-income earners and flu rates can potentially be explained. Additionally, greater income also insinuates more and better health care; this wage bracket may show a negative relationship due having enough money to prevent and treat influenza without hospital intervention.

Similarly, the religious variable for Evangelical Protestantism also has an inverse relationship with flu counts. Pennsylvania has become particularly well known for its populations of Evangelical Protestants, especially those of an Amish or Mennonite following. Members of these groups primarily live in settlements detached from the larger population and have little interaction with others outside of their religious group. An increase in any of these religious groups still represents a number of people who are only socializing within that small, independent population. Although these religious groups are, in general, a small portion of the population at large, their extreme preferences for "like me" socialization leading to living in isolated, homogenous communities could provide one possible explanation for the results discovered in this project.

Overall, this model does not account for every facet originally expected to influence flu rates. The significant results that do exist, however, express some support

for the assumed "like me" preference and prompt interesting reconsiderations of this project's original hypothesis, which will be discussed in the following chapter.

4.2 HCUP Nationwide Inpatient Sample

4.2.1 Description of Data Sources

With the Pennsylvania county level dataset now thoroughly examined, a second analysis was performed on nationwide data. This data comes from the Department of Health and Human Services' Healthcare Cost and Utilization Project (HCUP) and will provide individual level analysis to supplement the county level data previously investigated. The specific dataset used, the Nationwide Inpatient Sample (NIS), consists of hospital visit records from participating states across the country. Containing over seven million observations, the dataset boasts a large collection of data about hospital visits, including information on diagnoses, procedures, patient demographics, and hospital characteristics.

Due to privacy concerns, the dataset uses records based on individual hospital visits, not based on each patient. In other words, a single patient could possibly account for multiple records shown in this collection if he or she made multiple hospital visits. Additionally, the HCUP withholds hospital indicators, such as specific location, that could lead to identification of the hospitals. These policies, although to ensure privacy rights, do limit the extent of analysis that can be performed on the data relating to this project.

4.2.2 Summary of Selected Variables

The HCUP dataset is a vast compilation of characteristics for over seven million hospital discharge records. First and foremost, a dependent variable needed to be identified from within this dataset. All of the records contained in this dataset are people diagnosed with some illness or injury, as all observations come from hospital discharge records. Thus, in order to produce this variable, those individuals diagnosed specifically with influenza were pulled from the dataset, creating a dummy dependent variable indicating a flu infection (*Flu*). Unfortunately, doing so means those with the flu will only be compared with other individuals in the hospital system and not the whole population.

From the long list of variables included in the HCUP dataset, 31 additional variables were selected as part of the analysis. This selection attempts to account for location, demographics, and socio-economic status as much as possible given privacy restrictions on the data. All of the independent variables, with the exception of the patient's age (*Age*), are in the form of dummy variables, which was necessitated by the format of the original dataset.

Table 4.3 at the end of this section displays the full list of variables along with the frequency in which these variables occur within the dataset. The socioeconomic variable groups patients into four brackets (*Income1 – Income4)* based on the median income of the patients' ZIP codes as identified in Table 4.3. Median income provides less information on the patients than individual income earnings would, but still offers some insight into the neighborhoods in which these people live. The next variable set indicates patient location, determined by the populations of the metropolitan areas in which they

live (*Patloc1 – Patloc6*) rather than by geographical location. This variable lends itself to establishing living proximity and urban versus rural areas. In addition to the patient location variable, a group of hospital indicators specify the geographic locations of the hospitals by Census Division (*Hosploc1 – Hosploc9*). Again, this variable is on a large-scale to protect privacy, but geographic location by Census Division may still offer additional insight. Unlike in the county level data, this individual dataset now includes a time variable to identify the yearly quarter when a patient was discharged (*Quarter1 – Quarter4*). The time of year may also help to account for fluctuations in flu rates. Lastly, a series of race variables denote the racial affiliation of each patient (*Race1 – Race6*). Together, these groups of variable on an individual level will hopefully provide more information on the relationship between location-spurred socialization and flu transmission.

Table 4.3. Summary of HCUP Data

Variable	Description	Frequency	Percent
Flu	Positive Flu Diagnosis	7,122	.10%
Age	Age of patient (1-124)	n/a	n/a
Female	Sex of patient	4,208,733	57.68%
Income1	$1-$38,999	2,189,203	30.00%
Income2	$39,000-$47,999	1,775,913	24.34%
Income3	$48,000-$62,999	1,685,328	23.10%
Income4	> $63,000	1,484,010	20.34%
Patloc1	Central county, ≥ 1 million	2,126,757	29.15%
Patloc2	Fringe county, ≥ 1 million	1,727,785	23.68%
Patloc3	250,000-999,999	1,412,535	19.36%
Patloc4	50,000-249,999	683,405	9.37%
Patloc5	Micropolitan	786,879	10.78%
Patloc6	Neither metropolitan nor micropolitan	529,917	7.26%
Hosploc1	New England	342,492	4.69%
Hosploc2	Middle Atlantic	1,053,839	14.44%
Hosploc3	East North Central	1,146,546	15.71%
Hosploc4	West North Central	501,101	6.87%
Hosploc5	South Atlantic	1,463,015	20.05%
Hosploc6	East South Central	504,873	6.92%
Hosploc7	West South Central	854,729	11.71%
Hosploc8	Mountain	446,178	6.11%
Hosploc9	Pacific	984,195	13.49%
Quarter1	January-March	1,860,903	25.50%
Quarter2	April-June	1,807,445	24.77%
Quarter3	July-September	1,814,591	24.87%
Quarter4	October-December	1,806,402	24.76%
Race1	White	4,551,845	62.38%
Race2	Black	1,014,697	13.91%
Race3	Hispanic	815,336	11.17%
Race4	Asian or Pacific Islander	186,612	2.56%
Race5	Native American	50,226	0.69%
Race6	Other	259,780	3.56%

4.2.3 Empirical Model

With the selected variables now defined, a model can be built to test the effects of these indicators on flu rates. Different from the former regression, this dataset now deals with a dummy dependent variable and primarily dummy independent variables, requiring a different type of regression to better fit the results within the constraints of these dummy variables. A probit regression was used for this dataset wherein the dependent variable, or the successful outcome, is a positive flu diagnosis. The model then tests the independent variables for their influence in generating that desired outcome and assesses the probability of the successful outcome occurring due to the independent variables of interest. The following model outlines the variables that will be tested through this regression:

$$Flu_i = \beta_0 + \beta_1 Age_i + \beta_2 Female_i + \beta_3 Income2_i + \beta_4 Income3_i + \beta_5 Income4_i + \beta_6 Patloc2_i + \beta_7 Patloc3_i + \beta_8 Patloc4_i + \beta_9 Patloc5_i + \beta_{10} Patloc6_i + \beta_{11} Hosploc2_i + \beta_{12} Hosploc3_i + \beta_{13} Hosploc4_i + \beta_{14} Hosploc5_i + \beta_{15} Hosploc6_i + \beta_{16} Hosploc7_i + \beta_{17} Hosploc8_i + \beta_{18} Hosploc9_i + \beta_{19} Quarter2_i + \beta_{20} Quarter3_i + \beta_{21} Quarter4_i + \beta_{22} Race2_i + \beta_{23} Race3_i + \beta_{24} Race4_i + \beta_{25} Race5_i + \beta_{26} Race6_i + e_i$$

Again, one variable from each set has been left out to serve as a baseline: the lowest income quartile, the first patient and hospital locations, the first quarter of the year, and the white race variable. Here, "i" now denotes the use of individual level data. The model has been built and explained, so the 26 independent variables can be tested for their effects on flu infection.

4.2.4 Results

With the probit regression estimated, the results will help determine what factors contribute to the likelihood of a flu diagnosis and hopefully provide some insight on potential roles of location and socialization on this outcome. The results of the second test on individual level data can be viewed below in Table 4.4.

Table 4.4. Regression Results, HCUP

Variable	Coefficient	Z-Statistic
Age	0.00000157	5.13*
Female	-0.000147	-9.30*
Income2	-0.000047	-2.10*
Income3	-0.0000111	-0.45
Income4	0.000039	1.37
Patloc2	-0.0000256	-1.00
Patloc3	0.000097	3.68*
Patloc4	0.0001465	4.27*
Patloc5	0.000128	3.84*
Patloc6	0.0002786	6.93*
Hosploc2	-0.0002866	-7.55*
Hosploc3	-0.0000556	-1.35
Hosploc4	0.0002091	4.06*
Hosploc5	-0.0000459	-1.14
Hosploc6	-0.000093	-2.01*
Hosploc7	-0.0001611	-3.95*
Hosploc8	-0.0000651	-1.37
Hosploc9	-0.0003137	-8.23*
Quarter2	-0.0002733	-12.10*
Quarter3	-0.0009063	-31.15*
Quarter4	0.0006219	25.88*
Race2	0.0002	7.15*
Race3	0.0003531	10.68*
Race4	0.0001491	2.36*
Race5	0.0002532	2.37*
Race6	0.000151	2.93*
Pseudo R-squared	0.04	

*denotes statistical significance at a 95% confidence level

Presumably due to the extensive sample size of the dataset, the majority of the tested variables returned significant results at a 95% confidence level, though all of the coefficients indicate only minimal effects on flu infection.

The significant age variable shows the expected, logical result. With a positive coefficient, the variable shows that older individuals have a higher flu infection rate. It is likely that older individuals suffer from weaker immune systems, thus making them more susceptible to the flu. This result could also be linked to the elderly having more severe symptoms or being more likely to visit the hospital for any symptoms, while other younger generations might not seek professional medical advice. To a lesser degree, some of this increased flu spread among the elderly could possibly be attributed to the notion of community as well. Many elderly individuals move into group homes and communities where they surround themselves with other senior citizens. The combination of homogenous living and weaker immunity could possibly mean greater spread of the flu in these groups.

Several variables returned significant results, but do not provide much conclusive support or understanding. Curiously, the female indicator returned with a negative relationship to flu rates, revealing that women have fewer flu diagnoses than men. There could be a variety of reasons for this result, ranging from potential biological differences to women simply visit the hospital for treatment less than men. Unfortunately, the income variable set represents the median income of the patients' ZIP code areas, so it does not translate directly to the actual socioeconomic status of each patient. In contrast to the previous county level data where the significant income variable was the highest income bracket, this dataset shows only that the second income quartile is less indicative

of a flu diagnosis than the first quartile of the lowest income areas. Similarly, the hospital location set was also used as a control and implies that some variation in flu outcomes may be caused by general geographic location in the country.

All but one of the patient location variables, which are better described as indicators of living proximity than actual location, are significant. Once again, the proximity variables deviate from the expected logic of more people leading to more flu infection; these variables show a positive correlation between the less populated locations and higher flu rates. These regions may follow the same explanation described above with regards to the county level persons per square mile variable.

The time variables all produced significant results that fit expectations of flu infection, concurring with common knowledge about the flu. The three quarter variables used in the regression are compared against the first quarter of the year, or the months of January, February, and March. Influenza is known to be more pervasive in the cold winter months, so the *Quarter2* and *Quarter3* variables accounting for spring and summer months display negative coefficients, which indicate that flu rates decrease in these summer months. Oppositely, the final time variable, *Quarter4*, represents October, November, and December, typically associated as being the coldest winter months. This variable has a positive relationship with flu rates, indicating that during this time period, the number of flu infections rises in comparison to the first quarter.

Interestingly, all of the race variables returned significant results and positive relationships with flu rates. These results indicate that, compared to the white majority, greater distributions of these minority races produce greater transmission of the flu. While the county level data did not produce an observable effect between the African

American population and flu rates, this extensive bank of individual level data now shows significant correlations with all minority races. In relation to this project, the significant and positive coefficients on all of these variables potentially mean that all minority races express a larger "like me" preference in comparison to the white population, as previously explored in the county level analysis. One possible explanation for the emergence of this trend could be that minorities express varying degrees of a "like me" preference compared against the white population; using a dataset as large as the HCUP collection, even smaller degrees of this preference become detectable.

Although this dataset allows individual level analysis, the need for privacy protection prevents the HCUP from releasing more information that might lead to a better application of these results to the issue of neighborhood socialization. In the next chapter, however, some trends from both the county level dataset and the HCUP individual level data can be assessed together to draw final conclusions for this project.

Chapter Five: Conclusion

This project has sought to determine how urban economic theory can be applied to disease transmission. Socialization is a major function of every day life and has inherent connections to location. Disease spread, especially of ordinary communicable diseases like colds and the flu, remains a common negative consequence of socialization. Through these relationships, this project purported that there may then exist a connection between location and flu incidence. Under the assumption of a "like me" preference for socialization, homogenous neighborhoods would theoretically see greater flu prevalence due to the inclinations of their people to socialize more within that group.

Previous research and published literature created a foundation from which to launch this project. Separate studies asserted the individual theories on flu transmission and neighborhood choice, while other studies then began to explore the relationship between location and socialization. Finally, these ideas were synthesized into theoretical and empirical application. Following the examples of those former studies, this project then also attempted to establish a theoretical framework and follow through with regression analysis. The theory drew from urban neighborhood choice theory and social network analysis. Neighborhood choice speaks to how different neighborhoods, segregated, integrated, and mixed, form, which leads to discussion of how these different types of neighborhood have different implications for socialization. A segregated, or homogenous, neighborhood would see more social interaction and flu spread under the "like me" theory than mixed or integrated neighborhoods. In addition, social network analysis compliments this theory by mapping the pathways of interaction; by being able

to build a matrix of an individual's interaction patterns, it is possible to track potential routes of disease infection.

Empirical analysis did provide some support for the ideas espoused by this project. Application of the county level results shows the possibility of the "like me" socialization preference that was first established through neighborhood choice theory. In conjunction with the results relating to the highest income earners, the relationship found between Evangelical Protestantism and flu rates introduces an issue as of yet unaccounted for in this project. While people may indeed be socializing based on a "like me" preference, homogenous populations may in fact lead to less flu incidences if those populations are able to detach themselves from others. Through this lens, the results of the Asian and other races, highest income, and Evangelical Protestant variables can be reevaluated. The results exhibit support for the actual presence of the "like me" preference assumed in this project; however, the effect of this preference may be more complicated than originally hypothesized. If people have a primary desire to socialize with others like themselves, high-income earners have the monetary ability to separate themselves and indulge in activities with others of their income class. For example, highest-income households may live in gated, high-income communities, interact professionally with other high-income workers, send children to expensive and exclusive schools, indulge in expensive recreational activities, or even have the ability to avoid public transportation, all of which limit their interaction with the general population and serve to isolate them as a homogenous group. Similarly, members of certain Evangelical Protestant sects in the state have established their own communities and societies, thus

separating themselves from the main population and allowing them to also keep interaction restricted to their religious peers.

Conversely, the variable accounting for Asian and other races shows a positive correlation with flu rates. Based on this perspective, the result could be due to the fact that this population of people is majorly driven by a "like me" preference, but is, as a homogenous whole, unable to separate into a small cluster. Asian and other minority races may indeed express a greater willingness to locate near each other, allowing them the increased socialization with others of the same group, as suggested by the underlying theory. Unfortunately, especially at the county level, this division of race is not apparent. While on a smaller level there may exist isolated neighborhoods of any of the minority races included in this variable, they are still largely integrated with the general population through working, shopping, transportation, and other activities. As such, while the "like me" preference may promote greater socialization within the group, that increased socialization is in addition to unavoidable interaction with others outside of the group thus raising overall disease transmission. The HCUP dataset results further lent support to this theory as the minority race variables all showed a significant contribution to flu infection compared to the white majority.

Because these first two groups have the ability to isolate themselves from others, their preference for "like me" interactions not only promotes socialization within the group, but also then keeps them confined in these small populations, so flu spread becomes limited as well. The original statement regarding homogeneity promoting socialization and thus disease transmission did not account for the ability of these homogenous groups to socialize separate from the main population in small independent

populations, and so avoiding many avenues of flu infection. These results would thusly suggest that the effect of neighborhood homogeneity on flu transmission hinges on the size and total social connectivity of that neighborhood. Homogeneity and increased socialization in a large area can quickly prompt disease transmission. In a small or secluded community, however, the desire to socialize with similar people presumably stifles other social interactions and ultimately keeps disease confined, meaning lower transmission rates.

By itself, the individual level HCUP data lacked direct applicability to the issue of neighborhood socialization because of restrictions on person and hospital identifiers. The data still supplemented the original analysis on county level data and helped provide the most well-rounded analysis possible for this project. This individual level data added support to the primary theory that emerged through this project about greater degrees of "like me" preferences across races. While neither dataset provides the neighborhood-level data necessary to make complete assumptions about the original theory behind this project, finding potential support of this "like me" socialization affecting flu transmission could possibly be linked back to neighborhood makeups if the proper data existed.

Naturally, there may be relationships between flu incidence and social factors yet to be determined. There are also a plethora of other factors to consider, such as many cases wherein social preferences are not a primary driver in location. The urban economic theory at work in this project specifically investigates households' willingness to pay for certain qualities in their neighbors, but does not explore households choosing location based on characteristics other than neighbors. Further study may reveal

connections between locational interaction and disease transmission not discovered through this project.

Bibliography

Bansal, S., Pourbohloul, B., Hupert, N., Grenfell, B., & Meyers, L.A. (2010). The shifting demographic landscape of pandemic influenza. *PLoS ONE, 5*(2): e9360, 1-8.

Bandyopadhyay, S., Rao, A.R., & Sinha, B.K. (2011). *Models for social networks with statistical applications*. USA: SAGE Publications, Inc.

Clark, W.A.V, & Ledwith, V. (2007). How much does income matter in neighborhood choice? *Population Research and Policy Review, 26*(2), 145-161.

Emch, M., Root, E.D., Giebultowicz, S., Ali, M., Perez-Heydrich, C., & Yunus, M. (2012). Integration of spatial and social network analysis in disease transmission studies. *Annals of the Association of American Geographers, 102*(5), 1004-1015.

Keeling, M. J., & Eames, K.T.D. (2005). Networks and epidemic models. *Journal of the Royal Society Interface*, *2*, 295-303.

Tam, K., Yousey-Hindes, K., & Hadler, J.L. (2014). Influenza-related hospitalization of adults associated with low census tract socioeconomic status and female sex in New Haven County, Connecticut, 2007-2011. *Influenza and Other Respiratory Viruses, 8*(3), 274-281.

Van Kerckhove, K., Hens, N., Edmunds, W.J., & Eames, K.T.D. (2013). The impact of illness on social networks: implications for transmission and control of influenza. *American Journal of Epidemiology, 178*(11), 1655-1662.