Socially Optimal Vaccination Policies

Louis Boguchwal
Department of Economics
Hamilton College
Clinton, New York 13323
email: louis.boguchwal@gmail.com

Abstract: This research provides a novel framework for modeling infectious disease propagation throughout a population, and ultimately for determining socially optimal vaccine allocation. Given a fixed stockpile of vaccines, I determine the best geographic distribution that minimizes the number of infections in the population. This approach integrates network modeling and statistical analysis. The analysis reveals both the network attributes as well as the demographic characteristics that are critical in determining whether one region is more important than another in terms of vaccination. Finally, a framework using optimization methods is proposed for finding the nearest-to-optimal policy that would be amenable to policy-makers representing different constituencies.

1. Literature Review and Introduction

1.1 – Infectious Disease Modeling for a Single Population

Infectious disease continues to be a great societal concern at the local, national, and international levels. Diseases such as influenza affect large proportions of the world's population and are responsible for 250,000-500,000 deaths each year (World Health Organization, 2009). Governments as well as health organizations are interested in understanding the spread of infectious disease and steps that can be taken to minimize the loss of life. The dominant approach to combating infectious disease has been vaccination. However, the vaccination method is imperfect, as the number of susceptible individuals within a particular population generally exceeds the number of available vaccines. Thus, governments around the globe are faced with the problem of allocating a limited number of vaccines in order to maximize social welfare.

The first academic to formally investigate the impact of infectious disease on a population was Daniel Bernoulli. He attempted to estimate the increase in life expectancy if smallpox were eradicated from the population (Bernoulli, 1766). These increases were compared to the benchmark of inaction on the part of the government. This work was a persuasive piece, which argued for the mass-vaccination of the population. Bernoulli performed a mathematical cost-benefit analysis of smallpox vaccines, and found that the benefits of mass-vaccination outweigh the risks associated with it. Bernoulli calculated that such a policy would increase life expectancy by over three years.

About two centuries later, Kermack and McKendrick published their seminal paper in epidemic modeling (Kermack & McKendrick, 1927). They introduced Susceptible, Infected, and Removed (SIR) ordinary differential equations models for a single population. These models formed the foundation of epidemic theory and mathematical epidemiology. An SIR model classifies the individuals within a population into one of three disease compartments: susceptible, infected, and removed. Susceptible individuals are disease-free, but have the potential to contract the disease should they come into contact with an infected individual. Infected individuals have the disease and can spread it to others. Removed individuals have either developed immunity to the disease or have died. Hence, they are removed from the population in terms of disease propagation potential. An individual transitions through the disease states in the order susceptible, infected, removed. Kermack and McKendrick's model is given below, where $x(t)$, $y(t)$, and $z(t)$ are respectively, the susceptible, infective, and removal functions of time, and β and γ are the infection and removal parameters respectively. Note that $\beta, \gamma > 0$.

$$\frac{dx}{dt} = -\beta xy$$

$$\frac{dy}{dt} = \beta xy - \gamma y \tag{1}$$

$$\frac{dz}{dt} = \gamma y$$

The population is assumed to be of fixed size *N* for all time. That is, $x(t) + y(t) + z(t) = N$ for all *t*. Further, this single population SIR model assumes homogeneous mixing. In other words, all susceptible individuals in the population are equally likely to contract the disease and all infected individuals in the population are equally likely to transmit the disease (Frauenthal, 1980). These disease dynamics lead one to inquire as to the circumstances under which disease eradication is possible. Does there exist a scenario in which the epidemic would end when there are a positive number of susceptibles left in the population? The Threshold Theorem reveals that there is. It demonstrates that if $x_0 > \frac{\gamma}{\beta}$, the relative removal rate, then the epidemic will spread (Daley & Gani, 1999). Otherwise, the epidemic will cease to exist within the population. Hence, if the initial number of susceptibles could be reduced to some number below the relative removal rate, by some measures such as vaccination, then the impact of the epidemic would be greatly lessened.

Vaccination is the prevailing method of halting a mass epidemic. However, the positive externalities associated with vaccination imply that it is unnecessary to vaccinate the entire population in order to curb an epidemic. If an individual is vaccinated, he can no longer spread disease to anyone with whom he comes into contact. Based upon these externalities, there exists a critical proportion of the population receiving vaccines for which disease propagation will slow over time. Anderson and May (1991) derived the critical proportion figure through analysis of the average number of secondary cases produced by an initially infected individual into an otherwise susceptible population. This is known as the basic reproduction number R_0 (D. Mollison (Ed.), 1995). It is thought that $R_0 \in [1,3]$ (Wu, Riley & Leung, 2007). In context of the SIR model, $R_0 = \beta/\gamma$ (Sattenspiel, 2009). For a disease to spread throughout a population, the basic reproduction number must be greater than one. If $R_0 < 1$, then the epidemic will be unable to sustain itself as individuals are removed from the population faster than they are infected. Assuming that vaccination confers full immunity, Anderson and May determined the critical proportion of vaccination $p_c = 1 – 1/R_0$.

1.2 – Multi-City SIR Models and Spatial Heterogeneity

While ordinary differential equation SIR models provide insight into the dynamics of infectious disease, they are limited in scope to a single, immobile population. A model that allows for spatial heterogeneity better captures disease spread, as people are highly mobile in today’s modern age. Hyman and LaForce (2001) proposed a multi-city SIR model in which individuals are free to travel from one city to another, which generalized the work of Kermack and McKendrick. Thus, disease propagation in a particular city depends on the inflow and outflow of individuals with respect to that city. Intra-city disease dynamics are assumed to operate under homogenous mixing, while inter-city disease dynamics operate under heterogeneous mixing. The aggregate population of all cities under consideration was assumed to be constant for all time. The inter-city model was generalized to *n* cities via the use of a mobility matrix $m = (m_{ij})$, where m_{ij} is the number of people per unit time who move from city *i* to city *j*. For a multi-city network, Hyman and LaForce determined that the basic reproduction number for each city is found by computing the appropriate eigenvalue of the Jacobian matrix that corresponds to the spatially heterogeneous SIR model. The maximum eigenvalue of matrix *m* provides a rough estimate of R_0 over all cities in the network (Sattenspiel, 2009).

1.3 – Transportation Network Models and Optimal Vaccine Allocation

Shaw et. al. (2010) modeled influenza spread using a rudimentary air transportation network in the United States. In their model, intra-city disease dynamics are governed by a variant of the single population SIR model, a system of ordinary differential equations, while inter-city disease dynamics are governed by a Markov transition matrix generated from the air transportation network. The Markov transition matrix is derived from the mobility matrix corresponding to the weighted adjacency matrix of the air transportation network. The aggregate population of all cities under consideration is assumed to be constant for all time. Over several flu seasons, the authors accurately predicted the geographic spread of influenza throughout the United States through an “importance factor” derived out of city population as well as network-connectedness. The importance factor was confirmed via weekly influenza maps provided by the Center for Disease Control. Counter-intuitively, the authors found that as population size increases, likelihood of spread decreases. The spread of an epidemic is fast in smaller cities, as it is easier to engulf a small city with disease in comparison to a large city. Therefore, the likelihood of an infected individual flying out of a small city is greater than the likelihood of an infected individual flying out of a large city. The authors also found that cities of high incoming traffic were disease propagation hubs because high incoming traffic implies a high likelihood of arriving infectives.

Wu et. al. (2007) attacked the issue of spatial heterogeneity in a slightly different manner. Rather than partitioning the network into nodes by city or airport, they partitioned the network into ten nodes by regions, as defined by the United States Office of Management and Budget. The entries of the mobility matrix $m = (m_{ij})$ represent "the average proportion of time that a resident of population i spends in population j over one year." The mobility matrix is more realistically defined than in other works, as it incorporates multiple modes of transportation as opposed to only one. While the mobility matrix of this model itself is precisely defined, the transportation network is imprecisely defined. The larger the regional definition of a node, the less the model can hone in on spatial heterogeneity. Spatial heterogeneity at the city level is lost and assumed to be spatial homogeneity when cities are grouped together into larger regions.

An additional approach to modeling optimal vaccine allocation under spatial heterogeneity is gravity transportation theoretic models. In the twenty-first century, human mobility spans both short and long distances. Transportation flows governed by representative transportation modes for both short and long distance travel form a more realistic basis for the geographic spread of infectious disease. Balcana et. al. used global commuting patterns as well as global air traffic flows to model short and long distance travel, respectively (Balcana, Colizza, Goncalvesa, Hud, Ramascob, & Vespignani, 2009). Their analysis revealed that despite the comparatively large volume of commuting flow compared to air traffic flow, the most important network when considering infectious disease propagation is the long distance air transportation network.

Colizza et. al. constructed a stochastic global air transportation network model using data from the International Air Transport Association (IATA). An SIR model governed the within-node disease dynamics. Their network consisted of a weighted graph in which nodes represented airports and weighted edges represented passenger flows between airports (Colizza, Barrat, Barthe´lemy, & Vespignani, 2006). The air transportation network under consideration accounted for 99% of worldwide air transportation traffic. As one would expect, this network exhibited spatial heterogeneity. The authors investigated the effects of mobility along the worldwide air transportation network and its degree distribution, i.e. the frequency distribution of node degree, on the spread of infectious disease. Passenger flows between cities were governed by a stochastic transport operator. This operator described the net balance of individuals of each disease compartment who left and entered a particular city.

1.4 – Network Attributes

In contrast to multi-city models, infectious disease propagation can be explored through a network in which nodes represent individuals. At any given time, each node bears a disease status as susceptible, infected, or removed. Salathé et. al. examined the effects of degree distribution, community structure, betweenness centrality, and random-walk centrality on the spread of infectious disease along a network (Salathé & Jones, 2010). The authors postulated that in networks of high community structure, nodes acting as community bridges are important to vaccinate, where community structure is measured by modularity and community bridges are identified via betweenness centrality, which is defined below. Further, the authors concluded that, in general, a vaccination policy based upon betweenness centrality would result in fewer total infections throughout the population than a vaccination policy based solely upon degree. Hence, analysis of degree alone is insufficient to form the basis for vaccine allocation policy decisions. Nodes were ranked for vaccination based upon degree, betweenness centrality, and random-walk centrality in decreasing order for each criterion.

In order to enact socially optimal vaccination policies, policy-makers not only need to be well-informed with regard to infectious disease dynamics, but also with regard to vaccination policy analysis. Markov chains, transportation network analysis, and simulation have contributed to the most notable developments in efficient vaccine allocation methods. In the model presented above, Shaw et. al. measured the impact of a pathogen on a population by the aggregate number of sick days over 450 total days, through computer simulation. Evolutionary algorithms were used to minimize the total number of sick days taken throughout the population (Shaw, Spears, Billings, & Maxim, 2010). The resulting vaccination policies were then compared to two intuitive benchmark policies, namely uniform allocation by city and proportional allocation by city population. Ultimately, Shaw et. al. found that more efficient vaccination policies depend on the current disease distribution, the in-degree structure of the transportation network, and city population. Cities were ranked in decreasing order based upon importance factors

$$imp_i = \frac{S_{t,i}}{n_i} \qquad (2)$$

where $s_{t,i} \equiv$ the proportion of the disease distribution of city *i* at time *t* and $n_i \equiv$ the population of city *i*. This ranking proved to be surprisingly indicative of true city-importance regarding vaccination. The model presented demonstrated the importance of network structure on the effectiveness of a vaccination policy. However, this research only began to explore the effects of transportation network structure on disease spread. Numerous network attributes such as various node centrality measures were left unexplored that potentially play critical roles in disease propagation.

Wu et. al. viewed vaccine allocation as a nonlinear optimization problem in which the objective was to minimize the total number of infections across a population (Wu, Riley & Leung, 2007). A portion of the total vaccine supply was designated to be distributed pro-rata, which is equitable distribution by population, and the other portion of the vaccine supply was to be discretionarily distributed. The authors found that the purely equitable distribution was the least efficient allocation in terms of minimizing the total number of infections in the population. The purely discretionary policy is efficient, but highly inequitable. The improvement over the pro-rata policy is trivial, for a large "inequality cost." Hence, the authors concluded that the gain in efficiency is insufficient to justify the massive inequity corresponding to the completely discretionary allocation policy.

A limitation of the network models presented above is that they are all specific to disease propagation. However, network models applied to more general contexts also provide insight into the vaccine allocation problem. Guimerá et. al. considered the worldwide air transportation network using Official Airline Guide (OAG) data. Nodes were defined by city, rather than airport. They sought to evaluate the global importance of a particular city in reference to the entire network using shortest path length, clustering coefficient, betweenness centrality, community structure as defined by modularity, within-community degree, and total degree (Guimerá, Mossa, Turtschi, & L.A.N. Amaral, 2005). Using the above attributes, the authors proceeded to classify nodes by their connectedness as well as their "participation" in the network. The main limitation of this framework is that many attributes described in the research depend upon a well-defined notion of community structure. However, it is exceedingly difficult to detect and distinguish communities within a larger network.

1.5 – Policy Evaluation, Feasibility, and Implementation

While the aforementioned methods provide a framework for determining optimal vaccine allocation, it is imperative to investigate the accuracy of disease propagation model predictions as well as the feasibility of implementation of theoretical findings. First, policy recommendations based upon disease spread models should be tested empirically. For example, Shaw et. al. confirmed their importance factor using Center for Disease Control map data (2010). Next, it may be unreasonable to enact a policy recommendation rooted in theory. For example, the increase in social welfare of a given policy might be trivial compared to the inequity associated with it. Hence, such a policy is unlikely to pass in a real-world situation, as Wu et. al. hypothesized (2007).

One useful method of evaluating vaccination policy feasibility is to compare a particular allocation to what individuals would do if left to their own devices. It is important to note that all individuals would not vaccinate themselves under their own volition because of the free-rider problem, where it is not always in the best interest of an individual to vaccinate (Heal & Kunreuther, 2005). Free-riding behavior leads to situations that are not necessarily socially optimal. Heal and Kunreuther investigated the conditions under which individuals would act in accordance with socially optimal vaccination behavior.

Galvani et. al. explored the differences between perceived risk of infectious disease and actual risk of infectious disease. They found that there is a "large discrepancy between people's perceptions and the actual epidemiological facts and figures" (Galvani, Reluga & Chapman, 2007). Therefore, individuals likely make great errors when making their choices for vaccination. Thus, these errors must be taken into account when considering policy actions to be taken to move toward a socially optimal allocation.

1.6 – Network Attribute Extensions

In order to analyze the relationship between network structure, disease propagation, and ultimately vaccination policy, one must define and establish the important relevant network attributes. I believe that degree centrality, closeness centrality, betweenness centrality, and node significance are important network attributes to consider in context of disease spread. To my knowledge, these properties have never been examined simultaneously in the context of a multi-region spatially heterogeneous disease propagation model before. Spatial heterogeneity implies the existence of target regions for vaccination. Understanding which regions are most important to the network would provide insight as to the regions most critical for vaccination such that the impact of the infectious disease on the population is minimized. Node centrality captures the importance of a node to a network.

First, define the degree centrality of a node *i* to be the number of adjacencies to other nodes *i* has, where an adjacency is a node that is connected by an edge to *i*. Denote the degree centrality of *i* $c_D(i) = k_i$, for an unweighted and undirected network (Jackson, 2008). If a node *i* has numerous neighbors, it would have high degree centrality. Next, generalize degree centrality to weighted and directed networks. A weighted network is a graph in which the edges are labeled with non-uniform values that represent some attribute, such as passenger flow between two nodes. A directed network is a graph such that each edge bears a direction. In other words, the existence of an edge *(i,j)* does not imply the existence of an edge *(j,i)* in general. Further, if edges *(i,j)* and *(j,i)* exist, then they differ in weight in general. When evaluating the degree centrality of a node *i* in a weighted network, both the number of adjacencies as well as the aggregate weight of those adjacencies should play a role. Opsahl et. al. proposed an intuitive and useful degree centrality measure for weighted networks:

$$C_D^{w\alpha}(i) = k_i \times (\frac{w_i}{k_i})^{\alpha} \tag{3}$$

where k_i is defined as before, w_i is the aggregate weight of node *i*'s adjacencies, and α is a tuning parameter that accounts for the desired impact of degree and weight relative to one another on the degree centrality measure (Opsahl, Agneessens, & Skvoretz, 2010). If a node *i* has both numerous neighbors and high aggregate edge-weight, then it would have high degree centrality. This notion of degree centrality can be generalized for in-degree and out-degree by computing ${}^{IN}C_D^{w\alpha}(i)$ and ${}^{OUT}C_D^{w\alpha}(i)$.

Now, define the closeness centrality of a node *i* to be the inverse sum of the distances to all other nodes in the network. Denote closeness centrality

$$c_C(i) = \frac{1}{\sum_{j=1}^{n} d_{ij}} \tag{4}$$

where d_{ij} is the shortest distance from node *i* to node *j* (Golbeck). If a city *i* is close to all other cities in the network, it would have high closeness centrality. The sum of the distances from *i* to all other nodes would be small, implying that the reciprocal of this sum would be high. This centrality measure readily generalizes to weighted and directed networks without modification.

Next, define the betweenness centrality of a node *i* to be the fraction of shortest paths between nodes *s* and *t* that include node *i*. Denote betweenness centrality

$$c_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \tag{5}$$

where $\sigma_{st}(i)$ is the number of shortest paths from node *s* to node *t* that include node *i* and σ_{st} is the total number of shortest paths between nodes *s* and *t* (Salathé & Jones, 2010). If a node lies on most shortest paths between node pairs in the network, it would have high betweenness centrality. This measure indicates the likelihood of traversing a given node when traveling along the network. This centrality measure readily generalizes to weighted and directed networks without modification.

Finally, define node significance of a node *i* to be the difference in shortest path length between nodes *s* and *t* in the network excluding node *i* and the shortest path length between nodes *s* and *t* in the network including node *i*. Denote node significance

$$c_S(i) = \sum_{s \neq t \neq i} d_{st}(G \setminus \{i\}) - d_{st}(G) \tag{6}$$

where *G* is the network and d_{st} is the shortest path distance from *s* to *t*. Node significance is related to, but more precise than, betweenness centrality. Rather than merely measuring the proportion of shortest paths a node *i* is a part of, a binary concept, node significance measures the level to which shortest path distances are penalized. If a node *i* lies on a shortest path from *s* to *t*, the deletion of *i* from the network, in general, increases the shortest path length, d_{st}. To the contrary, if *i* does not lie on the shortest path from *s* to *t*, then d_{st} is unchanged in $G\backslash\{i\}$ relative to *G*. Two nodes *i* and *j* may have the same betweenness centrality values, but very different node significance values. For example, *j* might be more important to the network because its deletion implies massive increases in shortest path lengths, despite its betweenness centrality being identical to *i*'s. If a node is critical in the formation of a shortest path between several node pairs in the network it will have high node significance. This centrality measure readily generalizes to weighted and directed networks without modification. My notion of node significance is a variation of the one proposed by Mark and Rushton (2005).

1.7 –Thesis Outline

I expand upon the optimal vaccine allocation research presented above by thoroughly investigating the role of air transportation network structure on disease propagation, and ultimately on socially optimal vaccination policies. This research seeks to determine optimal vaccine allocation methods given a fixed stockpile of vaccines in a particular country, in this case the United States. The network under consideration is the air transportation network of the United States, generated from the United States Bureau of Transportation Statistics T-100 Domestic Segment dataset (BTS T-100 dataset). The disease under consideration is influenza, defined as any influenza-like-illness (ILI).

I examine multi-city network attributes and demographic characteristics in an effort to reveal the attributes relevant to disease incidence, through statistical analysis. Further, I construct a dynamic model that incorporates time and current disease incidence to predict activity levels in future periods. The two methods allow me to predict disease incidence. Based upon these predictions, regions can be ranked by importance for vaccination in order to maximize social welfare.

Finally, the allocation policies proposed will be evaluated with regard to feasibility of implementation. Theoretical and empirical derivations provide a framework for solving the optimal vaccine allocation problem, but are insufficient solutions without adequate real-world policy assessment. I am concerned with optimal vaccination policies such that these policies are feasible in a real-world context. I propose a novel optimization framework to evaluate a socially optimal, yet potentially inequitable, policy with regard to political feasibility.

2 – The Models

Recall, the infectious disease under consideration is influenza, defined as any influenza-like illness (ILI). However, the disease propagation models presented could apply to many other infectious diseases, provided that the disease spreads through contact-infection. I model disease spread between different states within the U.S. Disease dynamics are modeled using statistical and dynamic approaches, both of which relate to passenger flows along the United States air transportation network.

2.1 – Inter-state Disease Models

Within the United States, inter-state disease dynamics are modeled using both statistical and dynamic methods. These approaches are discussed in Section 3. The domestic air transportation network of the United States, derived from the Bureau of Transportation Statistics T-100 Domestic Segment 2011 dataset, is an integral part of each.

To construct the network, I created a mobility matrix in Microsoft Office Excel 2010 by United States airports via the T-100 dataset. This mobility matrix[1] is a weighted adjacency matrix *m*, where each entry m_{ij}, represents the number of passengers who traveled from airport *i* to airport *j* in the first ten months of the year 2011. However, a mobility matrix defined by city, as in Guimerá et al., is more meaningful than a mobility matrix defined by airport when discussing infectious disease spread. Airports do not have residents or any social characteristics such as population, whereas cities do. Thus, I constructed a mobility matrix $m = m_{ij}$ by city passenger flows from a city *i* to a city *j* by subtotaling rows and columns of the airport-defined mobility matrix. If several airports are in the same city market, as defined by the Bureau of Transportation Statistics City Market ID, their passenger flows were summed in order to reflect true city passenger flows. Finally, all diagonals m_{ii} were made zero to reflect that there is no passenger flow from a city to itself. Some diagonals were initially nonzero because of layovers. For example, if a passenger is flying from San Francisco, California to Syracuse, New York he or she might have a layover in New York City. The first leg of the trip might start in

[1] Note that this matrix is not square because not all airports are both origins and destinations, as an airport may be exclusively one or the other. But in order to construct the dynamic model discussed in Section 3.2, the mobility matrix must be square. Therefore, I added zero-rows and zero-columns to the matrix to incorporate any purely destination or purely origin airports, respectively. Thus, the rows and columns of the matrix contain all currently operational commercial airports in the United States.

San Francisco Airport and end in Newark International Airport. However, the second leg of the trip might start in John F. Kennedy Airport and end in Syracuse airport. The passenger flow of interest of this individual is San Francisco to New York City to Syracuse. Hence, it is sensible to make all diagonal entries m_{ij} zero.

The resulting mobility matrix $m = (m_{ij})$ is square and 1060 x 1060 in dimension. A 5 x 5 matrix, an extraction from the complete mobility matrix, can be found in Table 1 for illustrative purposes. Exploration of the network and associated node properties is found in Section 2.3.

		Kodiak	Homer	New York City	Pittsburgh	Bangor	ROW SUM
		AK	AK	NY	PA	ME	
Kodiak	**AK**	0	278	0	0	0	278
Homer	**AK**	242	0	0	0	0	242
New York City	**NY**	0	0	0	310,105	53,700	363,805
Pittsburgh	**PA**	0	0	312,054	0	99	312,153
Bangor	**ME**	0	0	57,921	98	0	58,019

Table 1: This small mobility matrix is constructed from the mobility matrix m_{ij} for illustrative purposes. Each entry represents the passenger flow from city *i* to city *j* in the first ten months of the year 2011.

2.2 – Network Attributes

The United States air transportation network and its corresponding mobility matrix not only provide passenger flow information, but also provide network centrality information. For purposes of precision and because the data was available, network attributes were computed with respect to cities, rather than states. A city's importance for vaccination greatly depends upon the overall network structure and its position in the network. From a network perspective, the salience of a city can be measured using several network attributes. Therefore, I establish and compute various measures of node centrality in order to understand which network attributes are critical to disease propagation.

Let $G = (V,E)$ be the entire United States air transportation network, V = the vertex set consisting of all cities in the U.S. air transportation network, E = the edge set consisting of all nonzero passenger flows (i,j) from i to j in the network, and $|V| = h =1060$ be the number of cities in the network. The network attributes of interest are degree centrality[2], closeness centrality, betweenness centrality, and node significance. Note that paths and distances between nodes are defined by passenger flows, as each edge *(i,j)* bears a label denoting the number of passengers who flew from city *i* to city *j* in the first ten months of the year 2011. Since nodes of high passenger inflow and outflow are of particular interest, all edge labels have been inverted[3] prior to computing the closeness centrality, betweenness centrality, and node significance so that high passenger flow paths have short distance. Recall the definitions of these attributes in equations (7) through (10):

$$C_D^{w\alpha}(i) = k_i \times (\frac{w_i}{k_i})^{\alpha} \qquad (7)$$

where k_i is the number of adjacencies of node *i*, w_i is the aggregate weight of node *i*'s adjacencies, and α is a tuning parameter that accounts for the desired impact of degree and weight relative to one another on the degree centrality measure (Opsahl, Agneessens, & Skvoretz, 2010). The in-degree and out-degree centrality measures ${}^{IN}C_D^{w\alpha}(i)$ and ${}^{OUT}C_D^{w\alpha}(i)$ were computed for greater specificity.

$$c_C(i) = \frac{1}{\sum_{j=1}^{n} d_{ij}} \qquad (8)$$

where d_{ij} is the shortest distance from node *i* to node *j* (Golbeck).

$$c_B(i) = \sum_{s \neq t \neq i} \frac{\sigma_{st}(i)}{\sigma_{st}} \qquad (9)$$

[2] Degree centrality is investigated in three categories: in-degree centrality, out-degree centrality, and net-degree centrality. These measures correspond to incoming edges, outgoing edges, and all edges, respectively.

[3] Inverting an edge refers to taking its reciprocal. For example, if an edge had weight 4 then its inverted edge would have weight 0.25.

where $\sigma_{st}(i)$ is the number of shortest paths from node s to node t that include node i and σ_{st} is the total number of shortest paths between nodes s and t (Salathé & Jones, 2010).

$$c_S(i) = \sum_{s \neq t \neq i} d_{st}(G \setminus \{i\}) - d_{st}(G) \tag{10}$$

where G is the network and d_{st} is the distance of the shortest path from s to t. These network attributes, as well as their components, are all bounded below by zero. This feature implies that no particular component of a centrality measure can impact another. A negative component to a sum would lower the overall sum, whereas a zero component would merely fail to make it greater. Thus, the above centrality measures accurately reflect a city's importance from a network perspective because no component cancels another. Observe that degree centrality depends upon the choice of α, and closeness centrality and node significance depend upon the choice of μ, the distance between nodes for which there is no path. The reader is directed to Appendix 8.7, where these parameter-sensitivities are explored and addressed.

The large size of the United States air transportation network as well as the complexity of the network attributes necessitate that network attribute data be generated by way of computer programs. The degree centrality, closeness centrality, betweenness centrality, and node significance of each city $i \in V$ are computed via the Network Attribute Suite written in the C++ programming language, created by Lindsay Shankman and Julian Aronowitz for this research project (Shankman and Aronowitz, 2012) based on my requirements and design. Given a weighted adjacency matrix $m = (m_{ij})$, programs composing this suite compute the aforementioned attributes. The generated network characteristics serve as data for the regression analysis portion of the project, discussed in Section 3.

2.3 – Social and Demographic Attributes

In addition to network attributes, social attributes are relevant to infectious disease spread. The social attributes under consideration are total population, average family size, percentage of the population older than 65, median income, percentage of the population possessing a bachelor's degree or higher, percentage of the population younger than 5, and average number of vehicles per household. My rationales for testing these specific demographic characteristics are found below.

1. Sheer population size is a pertinent social variable because disease spread likely depends upon the potential number of travelers available. More individuals allow for more social interactions, and thus greater potential for disease spread.
2. Family size is an appropriate social variable because, to an extent, disease spread likely depends upon family interactions. If one family member contracts influenza, then others likely will. Larger families might imply a higher probability having an infected family member, thereby implying an increased probability of disease spread. Alternatively, single individuals might be more likely to travel than families, and thus more likely to contract disease. Both hypotheses suppose that average family size is a pertinent to disease propagation.
3. The proportion of the population sixty-five years or older is a relevant demographic variable because retirees are likely the most susceptible to infection in a population because of weak immune systems.

4. The proportion of the population younger than five years old is a pertinent demographic characteristic because young children typically spread germs to one another while in school.
5. The number of cars per household is a pertinent demographic variable because it serves as a proxy for short-distance transportation. A greater number of cars might imply more short-distance travel, as compared to long-distance travel indexed by the air transportation network, and thus greater opportunity for disease spread.
6. Median income and proportion of the population possessing a bachelor's degree or higher serve as proxies for the population's hygienic habits. Good hygienic habits might lower the probability of an individual contracting the disease.

Data for the social characteristics mentioned above was obtained from the Census Bureau's American Factfinder. All data are from the 2006-2010 American Community Survey and are taken at the state level (U.S. Census Bureau).

The dependent variable for this study is the degree of disease spread. Data on influenza spread was obtained through Center for Disease Control weekly influenza activity maps. Activity level definitions were obtained through the Center for Disease Control weekly overview. The maps (CDC, 2011) portray the United States and color code each state by its disease activity level: no report, no activity, sporadic, local, regional, and widespread. These activity levels are mutually exclusive. Sporadic activity is defined to be "Small numbers of laboratory-confirmed influenza cases or a single laboratory-confirmed influenza outbreak has been reported, but there is no increase in cases of ILI." Local activity is defined to be "Outbreaks of influenza or increases in ILI cases and recent laboratory-confirmed influenza in a single region of the state." Regional activity is defined to be "Outbreaks of influenza or increases in ILI and recent laboratory confirmed influenza in at least two but less than half the regions of the state with recent laboratory evidence of influenza in those regions." Widespread activity is defined to be "Outbreaks of influenza or increases in ILI cases and recent laboratory-confirmed influenza in at least half the regions of the state with recent laboratory evidence of influenza in the state" (CDC, 2011). I excluded the "no report" category, as it represents missing data.

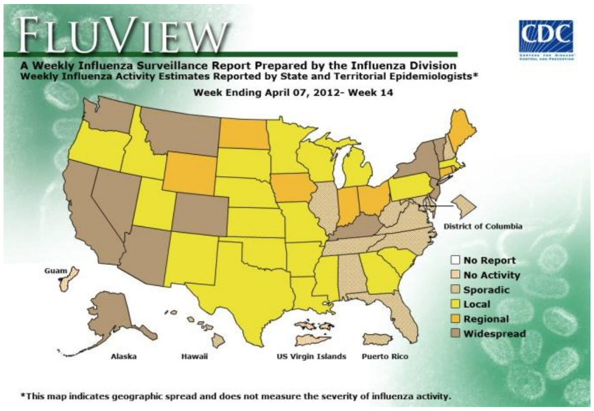
FLUVIEW
CDC
A Weekly Influenza Surveillance Report Prepared by the Influenza Division
Weekly Influenza Activity Estimates Reported by State and Territorial Epidemiologists*
Week Ending April 07, 2012- Week 14
District of Columbia
Guam
No Report
No Activity
Sporadic
Local
Regional
Widespread
Alaska
Hawaii
US Virgin Islands
Puerto Rico
*This map indicates geographic spread and does not measure the severity of influenza activity.

Figure 1: CDC weekly influenza activity map.

3 – Methodology

In order to determine optimal vaccination policies, I must first explore disease dynamics. There are two approaches to understanding disease dynamics: time-invariant modeling and time-variant modeling. The former is analyzed through statistical methods and the latter is analyzed through dynamic methods; these are presented in Sections 3.1 and 3.2, respectively. The statistical analysis provides insight into the relationship between time-invariant factors, namely the demographic and network characteristics, and disease incidence. However, this method pays no heed to how disease spreads from place-to-place in real time. The dynamic analysis predicts disease incidence in future times based upon current activity levels, but does not directly incorporate important time-invariant factors in its predictions. Both methods predict disease activity levels for each state in the United States. A socially optimal vaccination policy can be issued based upon the understanding of disease dynamics gained from both modeling techniques. The areas predicted to have especially high disease incidence should be the first to receive vaccine.

3.1 – Statistical Analysis

Statistical analysis provides insight into the characteristics, both network and demographic, that play critical roles influencing disease incidence. However, the observational units of the network data initially did not match those of the disease data. City is the observational unit of the network data, whereas state is the observational unit of the disease data. State figures were computed based upon the state's constituent cities' sum, maximum, average, median, and geometric mean. All five transformations were analyzed in an effort to determine which was not only most representative of the state level observational unit, but also which generates the greatest statistical relationship to influenza incidence.

Now both the independent variable as well as dependent variable observations bear the same units of observation, and statistical analysis is permissible. To determine the pertinent explanatory variables, multiple linear regression was used for its simplicity and ease of interpretation.

3.2 – Dynamic Disease Model

The disadvantage of the statistical models discussed above is that they are time-invariant. Their incidence level predictions remain constant for all time, regardless of the current disease distribution. The inherently static nature of the statistical models necessitates an additional dynamic component. This method attempts to predict the disease incidence level in each state at time *(t + 1)* based upon the incidence level at time *t*. In order to construct such a model, I make the following assertions: disease prevalence for a given state at time *(t + 1)* depends upon disease prevalence at time *t*, the number of infected individuals traveling to the state in question, and the overall growth rate (positive or negative) of disease spread. Applying these assumptions, the dynamic model is proposed below.

Let $\vec{s}_t = (s_{1,t}, s_{2,t}, ..., s_{52,t})$ be the vector of disease activity levels for the United States, where a component $s_{i,t}$ denotes the disease activity level of state *i* at time *t*. Define a transition probability matrix $Q = (Q_{ij})$, where each entry Q_{ij} represents the probability that a passenger travels from city *i* to city *j*. That is, given that an individual traveler resides at *i*, the entry Q_{ij} represents the probability that he or she will travel to *j*. The matrix *Q* is constructed by taking each entry m_{ij} from the mobility matrix $m = (m_{ij})$ and dividing by its row sum. The transition probability matrix corresponding to Table 1 can be found in Table 2 for demonstrative purposes. Let θ_t be the average disease incidence level over the United States at time *t*. Finally, let $\gamma \in [0,1]$ be a parameter for purposes of computing a weighted average. Then the dynamic disease model is the following:

$$\vec{s}_{t+1} = [\gamma \cdot \vec{s}_t + (1-\gamma) \cdot Q^T \vec{s}_t] \frac{\theta_t}{\theta_{t-1}} \qquad (11)$$

This model computes an expected disease incidence vector for period *(t + 1)*, given disease incidence in period *t*. Each component of this vector is the expected value of disease incidence for its corresponding state. Put another way, based upon the disease activity levels of the current period, this model predicts disease incidence next period. The $Q^T\vec{s}_t$ term represents the influence of incoming infected passengers on disease incidence in time *(t + 1)*. Each entry of $Q^T\vec{s}_t$ denotes the expected value of disease incidence as a result of air transport. The transition probability matrix *Q* is transposed so that each entry correctly represents the probability of passenger arrival into *j*, rather than the probability of passenger departure from *i*. Disease activity level in *j* depends on those arriving into *j*, not those leaving *j* to go elsewhere. Ultimately, a weighted average of the expected value of disease activity level due to air travel and the disease activity level in the current period is taken, based upon *γ*. Finally, the factor $\frac{\theta_t}{\theta_{t-1}}$ is the growth rate of disease propagation with respect to the current period and the one before. It acts as a multiplier to inflate or deflate the expected disease vector based upon the dynamic growth rate of disease incidence over time. Through empirical testing and error analysis, I determine *γ* such that error is minimized between the predicted $\vec{s}_{t+1}$ and the actual $\vec{s}_{t+1}$.

		Kodiak	**Homer**	**New York City**	**Pittsburgh**	**Bangor**
		AK	**AK**	**NY**	**PA**	**ME**
Kodiak	**AK**	0	1	0	0	0
Homer	**AK**	1	0	0	0	0
New York City	**NY**	0	0	0	0.852393	0.147607
Pittsburgh	**PA**	0	0	0.999683	0	0.000317
Bangor	**ME**	0	0	0.998311	0.001689	0

Table 2: This transition probability matrix is constructed from the mobility matrix found in Table 1. Each entry represents the probability of traveling from city *i* to city *j*.

4 – Policy Evaluation Methodology

Given a socially optimal vaccination policy determined by the machinery provided above, the fundamental question of "political feasibility" remains. The socially optimal policy might not be ratified by a governing body if it is deemed too inequitable. For example, if too many states[4] receive an insufficient number of vaccines, relative to their constituents' vaccine demands, then the optimal allocation would be voted down. It is important to note that the vaccine supply chain is vastly more complex than simply ratifying a bill, but states acquiring vaccines for further redistribution is a part of it. Thus, such analysis still provides insight into the "political feasibility" or "passability" problem: Given a socially optimal allocation, what is the nearest-to-optimal allocation such that it is ratified by Congress? This politically feasible vaccination policy is a slight reallocation of vaccines from the socially optimal one.

Reallocated vaccines can be viewed as deviations from the socially optimal allocation. Each state has its own corresponding deviation d_i, representing the difference in number of vaccines of the initial optimal allocation and the politically feasible one. Deviations can be either positive or negative. The goal of the political feasibility problem is to minimize the aggregate deviations from the socially optimal vaccination policy. However, there are several constraints as well. First, no state's deviation can be greater than the number of vaccines prescribed by the socially optimal policy. That is, the minimum number of vaccines a state can have is zero. Hence, one cannot reallocate more vaccines away from a particular state than it has to lose. Next, a state's representative, e.g. a senator, will only vote for an allocation if the state's initial

[4] The region under consideration in this research is the state. However, this optimization machinery could be readily applied to a region of any size, in which allocation decisions are made by a representative-voting process.

allocation plus its deviation are greater than or equal to the state's vaccine demand. Constraints of this form have no unreasonable assumptions associated with them. The only assumption is that a representative will approve a policy if his or her demands are satisfied. This forms a well-defined optimization problem:

$$\begin{aligned}
&\min A = \sum_i E(\cdot)_i d_i^2 \\
&s.t. \\
&d_i \leq \alpha_i \\
&\alpha_i + d_i \geq p_i \\
&\& \, d_i \in Z
\end{aligned} \tag{12}$$

where d_i denotes the vaccine allocation deviation of State *i* from the socially optimal policy, α_i denotes the initial socially optimal vaccine allocation to state *i*, p_i denotes the vaccine demand of state *i*, and $E(\cdot)_i$ denotes the expected disease activity level of state *i*.

Deviations are squared in the objective function so that the objective function is nonzero in general. If deviations were linear in the objective function, then the number of vaccines reallocated away from one state and reallocated to another state would always net to zero. Thus, the quadratic deviation terms in the objective function penalize, i.e. increase, the objective value for either a negative or a positive deviation. An objective value of zero means that the socially optimal allocation is also politically feasible.

If the objective function were simply $\sum_i d_i^2$, then any state's deviation would be treated the same. Reallocating a vaccine away from a highly salient state would have the same impact on the objective function as reallocating a vaccine away from a relatively unimportant state. This assumption would be unrealistic. It is far worse to reduce the vaccine supply of the highly salient state. Therefore, it is appropriate to weight each term of the objective function by its corresponding salience. The most intuitive measure of a state's salience is its predicted disease activity level. To incorporate the predictive power of the two frameworks presented above, some weighted average of the predicted levels could be computed. Hence, the objective function is penalized more for a deviation from an important state, and takes the form $A = \sum_i E(\cdot)_i d_i^2$.

All constraints of the form $\alpha_i + d_i \geq p_i$ need not be satisfied. A solution $\vec{d} = (d_1, d_2, ..., d_h)$ is politically feasible if and only if *T* constraints of that form are satisfied, where *T* is the number of votes necessary to pass a bill into law. In a simplified example using the U.S. senate, *T = 26* where senators from the same state are assumed to vote in the same way. Thus, this optimization problem cannot be solved using conventional methods. Classical methods can solve optimization problems in which all constraints must be satisfied. Therefore, the optimization problem presented above must be broken into several solvable problems so that a solution can be reached. Since only 26 constraints of the form $\alpha_i + d_i \geq p_i$ must be satisfied in the United States, construct all possible optimization problems consisting of 26 such constraints that indeed must be satisfied. All possible optimization problems can be constructed by taking all combinations $\binom{50}{26}$ of constraints of the form $\alpha_i + d_i \geq p_i$. Now each problem is in a form in which all constraints must be satisfied, and is thus solvable by classical methods. Each problem has a respective optimal solution. Once all problems are solved, the optimal solution of the original problem is found by determining the solution $\vec{d}$ that satisfies $A(\vec{d}) = \min_{A \in A^*}\{A\}$, where *A** is a minimum objective value to one of the corresponding optimization problems constructed. This solution corresponds to the minimum objective value over all such minima for all $\binom{50}{26}$ optimization problems. Therefore, it is the optimal solution for (12). Thus, the nearest-to-optimal politically feasible allocation policy has been found. Hence, the optimization problem presented in (12) is mathematically decidable, but computationally impractical because of the immensely large number of intermediate constituent optimization problems that must be solved.

The political feasibility optimization problem and solution method not only apply to vaccine allocation, but also to any scarce resource allocation decided by government. There are likely unique methods of determining the socially optimal allocation of the particular resource. Yet even after such a policy has been determined, the issue of political feasibility potentially remains in conflict with social optimality. Thus, (12) and its solution technique apply in order to determine the allocation that is nearest to socially optimal but also politically feasible.

5 – Data and Results

This section details the results of the two modeling methods described in Section 3: statistical modeling and dynamic modeling.

I use a multiple linear regression to determine the salient factors with regard to both disease incidence and propagation. Note that node significance is not included in any regression because it proved to be computationally inefficient to compute. Consequently, no node significance data was available for analysis. I am currently working on refining the implementation discussed in Appendix 8.6.

A regression coefficient represents the change in average influenza activity level on a five point scale, from a one unit change in its corresponding independent variable. First, I included only demographic characteristics as my control variables. Then, I ran a regression with only network variables. Additionally, I used the dynamic model proposed in Section 3 to account for influenza incidence changing with time.

5.1 – Descriptive Statistics

Descriptive statistics for the demographic and network characteristics can be found in Tables 3-8, respectively. Notice that the percentage of the population younger than five and older than sixty-five have means quite different from one another, 6.5% and 13.3% respectively. Further, the mean percentage of the population possessing a bachelor's degree or higher is nearly double that of the retirees, at 27.5%. Network attribute mean comparisons are discussed with respect to averages. Closeness centrality has an extremely small mean of 0.00127, while the other measures are many orders of magnitude larger. Out-degree centrality has a mean of 8,8914.3, nearly double that of betweenness centrality, 4,892.17.

Variable	Obs	Mean	Std. Dev.	Min	Max
FluAvg	52	3.130134	0.406328	2.086957	4.04
VehAvg	52	1.718764	0.17839	0.885344	1.996339
TotPop	52	6009064	6764460	563626	3.73E+07
FamilyAvg	52	3.077692	0.140177	2.83	3.56
PctPoplt5	52	6.513462	0.763128	5.1	9.5
PctPopgt65	52	13.29423	1.671141	7.8	17.4
PctPopBach	52	27.46923	5.646743	17.3	49.2
MedInc	52	51142.06	9455.192	18791	70647

Table 3: Summary statistics for the demographic variables. There are only 52 observations because each state, including Washington DC and Puerto Rico, is one data point.

Variable	Obs	Mean	Std. Dev.	Min	Max
SumofInDeg~y	52	925005.2	1208013	258.4205	5102152
SumofOutDe~y	52	934258	1223175	208.0227	5202040
SumofNetDe~y	52	1859862	2431991	467.209	1.03E+07
SumofClose~y	52	0.025599	0.063293	0.001517	0.4674423
SumofBetwe~y	52	85300.4	192826.3	0	1057671

Table 4: Summary statistics for network attribute sums as representative state figures.

Variable	Obs	Mean	Std. Dev.	Min	Max
MaxofInDeg~y	52	635310.5	776792.1	162.5822	2710146
MaxofOutDe~y	52	642068.5	784706.3	122.4467	2721187
MaxofNetDe~y	52	1277649	1561760	285.1145	5431475
MaxofClose~y	52	0.001515	1.43E-05	0.001414	0.0015169
MaxofBetwe~y	52	62548.77	123867.2	0	513381

Table 5: Summary statistics for network attribute maxima as representative state figures.

Variable	Obs	Mean	Std. Dev.	Min	Max
AverageofI~y	52	88314.3	274135.5	64.60512	1988873
AverageofO~y	52	88914.3	275140.7	52.00568	1995914
AverageofN~y	52	177265.2	549311.6	116.8023	3985048
AverageofC~y	52	0.001265	0.000171	0.000865	0.0015169
AverageofB~y	52	4892.173	13513.94	0	93906

Table 6: Summary statistics for network attribute averages as representative state figures.

Variable	Obs	Mean	Std. Dev.	Min	Max
MedianInDe~y	52	40619.38	275497.4	4.65317	1988873
MedianOutD~y	52	40744.84	276474.5	5.278152	1995914
MedianNetD~y	52	81367.16	552008.2	9.815749	3985048
MedianClos~y	52	0.001459	0.000125	0.000861	0.0015169
MedianBetw~y	52	1842	13019.75	0	93906

Table 7: Summary statistics for network attribute medians as representative state figures.

Variable	Obs	Mean	Std. Dev.	Min	Max
Geo~eCentral	52	39178.41	275683.2	0	1988873
G~OutDegre~a	52	39209.45	276673.8	0	1995914
G~NetDegre~a	52	78524.37	552374.9	0	3985048
Geo~sCentral	52	0.0011	0.000304	0	0.0015169
Geom~sCentra	52	1805.885	13022.42	0	93906

Table 8: Summary statistics for network attribute geometric means as representative state figures.

5.2 – Demographic Regression

The statistical model under consideration is given below:

$$FluAvg = \beta_0 + \beta_1{*}TotPop + \beta_2{*}FamilyAvg + \beta_3 PctPopgt65 + \beta_4{*}MedInc + \varepsilon \quad (13)$$

The small number of observations and potential for multi-collinearity required that several demographic variables be excluded from the model. The results for this regression are reported in Table 9. The symbols *,**,*** indicate significance at the 10%, 5%, and 1% levels, respectively.

The results in Table 9 suggest that a state's total population, average family size, and median income all contribute to a State's influenza activity level. Contrary to the results found in Shaw et. al (2010), larger populations are found to increase disease incidence, rather than decrease it. Shaw et al. justify their findings by arguing that a small group of infected individuals forms a greater proportion of a small population as compared to the same small group in a large population. This finding is thought to explain why disease spreads more quickly within smaller populations. However, my contrasting finding is intuitive because larger populations likely contain more susceptible individuals by nature of population composition. Further, there is greater potential for more social interactions within a large population. Thus, an infected individual has more opportunity to spread disease and ultimately infect more people. This process is assumed to occur iteratively. The results in Table 9 show that a state with a population 10 million people greater than another state is expected to have an average influenza activity level 0.245 points higher on a scale from 1 to 5.

Additionally, the positive effect of median income on disease incidence agrees with theory if the time-frame is before vaccines have been distributed. A state with median income $10,000 higher than another state is expected to have an average disease incidence 0.198 points higher on a scale from 1 to 5. Wealthier individuals are more likely to travel and engage in numerous social interactions. Thus, if infected, wealthier individuals are more likely to spread disease. However, if the time-horizon is during or after vaccine-distribution, then it would make intuitive sense to assume that wealthier individuals are the least effective vehicles of disease spread. As they have access to better healthcare, they should be able to afford medicine and vaccines. Then increases in median income would lead to expected decreases in disease incidence level. In this situation, the median income relationship found is inconsistent with theory.

The effect of average family size on disease incidence acts in contrast to my expectations. It would be sensible to assume that larger families are more likely to contract influenza, as more individuals engage in more social interactions and can then spread the disease contracted throughout the household. However, single individuals may travel short distances within and across states more often than those who are part of a larger family. Thus, areas with high singles populations may be expected to have higher disease incidence levels, resulting from additional short-distance travel. This explanation would validate the statistically significant negative coefficient found for average family size. Further, the negative effect of the proportion of individuals older than sixty-five years of age on disease incidence also acts in contrast to theory's predictions, as one might predict that older people are more susceptible to influenza. However, the coefficient on this variable is not statistically significant at any conventional level.

FluAvg	Coef.	Std. Err.	t	P>t
TotPop/10^{-8}	2.45	.798	3.07	0.004***
FamilyAvg	-1.521163	0.480806	-3.16	0.003***
PctPopgt65	-0.0378434	0.037014	-1.02	0.312
MedInc/10^{-5}	1.98	.549	3.6	0.001***
_cons	7.15604	1.78466	4.01	0

Table 9: Multiple regression results for demographic characteristics. All coefficients are statistically significant at the 10%, 5%, or 1% levels.

5.3 – Raw Network Attribute Regression

Observe that closeness centrality values are extremely small with little variation, while the rest of the network attributes have very large values with great variation. As seen from Tables 10-12, this trend is not sensitive to the aggregation technique.

A regression was run for each of the state-aggregation techniques. Unsurprisingly, all three degree centrality measures were perfectly correlated with one another for all state aggregated figures, namely sum, maximum, average, median, and geometric mean. For all state aggregated figures, closeness centrality was found to be weakly correlated with all degree centrality measures. Betweenness centrality was found to be relatively correlated with all three degree centrality measures for the summation and maximum state representative figures. However, betweenness centrality was found to be nearly perfectly, or perfectly, correlated with all three degree centrality measures for the remaining state representative figures of average, median, and geometric mean.

Network attributes were expected to have positive effects on disease incidence. That is, an increase in a network variable was expected to increase disease activity level. While I found that several network coefficients bear the sign theory would predict, the standard error on each tended to be large. Across all regression models, betweenness centrality and degree centrality seemed to have the greatest explanatory power. The state-aggregation techniques that indicated to the most statistically significant relationships were sum, maximum, and average. The network characteristic regression results can be found in Tables 10-12.

FluAvg	**Coef.**	**Std. Err.**	**t**	**P>t**
SumofBetwe~y/10^{-7}	4.78	2.90	1.65	0.106
_cons	3.089364	0.060703	50.89	0

Table 10: Regression results for summation as state representative figures. This is the most informative regression of both simple and multiple linear regressions run involving summation data.

FluAvg	**Coef.**	**Std. Err.**	**t**	**P>t**
MaxofBetwe~y/10^{-7}	7.36	4.52	1.63	0.11
_cons	3.084076	0.062249	49.54	0

Table 11: Regression results for maxima as state representative figures. This is the most informative regression of both simple and multiple linear regressions involving maxima data.

FluAvg	**Coef.**	**Std. Err.**	**t**	**P>t**
AverageofB~y/10^{-5}	2.25	1.64	1.37	0.178
AverageofN~y/10^{-7}	-6.44	4.04	-1.59	0.118
_cons	3.134365	0.059723	52.48	0

Table 12: Regression results for averages as state representative figures. This is the most informative regression of both simple and multiple linear regressions involving averages data.

5.4 – Dynamic Model Results

The dynamic model proposed in Section 3.2 was constructed in Microsoft Office Excel 2010. Recall the dynamic disease model:

$$\vec{s}_{t+1} = [\gamma \cdot \vec{s}_t + (1-\gamma) \cdot Q^T \vec{s}_t]\frac{\theta_t}{\theta_{t-1}} \tag{11}$$

where $\vec{s}_t = (s_{1,t}, s_{2,t}, ..., s_{52,t})$ is the vector of disease activity levels for the United States at time *t*, and $Q = (Q_{ij})$ is the transition probability matrix. Disease vectors at time *t* input into the model came from the Center for Disease Control 2011 data.

In order to maximize the predictive power of the model, the parameter γ was varied from 0 to 1 in increments of 0.1 for four adjacent periods. I sought to find the γ for which predictive error was minimized, where error is defined to be the sum of squared error for each state totaled over the four periods. Error was minimized for $\gamma = 1$, implying that the transportation component of the dynamic model is irrelevant, and that future disease incidence in a state depends exclusively upon current disease incidence in that state.

This finding makes no intuitive sense, which prompted thorough investigation of the Center for Disease Control data. Not only does the data exhibit little variation from state to state by nature of the measurement scale, but also exhibits little to no variation across several adjacent weeks. Since the data was inadequate for analysis, no results are provided here. The lack of variation explains the nonsensical finding that the best prediction of disease incidence in period $(t + 1)$ is to predict the exact same disease incidence in period t. A more thorough exploration of the measurement error and general problems with the Center for Disease Control data is found in Section 6.1. The dynamic model would likely be best applied to a data set in which the observations are estimations of the number of influenza cases by state. Such a data set would exhibit more variation from week to week and from state to state.

6 – Discussion

6.1 – Data Limitations

While the statistical and dynamic methods presented above are sound, the influenza data proved to be extremely imprecise. The data exhibited little to no variation across time within states[5], and minor variation between states within a particular week. Recall that disease data was obtained from Center for Disease Control weekly influenza activity maps. The disease incidence scale is only measured in five broad categories, which by nature do not allow for much variation in the influenza data. The five disease activity levels are too general to provide any meaningful comparisons between states or weeks. For example, two states may both be classified as disease category 4, but in reality differ greatly from one another regarding disease incidence. Under the current measurement system, one state might truly be a category 4.2 and the other a category 4.9. Yet both are expressed as category 4. Further, if a given state has been at disease activity level 5 for four consecutive weeks, there is no way of determining whether disease is becoming more or less prevalent there over time. The broad nature of the disease activity level only allows for massive changes in disease incidence to be detected over time.

There are numerous measures of influenza incidence that are more informative than the one published in the Center for Disease Control weekly activity maps. The ideal measure would be the estimated number of cases or deaths within a state. Such a measure would exhibit the true concept of disease incidence, and represent the exact impact on a particular state at a particular time. An alternative measure more along the order of the current Center for Disease Control measure would be the percent of localities in a particular state infected beyond a certain threshold of infected individuals. While such data would exhibit measurement error, it is much more representative of disease incidence than broad categories.

[5] After I completed my analysis, I was alerted by the CDC of some errors in the data. Through email correspondence, the CDC specifically informed me that influenza maps have not always been updated on a weekly basis.

Next, the observational unit of the Center for Disease Control is the state, rather than the city. Smaller observational units lend themselves to more precise analysis. The size of the observational unit limits the analysis of disease dynamics to the unit itself, and larger. Thus, any insight into socially optimal vaccination policies that empirical analysis at the city level could provide is impossible, given the existing public influenza data. Hence, if influenza data were available in some form at the city level, more accurate statistical and dynamic modeling would be possible. City level data would allow for more observations regarding statistical modeling. Additionally, it would allow for disease vectors of more components, regarding dynamic modeling. Therefore, more precise error analysis could be conducted, making for a more accurate estimation of γ.

If more precise influenza data were available, the same statistical specifications could be applied, potentially resulting in statistically significant network coefficients. Further, the same dynamic model could be applied to influenza data containing representative variation. Then the true role of air transportation on disease propagation could be more reliably tested. It is imperative that the statistical and dynamic modeling techniques be applied to influenza data sets with minimal measurement error.

6.2 – Vaccine Allocation Policy

The statistical and dynamic frameworks presented in this research provide modes for understanding heterogeneous multi-region disease dynamics and, ultimately, how to rank areas by importance for vaccination. Areas that are expected to not only have high disease incidence, but also lead to disease propagation should be allocated vaccines first. Given the scarcity of vaccines, areas must be ranked by expected incidence level.

The statistical and dynamic approaches each offer distinct ranking criteria. The statistical models proposed take relevant regional characteristics into account for ranking, without any regard to time or current disease incidence. The dynamic model proposed takes time and current disease incidence into account for ranking, without any regard to relevant regional characteristics. Both methods predict disease incidence levels, and could be readily applied to data sets of the preferred type discussed in Section 6.1. Each addresses the shortcomings of the other. Thus, an informed policy decision must necessarily incorporate both approaches. Practically speaking, this means that policy-makers must take all facets of disease propagation into consideration. Mathematically speaking, one way to rank regions would be by the following decision rule:

$$a_f = \psi \cdot a_s + (1-\psi)a_d \tag{14}$$

where a_f, a_s, and a_d denote the final, statistically predicted, and dynamically predicted disease incidence levels, respectively, and $\psi \in [0,1]$ is a parameter representing the policy maker's weighting of the two predicted activity levels. The policy-maker would then proceed to rank regions based upon the activity level generated by (14).

However, the decision rule is not complete without a designated number of vaccines to be allocated to an area of high rank. If a region is selected for vaccination, the number of vaccines appropriated should be sufficient to curb the epidemic within the given region's population. Anderson and May (1991) demonstrated that this figure is exactly the critical proportion of vaccination:

$$p_c = 1 - \frac{1}{R_0} \tag{15}$$

where R_0 is the basic reproduction number, defined in Section 1.1. This proportion is based upon a single-population SIR model, which is discussed at length in Appendix 8.1.

For simplicity, the critical proportion is assumed to be uniform across all regions. To challenge this assumption, precise measurements of R_0 in each region would be necessary. While the proportion is uniform, the number of vaccines allocated to any two selected regions will be different because of the difference in population sizes.

Now, a complete decision process for socially optimal vaccination policy can be defined:

1. Rank regions[6] by importance for vaccination by their predicted incidence levels.
2. If a region is selected[7] for vaccination, vaccinate it level to p_c.
3. Return to step 1 and continue allocating vaccine until the supply is exhausted.

The dynamic element of this decision process integrates well with the complexity of the vaccine supply chain. All vaccine doses for a given season are not available at one time. Therefore, policy decisions for allocation cannot be made only once, but must be made several times over the course of a season. As new supplies of vaccines become available for distribution, the decision process can be run. Each time allocation decisions are made, the rankings can be updated based upon current disease prevalence.

The methods presented in this research can be readily applied to determining socially optimal vaccination policies with respect to contagious diseases other than influenza. The exact relationships between various factors and influenza incidence may not hold for some other infectious disease, but the same methods of analysis can be applied to determine the case-by-case relationships. Similarly, the optimization framework presented in Section 4 can also be applied to determining politically feasible,

[6] "Region" refers to the observational unit of the data. In this research, the CDC data limited me to the state, rather than the city. But the decision process most aptly applies to cities.

[7] The methods discussed could be applied to an observational unit of any size, depending on the data. As discussed in Section 6.1, the ideal data would be at the city level. Then every state would likely receive vaccines.

yet nearly socially optimal, allocations of any scarce resource. Therefore, the above decision process can be revised:

1. Rank regions by importance for vaccination by their predicted incidence levels.
2. If a region is selected for vaccination, vaccinate it level to p_c.
3. Return to step 1 and continue allocating vaccine until the supply is exhausted.
4. Apply (12) to reallocate some vaccines to reach a politically feasible solution.

But in order to apply the optimization technique, some socially optimal benchmark must first be declared. As discussed in Section 4, the solution technique proposed is intuitively appealing, but algorithmically impractical.

7 – Conclusion

This research provides novel analytical methods for determining socially optimal vaccination policies. The methods proposed capture both the time-invariant and time-variant components of disease propagation. Additionally, the techniques presented incorporate both demographic and network characteristics, as opposed to exclusively one or the other.

The computational and algorithmic methods proposed for computing the centrality measures themselves solve several algorithmic network problems. Also, a new and precise centrality measure, node significance, was proposed. The reader is directed to Appendices 8.2-8.6 for detailed descriptions of network centrality algorithms. The testing and comparison of these various network attributes to one another with respect to disease spread has not been done before. The ability to compute these centrality measures enables future research in the areas of epidemiology and network analysis.

This research proposes novel methods and machinery that can be readily applied to a myriad of problems in both network analysis and resource allocation. Regarding network analysis, any centrality measure can be tested for relevance to a diffusion process along a network using the statistical methods discussed. Additionally, any time-variant diffusion process can be analyzed through the dynamic methods discussed. Further, the optimization problem proposed in (12) can be applied to resource allocation problems, especially when the resource is scarce. Yet more efficient algorithms for solving (12) must first be developed. The machinery developed in (12) can also be extended to resource allocation problems facing any decision-making entity, from a non-profit organization to a for-profit corporation.

8 – Appendices

Appendix 8.1 describes the single-population SIR model indirectly used in this research. Appendices 8.2-8.7 describe the algorithms implemented for network attributes.

Appendix 8.1 – SIR Model

To model disease propagation within a particular state, I use an SIR ordinary differential equations model in which infected individuals can travel. The model consists of the standard disease compartments: Susceptible, Infected, Removed. Intra-state disease dynamics are described by the following system:

$$\frac{dx}{dt} = -\beta xy \tag{15}$$

$$\frac{dy}{dt} = \beta xy - \gamma y \tag{16}$$

$$\frac{dz}{dt} = \gamma y \tag{17}$$

where $x(t), y(t)$, and $z(t)$ are respectively, the susceptible, infective, and removal functions of time, and β and γ are the infection and removal parameters respectively. Note that $\beta, \gamma > 0$. I assume that the entire population of a given state is of fixed size N for the duration of any influenza season and that individuals mix homogeneously. The three equations below describe the disease compartment distribution for all time. The entire population of any state can be represented by these mutually exclusive disease compartments for all time, as seen in (18). Equation (19) specifies the same relationship at the state level. Finally, I assume that the aggregate population of all states under consideration is equal to the total United States population, as seen in (20).

$$x + y + z = N \tag{18}$$

$$x_i + y_i + z_i = n_i \text{ for all } i \tag{19}$$

$$\sum_i n_i = N \tag{20}$$

where n_i is the population of state i, and x_i, y_i, and z_i are the susceptible, infected, and removed populations of state i respectively. I also assume all state populations to be fixed for the duration of any influenza season.

This model differs from the SAIR model presented in Shaw et. al. (2010) in that it does not include an asymptomatic disease compartment. My model is the standard SIR model, and consequently assumes that removal is the terminal disease state. That is, once an individual leaves the susceptible compartment he cannot return to it and once an individual enters the removed disease compartment he cannot leave it. This model more aptly reflects a time horizon of a given influenza season because compartmental change from removal would not change until the following season. The transitions between states can be seen in Figure 1.

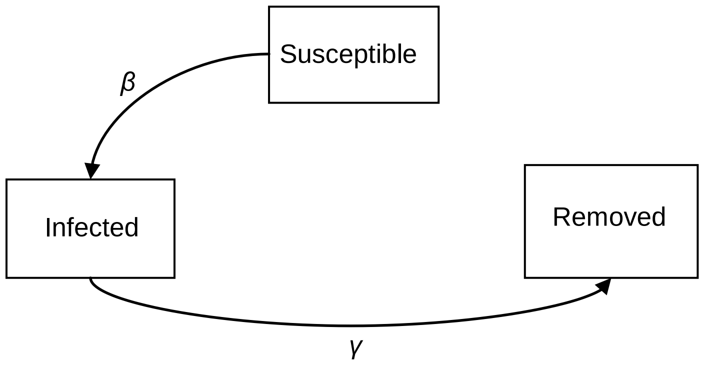

Figure 1: SIR state diagram. The infection parameters can also be viewed as probabilities of transitioning from one disease compartment to another.

Appendix 8.2– Degree Centrality

The degree centrality program requires determining two parameters and one input value: *Δα*, the size of the mobility matrix *m*, and the mobility matrix *m* itself, respectively. These parameters and inputs will be discussed below. The degree centrality measure has two components: the number of edges incident to a node *i*, and the cumulative weight of those edges. The input to the Degree Centrality program is the mobility matrix *m*. Without loss of generality, I discuss the algorithm for computing out-degree centrality in this appendix. The method for computing in-degree centrality of a node *i* is identical except that it involves a scan of column *i*, rather than row *i*.

To compute the out-degree centrality of a node *i*, scan through row *i* of matrix *m*. A zero entry implies that there is no edge from *i* to *j*. Each nonzero entry in row *i* represents an edge originating at node *i* and terminating at some other node *j*. Thus, to compute the number of edges, sum the number of nonzero entries in row *i*. Every nonzero entry detected in row *i* adds one to the running total of the out-degree edge counter for node *i*. Simultaneously, the program computes the cumulative out-degree edge-weight of node *i*. Each time the scan of row *i* detects a nonzero entry *(i,j)*, the value of the entry itself is added to the running out-degree edge-weight subtotal. Hence, when the row scan is complete, the program has computed the number of edges incident to node *i* as well as the cumulative edge-weight of edges originating at node *i*. Now, the program can substitute this information into equation (3).

The only remaining component of degree centrality left unspecified is the parameter α. For the purposes of this research, it is sensible to declare a node more central than another with the same cumulative edge-weight if the former has more neighbors. Therefore, the tuning parameter $\alpha \in (0,1)$, as discussed in Opsahl et. al. (2010). However, there is no particular value of $\alpha \in (0,1)$ that can be justified as the correct tuning parameter any more than any other. Thus, it is imperative to compute a value for degree centrality that is representative of the open interval (0,1). This is done by computing a degree centrality value for a node *i* for each value of α in (0,1) in increments of $\Delta\alpha$, which is a parameter required to be set before running the program. The degree centrality value for each α from $[0 + \Delta\alpha, 1 - \Delta\alpha]$ is computed so that the interval is closed rather than open. Otherwise, the program would be unable to compute the infinite number of values in the open interval (0,1). Once a degree centrality value has been computed for all values of α in $[0 + \Delta\alpha, 1 - \Delta\alpha]$ in increments of $\Delta\alpha$, the program takes an average of these values. This average degree centrality value is representative of the $\alpha \in (0,1)$ because a node that is frequently central for many choices of α should be treated as a highly central node, in terms of degree centrality.

The program performs the algorithm discussed above for each node $i \in V$, where *V* is the Vertex set. Recall, there are 1060 nodes in the vertex set of the air transportation network under consideration.

Appendix 8.3– Dijkstra's Algorithm and Shortest Path Construction

In order to compute the remaining three network attributes, namely closeness centrality, betweenness centrality, and node significance, shortest paths between all pairs of nodes *s* and *t* in the network must be established. Shortest paths between pairs of nodes are computed using Dijkstra's Shortest Path Algorithm (Dijkstra, 1959). Note that Dijkstra's Algorithm does not check for the uniqueness of shortest paths. Theoretically, there might exist several shortest paths from a source-node *s* to a terminal node *t*. Shortest path computation stops once one such shortest path is found. Hence, this implementation potentially does not account for all shortest paths in the network, as some may not be unique. I refer to this as the "shortest path uniqueness problem." The number of shortest paths in the network unaccounted for is negligible because of the size of the network and the variation in passenger flows.

One at a time, each node in *V* is designated the source node *s*. For each node *s*, Dijkstra's algorithm outputs a distance array $[d_1, \ldots, d_n]$ as well as a predecessor array $[p_1, \ldots, p_n]$. A component d_j of the distance array denotes the shortest path distance from the source node *s* to node *j*. Recall that all entries of the original mobility matrix *m* have been inverted. Hence, the shortest path distance from *s* to *j* is the sum of the reciprocals of the constituent elements of *m*, which represent the edges in the shortest path. Put another way, the shortest path distance from *s* to *j* is the sum of inverted passenger flows of intermediate edges in the path. Therefore, the maximum weight a particular edge can have, post-inversion, is 1. For example, if $d_j = 5$, the shortest path distance from *s* to *j* is 5. The distance d_s from *s* to itself is always zero.

It is convention in the field of network analysis to say that s is distance infinity from j if there does not exist a path from s to j. However, applying this convention would homogenize any centrality measure involving distance, thereby rendering the measure meaningless. If nodes j and k have no path to some other node in the network, both j and k will have distance infinity as a component of any centrality measure dependent on distance. Thus, both j and k would have a centrality measure of either ∞ or $1/\infty = 0$, depending on the measure under consideration. However, the other components of the centrality measure may differ greatly between j and k. But those differences are not reflected because they are "overpowered" by the infinity term. Thus, it is imperative to develop a notion for the "infinity distance" that addresses this homogenization problem. The infinity distance must therefore be a real number, but not so large that the homogenization problem happens anyway. Yet it must be sufficiently large so as to penalize a node's centrality measure more than a long path distance would. Additionally, the infinity distance should be based upon the network itself. I designate the infinity distance to be $d_\infty = \mu \cdot D$, where μ is a parameter set before running the program, and $D = \max_{ij} d_{ij}$, the longest length of any finite shortest path in the network. Since the infinity distance must necessarily be greater than the length of any shortest path in the network, $\mu > 1$. Depending on the network, one may have to experiment with varying values of μ in order to output a meaningful distance-dependent centrality measure. For purposes of this research, μ was set to 2. I discuss the rationale in Appendix 8.7. Finally, it is important to note that the infinity distance was not determined recursively. In order to compute $\mu \cdot D$, finite distances must exist for all pairs of nodes in the network. Algorithmically, if there does not exist a path from i to j, a distance $d_{ij} = -1$ is designated in the predecessor array. Then, to determine *D*, the program scans the distance array and locates the largest value. This value must be the greatest length of a finite shortest path in the network because all "infinite" distances have been set to -1.

A component p_j of the predecessor array denotes the predecessor node to *j* in the shortest path from *s* to *j*. For example, if $p_j = 8$, node number 8 is the node that comes before node *j* in the shortest path. Since predecessor array values are node numbers, they must be integers. A -1 in the predecessor array indicates that either that spot in the predecessor array is the source node, or that there is no path from *s* to *j*. If $p_j = -1$, then there is no path from *s* to *j*. To determine the path from *s* to *j*, we use the predecessor array to trace the path backwards from *j* to *s*. In the above example, if $p_j = 8$, the next iteration of path tracing is to go to component p_8 in the predecessor array. This is done iteratively until a -1 value is reached, indicating the path from *s* to *j* has been completely traced backwards from *j* to *s*.

Appendix 8.4 – Closeness Centrality

Closeness centrality is entirely dependent on shortest path distance. To compute the closeness centrality of a node *i*, the program sets *i* equal to the source node *s* in the Dijkstra framework. Dijkstra's algorithm now outputs a distance array with respect to the node *i*=*s*. This array represents the distance from *i* to all other nodes in the network. Summing all components in the array $d_{i1} + d_{i2} + ...d_{in} = \sum_{j=1}^{n} d_{ij}$ is the cumulative distance from *i* to all other nodes in the network. The closeness centrality of *i* is simply the reciprocal of this value. Note that closeness centrality is solely dependent upon shortest path distance, irrespective of the actual number of shortest paths. Hence, closeness centrality is unaffected by the shortest path uniqueness problem discussed in Appendix 8.3.

Appendix 8.5 – Betweenness Centrality

Betweenness centrality is not directly dependent on distance, but on paths themselves and their constituent nodes. Thus, the predecessor array, rather than the distance array, is used in the betweenness centrality program. The program constructs an output matrix that keeps track of the betweenness centrality of each node. As the algorithm performs, a running subtotal of the betweenness centrality of each node is maintained. A given predecessor array is with respect to a particular $s \in V$, and specifies the nodes in paths to all other $t \in V$. The program iterates on *s* from 1 to 1060, performing the operations described below for each $s \in V$. First, the predecessor array with respect to *s* is scanned. If there is no path from *s* to that particular *t*, then the program places a null value in component *t* of the predecessor array. This null component indicates that there is no path from *s* to *t.* If the node number of source node *s* is in component *j* of the predecessor array, then there is a direct edge from *s* to *j*. Further, this direct edge is the shortest path from *s* to *j*. If there is a node number for some node $w \neq s$ in component *j* of the predecessor array, then *w* is an intermediary node in the path from *s* to *j*. Now, add one to the running betweenness centrality subtotal for *w*. Every time a node *w* is detected as part of a path from some *s* to some *t*, the betweenness centrality subtotal for *w* is increased by one. This process is conducted by scanning through the shortest paths from *s* to all *t* in the predecessor array. Within a given predecessor array, a particular *t* is specified one at a time. The path from *s* to *t* is traced backwards and the betweenness centrality output matrix of all nodes *w* in the path from *s* to *t*, σ_{st}, is updated until node *s* is reached in the predecessor array. Now, for the same *s*, the program iterates on *t* until all *t* from 1 to 1060 have been investigated. Once this process is completed for all nodes *t*, *s* is increased by one and the process is repeated until all *s* from 1 to 1060 have been investigated. Note that shortest paths to all *t* are investigated for each *s*, and there is a predecessor array generated and scanned for each *s*. Thus, all possible node pairs *(s,t)* have been checked. Note that betweenness centrality is affected by the shortest path uniqueness problem discussed in Appendix 8.3.

Appendix 8.6 – Node Significance

The node significance program uses two adjacency matrices, multiple stacks, and an output matrix for its computations. The adjacency matrix elements are edge weights, which are passenger flows, post-inversion. One of the adjacency matrices, m_f, will be fixed and the other, m_c, changeable. The relationship between the two matrices is described below. All computations necessary for node significance are performed simultaneously with those for betweenness centrality. The node significance program computes based upon a running subtotal for each node. The node significance of all nodes is initialized to be zero in the node significance output matrix. First, all negative terms of the summation are found and added together. Then, all positive terms of the summation are found and added to the running subtotal, resulting in the node significance for node *i*.

Stacks are formed for each node $i \in V$, based upon the betweenness centrality output matrix. Initially, the stack corresponding to each node is empty. When *i* is detected to be in the shortest path from some *s* to some *t*, that particular pair of *s* and *t* are added to the top of the *i*th stack. When *s* and *t* are added to the *i*th stack, the program calls the distance array corresponding to the node on the top of the stack, denoted s_1 as it is in the first source-node position on the stack. This distance array tells us the distance from s_1 to all other nodes in the network. The t_1 component of the array, d_{t1}, is the shortest path distance from s_1 to t_1 in the original network *G*. Thus, the negation of this value is the negative part of the summation term for the path from s_1 to t_1. The positive part of this term is computed after all negative terms of the summation have been computed. The value $–d_{t1}$ is then added to the running subtotal for $c_s(i)$. This process is done iteratively for all $s,t \in V$.

Now, all negative terms of the node significance summations have been found. The current node significance values for any node $i \in V$ are either negative or zero. If the original stack size for node *i* was zero, then the node significance of *i* is zero because *i* is never part of a shortest path from some *s* to some *t* in the network. Thus, its removal from the network does not penalize the length of any shortest path. All nodes for which the *j*th stack was nonempty currently have a negative value for node significance. The positive terms must be computed and added to the summation in order to compute accurate node significance values. Now, the program scans all stacks in order of node number, from 1 to 1060. If a stack is empty, then the program continues scanning onto the next stack. If the *i*th stack is nonempty, the *i*th row and *i*th column of the changeable adjacency matrix m_c are set to zero. This “zeroing” of the *i*th row and column effectively removes *i* from the network because all edges incident to *i* are removed by replacing their weight with zero in m_c. Denote this matrix as m_c’. Hence, node *i* is now isolated in that there is no path from or to *i* from any other node. Dijkstra's algorithm is then applied to m_c'. Now, the distance array corresponding to the source node *s* on top of stack *i*, denoted s_1, is viewed. The distance array generated by Dijkstra's algorithm represents the distance from s_1 to all other nodes in the network, given that node *i* has been removed. The element d_{t1} of that distance array, corresponding to the terminal node on top of stack *i*, is the positive term of the node significance summation from s_1 to t_1. This term is added to the running subtotal for *i* in the node significance output matrix. Once this process is complete, s_1 and t_1 are removed from stack *i*. The process is repeated for s_2 and t_2, the new top of stack *i*. If at this point, $s_2 = s_1$, remove s_2 from the stack and check the associated t_2 on the stack. Since $s_2 = s_1$, the same distance array generated with respect to s_1 can be used to determine the positive term to be added to the node significance output matrix for *i*. Now, the element d_{t2} will be added to the node significance of *i*. Note that t_2 is necessarily different than t_1. The program iterates until stack *i* is empty. Then it proceeds to the next stack, *(i+1)*, iteratively. When all stacks have been checked, the node significance for each node in the network has been computed. Note that node significance is solely dependent upon shortest path distance, irrespective of the actual number of shortest paths. Hence, node significance is unaffected by the shortest path uniqueness problem discussed in Appendix 8.3.

Appendix 8.7 – Sensitivity Analysis of Network Attribute Parameters

Before any statistical analysis could be performed on the network attributes, the attributes themselves had to be accurately computed. The exact values of the network attributes degree centrality[8] as well as closeness centrality are affected by the choice of the parameters $\Delta\alpha$ and μ, respectively. Thus, I performed a sensitivity analysis for each measure to determine how the rank of a given city was affected based on the parameter change. It is important to note that for these sensitivity analyses no aggregation technique was used. The raw network data at the city level was analyzed. Hence, I am confident in the results of these analyses. All cities were ranked in descending order based upon their in-degree centrality, out-degree centrality, and net-degree centrality for $\Delta\alpha = 0.01, 0.001$. That is, the city of highest degree centrality was ranked number 1, the next highest ranked number 2, and onwards. Cities were also ranked in descending order based upon their closeness centrality for $\mu = 2,3$.

Next, I found the correlation between the rankings corresponding to each parameter of each of the degree and closeness centrality measures. The results of these sensitivity analyses can be found in Table 13. As seen in Table 13, the centrality measure rankings are not sensitive to changes in parameters. All rankings for a particular measure are nearly perfectly correlated, indicating that virtually no rankings change based upon a parameter change. Consequently, for precision $\Delta\alpha = 0.001$ was selected and for heterogeneity $\mu = 2$ was selected for purposes of the statistical analysis of this research. A small $\Delta\alpha$ ensures a precise measurement of degree centrality, as more outputs are computed and averaged to form a representative figure. A small μ ensures that a large, penalizing distance is assigned to pairs of nodes for which there exists no path, but not too large. If too large a distance had been designated the “infinity distance,” then the

[8] All degree centrality measures are affected by the choice of α, namely in-degree centrality, out-degree centrality, and net-degree centrality.

closeness centrality of most cities would be nearly identical. If this were the case, then the closeness centrality measure would hold no meaning.

Centrality Measure	**Correlation**
In-degree	0.99999737
Out-degree	0.99999713
Net-degree	0.99999681
Closeness	0.99999562

Table 13: Correlations between rankings corresponding to $\Delta\alpha = 0.01, 0.001$ for each degree centrality measure, and rankings corresponding to $\mu = 2,3$ for the closeness centrality measure.

Acknowledgments

I would like to thank these individuals and groups for their contributions to my thesis: Foremost, my advisor, Stephen Wu, for his time and unwavering support throughout this project – our stimulating discussions allowed this thesis to reach its full potential. The Hamilton College Economics Department, who funded the creation of the network centrality programs. Lindsay Shankman and Julian Aronowitz, whose network centrality suite allowed me to pursue a critical component of this project. Hamilton College President Joan Hinde Stewart and Dean of Faculty Patrick Reynolds, whose generosity made my West Point Network Science conference presentation a reality. Finally, my parents Lawrence and Freda Boguchwal, for their continued support throughout this endeavor.

References

[1] Anderson, R. & May, R. (1991). *Infectious Diseases of Humans*. Oxford Oxfordshire: Oxford University Press.

[2] Balcana, D., Colizza, V., Goncalvesa, B., Hud, H., Ramascob, J., & Vespignani, A. (2009). Multiscale mobility networks and the spatial spreading of infectious diseases. *PNAS 106*(51), pp. 21484–21489.

[3] Bernoulli, D. (1766). Essai d'une nouvelle analyse de la mortalité causée par la petite vérole, et des avantages de l'inoculation pour la prévenir. *Memoirs of the Royal Mathematical and Physical Academy of Sciences*, pp. 1-45.

[4] Center for Disease Control (CDC) (2011), CDC – Influenza Weekly Activity Maps. Retrieved from http://www.cdc.gov/flu/weekly/usmap.htm.

[5] Center for Disease Control (CDC) (2011), CDC – Influenza Weekly Overview. Retrieved from http://www.cdc.gov/flu/weekly/overview.htm.

[6] Colizza, V., Barrat, A., Barthe´lemy, M., & Vespignani, A. (2006). The role of the airline transportation network in the prediction and predictability of global epidemics. *PNAS*, *103*(7), pp. 2015-2020.

[7] Mollison, D. (Ed.) (1995), *Epidemic Models: Their Structure and Relation to Data*. Cambridge, England: Cambridge University Press.

[8] Daley, D. & Gani, G. (1999). *Epidemic Modelling*. Cambridge: Cambridge University Press.

[9] Dijkstra, E.W. (1959). A note on two problems in connexion with graphs. *Numerische Mathematik, 1(1)*, pp. 269–271.

[10] Frauenthal, J. C. (1980). *Mathematical modeling in epidemiology*. New York: Springer-Verlag.

[11] Galvani, A., Reluga, T., & Chapman, G. (2007). Long-standing influenza vaccination policy is in accord with self-interest but not with the utilitarian optimum. *PNAS*, *104*(13), pp. 5692-5697.

[12] Golbeck, J. (n.d.). *Network centrality*. Retrieved from http://www.cs.umd.edu/~golbeck/CMSC498N/blog/3.2.pdf

[13] Guimerá, R., Mossa, S., Turtschi, A., & L.A.N (2005). The Worldwide Air Transportation Network: Anomalous Centrality, Community Structure, and Cities' Global Roles. *PNAS 102(22)*, pp. 7794-7799.

[14] Heal, G., & Kunreuther, H. (2005). "The Vaccination Game." Working paper

[15] Hyman, J. M., & LaForce, T. (2001). Multi-city sir epidemic model. Los Alamos National Laboratory, U.S. Department of Energy. Working paper.

[16] Jackson, M. (2008). *Social and Economic Networks*. Princeton: Princeton University Press.

[17] Kermack, W. O., & McKendrick, A. G. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, *115*(772), pp. 700-721.

[18] Mark, S., & Rushton, S. P. (2005). The impacts of network topology on disease spread. *Ecological Complexity 2*(3), pp. 287-299.

[19] Opsahl, T., Agneessens, F., & Skvoretz, J. (2010). Node Centrality in Weighted Networks: Generalizing Degree and Shortest Paths. *Social Networks 32*, pp. 245–251.

[20] Research and Innovative Technology Administration (RITA), *Bureau of Transportation Statistics (BTS),* T100 Domestic Segment Data. (2011). Retrieved from http://www.transtats.bts.gov/DL_SelectFields.asp?Table_ID=259&DB_Short_Name=Air%20Carriers

[21] Salathé M, Jones J.H. (2010). Dynamics and Control of Diseases in Networks with Community Structure. *PLoS Computuational Biology 6(4)*, pp. 1-11.

[22] Sattenspiel, L. (2009). *The geographic spread of infectious diseases models and applications*. Princeton: Princeton University Press.

[23] Shankman, L., Aronowitz, J. (2012). *Network Attribute Suite.* C++ program.

[24] Shaw, L., Spears, W., Billings, L., Maxim, P. (2010). Effective vaccination policies. *Information Sciences*, pp. 1-17.

[25] United States Census Bureau, (2010). *American Factfinder.* Retrived from http://factfinder2.census.gov/faces/nav/jsf/pages/index.xhtml

[26] World Health Organization, (2009). *Influenza (seasonal)*. Retrieved from www.who.int/mediacentre/factsheets/fs211/en/

[27] Wu, J., Riley, S., Leung, G. (2007). Spatial considerations for the allocation of pre-pandemic influenza vaccination in the united states. *Proceedings of the Royal Society Series B, 274*, pp. 2811-2817.