The Spread and Persistence of Infectious Diseases in Structured Populations*

LISA SATTENSPIEL
Department of Anthropology, University of Missouri, Columbia, Missouri 65211

AND

CARL P. SIMON
Departments of Mathematics, Economics, and Public Policy
University of Michigan, Ann Arbor, Michigan 48109

Received 1 June 1987; revised 3 December 1987

ABSTRACT

A basic assumption of many epidemic models is that populations are composed of a homogeneous group of randomly mixing individuals. This is not a realistic assumption. Most actual populations are divided into a number of subpopulations, within which there may be relatively random mixing, but among which there is nonrandom mixing. As a consequence of the structuring of the population, there are several sources of heterogeneity within populations that can affect the course of an infection through the population. Two of these sources of heterogeneity are differences in contact number between subpopulations, and differences in the patterns of contact among subpopulations. A model for the spread of a disease in such a population is described. The model considers two levels of interaction: interactions between individuals within a subpopulation because of geographic proximity, and interactions between individuals of the same or different subpopulations because of attendance at common social functions. Because of this structure, it is possible to analyze with the model both heterogeneity in contact number and variation in the patterns of contact. A stability analysis of the model is presented which shows that there is a unique threshold for disease maintenance. Below the threshold the disease goes extinct, and the equilibrium is globally asymptotically stable. Above the threshold, the extinction equilibrium is unstable, and there is a unique endemic equilibrium.The analysis presents a sufficient condition for disease maintenance, which determines critical subpopulation sizes above which the disease cannot go extinct. The condition is a simple inequality relating the removal rate of infectives to the infection rate of susceptibles. In addition, bounds on the actual threshold and the effect of symmetry in the interaction matrix on the threshold are presented.

*Paper prepared for Los Alamos National Laboratory, Workshop on Nonlinearity in Medicine and Biology.

©Elsevier Science Publishing Co., Inc., 1988
52 Vanderbilt Ave., New York, NY 10017 0025-5564/88/$03.50

INTRODUCTION

A common assumption of many mathematical models for the spread of infectious diseases in a population is that the individuals within the population mix randomly. The population is generally divided into two or more groups, such as susceptible individuals, infective individuals, and recovered individuals, but these individuals constitute one large randomly mixing population, so that an infective individual has an equal probability of coming into contact with any one of the susceptible individuals in the population.

This assumption of random mixing among the individuals within the population is not realistic. Actual populations are divided into a number of smaller subpopulations, within which the mixing is more nearly random, but among which there is limited mixing. This structure in the population can be the result of many different factors, such as geographical separation of neighborhoods or villages, separation along lines of social interaction, division into host and vector species, or division into sexual groups.

The importance of this structuring of a population for the spread of a disease within a population can be illustrated with data on the incidence of hepatitis A in Albuquerque, New Mexico. Throughout 1979 there were over 700 cases of hepatitis A in Albuquerque, 28% of which were associated with day care centers (Bernalillo County Health Department, unpublished). The pattern of incidence of the disease among the city's day care centers markedly shows the influence of population structure on disease transmission within the day care population.

Figure 1 shows all centers with outbreaks of hepatitis, and indicates the size of the outbreak in each center. Five centers in Albuquerque were owned and operated by the same family and are linked together by a dashed line. Four of these centers had large outbreaks of the disease. One of the remaining two centers with large outbreaks occurred in the geographic center of the five family-owned centers, and the other one was located on the local Air Force base and had one student attending a center adjacent to a family-owned center in addition to the Air Force center. In both these latter two cases, there was likely a significant amount of contact within home neighborhoods of children attending these centers with children attending the family-owned centers.

These incidence data clearly indicate that the spread of hepatitis A among day care centers did not occur randomly. There was definite social localization among the centers owned and operated by the same family, probably as a result of shifting employees to fill in gaps that arose with illness as well as movement of children from one of the linked centers to another. The epidemiological data thus clearly indicate that structured models are a

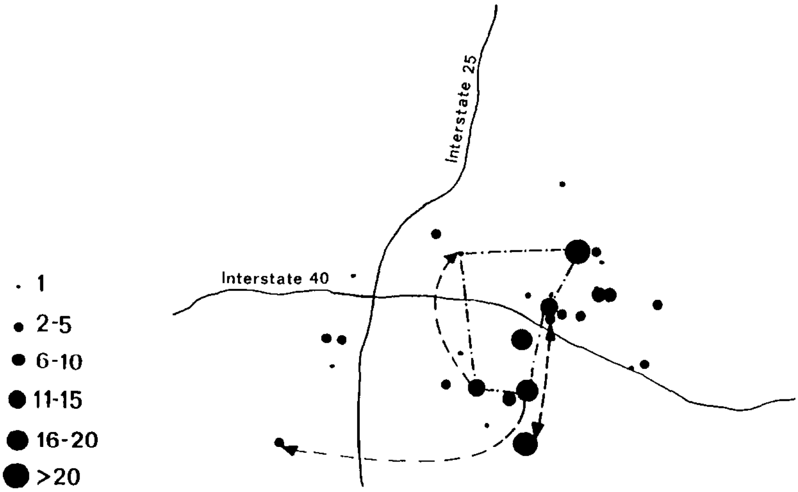

FIG. 1. Locations of day care centers in Albuquerque, New Mexico with at least one case of hepatitis A in 1979. Circles indicate size of outbreak in each center. Links between centers are indicated by dashed arrows showing the direction of transmission (–·–·–· = centers run by same family; — — — → = direct links between centers).

necessary starting point for the realistic modeling of actual disease patterns, at least on a local scale.

There are at least five important sources of heterogeneity in the infection process that derive from the population structure. First, there may be variation in relative susceptibility among individuals in different groups. Second, there may be variation in relative infectiousness of different individuals. Third, there may be nonrandom mixing among individuals because of the age structure of the population. Fourth, there may be heterogeneity due to a variation in the number of contacts made within and between groups. Finally, there may be heterogeneity due to the way in which the contacts are distributed among the different groups. This paper will be primarily concerned with describing a model which addresses the last two sources of heterogeneity, heterogeneity in contact number and heterogeneity in contact pattern.

There is a growing body of literature on models for the spread of disease in structured populations. These models can generally be divided into three types: those which consider a population divided into subgroups on the basis of age, those which consider a population divided into subgroups on the basis of sex, and those which consider a population divided into an arbitrary number of groups on the basis of social or geographic factors. The remainder of this paper will consider only models without age structure.

Most early models for the spread of disease in structured populations considered only two interacting groups [34, 11, 2, 3, 8]. These models are also appropriate for sexually transmitted diseases when there is only heterosexual transmission, so that the two groups are males and females [31, 32, 21]. For most diseases it is more realistic to consider models generalized to an arbitrary number of subpopulations. A model of this type was first proposed by Rushton and Mautner [25], who considered the spread of a disease for which there was no immunity and no recovery (commonly called the SI model, since the subpopulations consist only of susceptible and infective individuals). Watson [33], Lajmanovich and Yorke [17], Hethcote [12], Nold [23], Hethcote et al. [15], Hethcote and Van Ark [14], Post et al. [24], Rvachev and Longini [26], May and Anderson [19, 20], and Sattenspiel [27, 28] have extended and modified this model to take into account disease features such as removal of infectives, temporary immunity, latent periods, vital dynamics, immunization, and multiple types of social interaction among the individuals in the population. Recent reviews of models for structured populations can be found in [1], [5], and [7].

Most of these models for the spread of disease in n subpopulations consider heterogeneity in the contact rates within and between populations. The average number of effective contacts between individuals from different groups may vary from one group to another. For example, Hethcote [12] presents an SIRS epidemic model with immunizations and vital dynamics; Hethcote et al. [15] consider a gonorrhea model with the heterosexual population divided into four male and four female subpopulations on the basis of activity levels and presence or absence of symptoms; May and Anderson [19] and Hethcote and Van Ark [14] consider the effects of spatial heterogeneity in the design of immunization programs; and May and Anderson [20] present a model for the spread of AIDS in a heterogeneous homosexual population. These models and others of their type have provided significant inroads into the understanding of the effects of heterogeneity of contact number on the process of infection transmission.

Dietz and Schenzle [7] review much of the literature dealing with disease spread in structured populations. In addition, they evaluate critically the utility of these models for prediction of actual epidemiological patterns. They specifically address the importance of heterogeneity in contact rates among subpopulations for vaccination strategies and discuss reasons for the reluctance among epidemiologists to use the results from mathematical models in the practical control of diseases.

The fifth type of heterogeneity, variation in the patterns of contact among subpopulations, has been less well studied. Although May and Anderson [19, 20], Hethcote [12], and Hethcote and Van Ark [14] formulate their models in such a way that this heterogeneity could be studied, in general simplifying assumptions have been made about the patterns and there has been no analysis of the effects of varying these patterns. The most common of such assumptions is that the subpopulations undergo proportionate mixing, where the number of encounters between individuals of different subpopulations is proportional to the size of the subpopulations involved.

Hethcote et al. [15] and others explicitly address this heterogeneity due to mixing patterns. They recognize that the proportionate mixing assumption is unrealistic, and in addition to analyzing the proportionate mixing model, they consider a model combining a component due to proportionate mixing in the entire population with one due to proportionate mixing within activity levels only. However, the mixing matrices used are chosen to fit known patterns of incidence, rather than on theoretical grounds; and there is no exploration of the importance of variation in the mixing patterns to the transmission of infection through the population.

Rvachev and Longini [26] develop a model for the global spread of influenza that incorporates a transportation matrix giving the average number of individuals that travel from one population to another. Although the model is entirely general, there is only one particular pattern that is studied. The transportation matrix is estimated from data on the number of airline passengers traveling between the cities in a 24 hour period. In addition, the matrix is assumed to be symmetric. As in the case of Hethcote et al. [15], there is no exploration of the effects of varying the pattern of contact among subpopulations.

Heterogeneity in the patterns of contact among subpopulations may be an important factor affecting the spread of a disease throughout a population. For example, in the case of a sexually transmitted disease, the effect of having individuals in the population who are highly sexually active may depend significantly on whether these individuals interact only with other highly active individuals or with individuals from many different activity levels. Simply looking at the heterogeneity in number of contacts may not be sufficient to understand the relative importance of the variation in the number of contacts.

Sattenspiel [27, 28] and Travis and Lenhart [30] have begun to evaluate the importance of different patterns of contact on the transmission of infection in a subdivided population. The remainder of this paper will present Sattenspiel's model, will show the existence of a unique threshold for the maintenance of transmission of the disease, and will describe the effects of variations in the mixing patterns among subpopulations.

THE BASIC SUBDIVIDED MODEL

The model which Sattenspiel [27, 28] developed for the spread of an infection in a subdivided population incorporated a migration matrix to describe the patterns of movement of individuals between subpopulations. A migration matrix approach was first used by population geneticists to study the effects of population subdivision on the genetic structure of a population. This approach was apparently developed independently by Bodmer and Cavalli-Sforza [6] and C. A. B. Smith [29]. These models use a backward stochastic migration matrix, in which the elements m_{ij} give the probabilities that the parents of individuals in population i came from population j. The models are then used to derive the genetic variances and covariances among populations.

The infection transmission model developed by Sattenspiel [27, 28] uses a similarly defined "migration" or contact matrix to describe the probability of two individuals from different neighborhoods coming into contact. This matrix is a forward stochastic migration matrix, with each element m_{ij} giving the probability that an individual from population i moves to population j. In addition, the model is a hierarchical model which allows for the incorporation of two qualitatively different types of interaction among individuals: interactions within a local "neighborhood," where "neighborhoods" can be defined geographically, socially, or temporally, and interactions between neighborhoods, which are likely to involve only a portion of individuals living in a given neighborhood. Using this model, it is possible to assess the effects of varying the patterns of contact simply by exploring the effects of using different migration matrices in the model. Also, because of the hierarchical nature of the model, it is possible to consider what happens to the process of disease spread if some susceptible individuals are allowed to come into contact with infectives through multiple kinds of activities, while others have a limited possibility of contact with infectives.

Consider a population that is divided into n discrete neighborhoods. Within a neighborhood all individuals interact randomly, and in addition, a proportion of the individuals within a neighborhood engage in some kind of social behavior that allows them to come into contact with individuals from other neighborhoods. These two kinds of interaction can be modeled by considering each neighborhood to be further divided into two subneighborhoods: one consisting of all "nonsocial" individuals (who interact only with other individuals within the neighborhood), and one consisting of all "social" individuals (who interact with both nonsocial and social individuals from the same neighborhood and who also interact with social individuals from different neighborhoods). The patterns of interaction between subneighborhoods can then be described by a migration matrix, each element of which gives the probability of movement from one subneighborhood to another.

Each subneighborhood is composed of individuals who can be classified into three distinct groups with respect to their disease status: (a) susceptibles —those individuals who have not yet contracted a case of the disease and are therefore at risk for infection, (b) infectives—those individuals with an active infection who are capable of transmitting the infection to susceptible individuals, and (c) removed—those individuals with permanent immunity who can neither contract nor transmit the infection.

The following assumptions are made about the infection process:

(1) Births and deaths occur at a rate b_i in neighborhood i. This birth and death rate is equal for each of the two subneighborhoods within a neighborhood. All newborns are susceptible, but deaths occur among individuals in all classes. There is a constant total population size, so the total number of births in the population is equal to the total number of deaths.

(2) There is no permanent movement of individuals within or between subneighborhoods. However, there is temporary movement on a daily basis among the social subneighborhoods.

(3) The infection has no latency period.

(4) Recovery occurs at a constant rate γ and is proportional to the number of infectives. Recovery confers permanent immunity.

(5) The number of new cases is a proportion β of the total number of contacts between susceptible and infective individuals, corrected by a factor σ_i for differences in population size and density among the groups. Contact can occur within neighborhoods because of geographical proximity, and can occur within and between neighborhoods because of attendance at social functions.

In neighborhood i, we index members of the social subneighborhood by si and the members of the nonsocial subneighborhood by $0i$. Let N_{0i} and N_{si} represent the total population of the nonsocial subneighborhood and the social subneighborhood of neighborhood i, respectively. Let x_{0i}, y_{0i}, and z_{0i} be the numbers of susceptible, infective, and recovered individuals, respectively, in nonsocial subneighborhood $0i$. Let x_{si}, y_{si}, and z_{si} be the corresponding numbers for social subneighborhood si. The process of disease spread throughout the population can then be represented by the following system of $6n$ differential equations:

$$\frac{dx_{0i}}{dt} = b_i N_{0i} - b_i x_{0i} - \beta\sigma_i(x_{0i}y_{0i} + x_{0i}y_{si}) - \beta x_{0i}(\mathbf{MM}^T)_{0i}\mathbf{y}, \tag{1}$$

$$\frac{dx_{si}}{dt} = b_i N_{si} - b_i x_{si} - \beta\sigma_i(x_{si}y_{0i} + x_{si}y_{si}) - \beta x_{si}(\mathbf{MM}^T)_{si}\mathbf{y}, \tag{2}$$

$$\frac{dy_{0i}}{dt} = \beta\sigma_i(x_{0i}y_{0i} + x_{0i}y_{si}) + \beta x_{0i}(\mathbf{MM}^T)_{0i}\mathbf{y} - \gamma_i y_{0i} - b_i y_{0i}, \tag{3}$$

$$\frac{dy_{si}}{dt} = \beta\sigma_i(x_{si}y_{0i} + x_{si}y_{si}) + \beta x_{si}(\mathbf{MM}^T)_{si}\mathbf{y} - \gamma_i y_{si} - b_i y_{si}, \tag{4}$$

$$\frac{dz_{0i}}{dt} = \gamma_i y_{0i} - b_i z_{0i}, \tag{5}$$

$$\frac{dz_{si}}{dt} = \gamma_i y_{si} - b_i z_{si}, \tag{6}$$

where β is the transmission rate per unit contact, σ_i is a neighborhood-specific adjustment of the transmission rate, b_i is the birth (and death) rate, γ_i is the recovery rate in neighborhood i, and x_{0i}, x_{si}, y_{0i}, y_{si}, z_{0i}, and z_{si} are the numbers of individuals in each class in neighborhood i. y is a $2n \times 1$ vector with elements $(y_{01},\ldots,y_{0n},y_{s1},\ldots,y_{sn})$. The total number of births in each population is b_iN_i. All of these individuals are susceptible. Deaths in each class can be given by b_ix_{0i}, b_ix_{si}, b_iy_{0i}, b_iy_{si}, b_iz_{0i}, and b_iz_{si}.

The matrix $\mathbf{M}$ in these equations is a forward stochastic contact matrix, with each element, m_{ij}, representing the probability that a susceptible individual who lives in neighborhood i comes into contact with an infective individual who lives in nieghborhood j. Therefore, $0 \leqslant m_{ij} \leqslant 1$ for all i and j and $\Sigma_j m_{ij} = 1$. The rate of social contacts between susceptibles from subneighborhood si and infectives is given by the sum of the rates of contact between susceptibles from subneighborhood si and infectives from subneighborhood sj. This sum is given by $\Sigma_{j=1}^n[x_{si}m_{ij}(\Sigma_{k=1}^n y_{sk}m_{kj})]$, since $x_{si}m_{ij}$ is the rate that susceptibles from subneighborhood si visit subneighborhood sj, and $\Sigma_{k=1}^n y_{sk}m_{kj}$ is the total rate at which infectives visit sj. The matrix formulation of this sum is the desired social effect and is given by $-\beta x_{si}(\mathbf{M}_s\mathbf{M}_s^T)_i\mathbf{y}_s$, where $(\mathbf{M}_s\mathbf{M}_s^T)_i$ is the ith row of $\mathbf{M}_s\mathbf{M}_s^T$, and $\mathbf{y}_s$ is a vector giving the number of infectives in each social subneighborhood.

The entire movement matrix is a $2n \times 2n$ matrix of the following form:

$$\left[\begin{array}{c|c} \mathbf{0} & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{M}_s \end{array}\right],$$

where $\mathbf{M}_s$ is an $n \times n$ submatrix representing the patterns of movement among social subneighborhoods. All other elements are $n \times n$ submatrices with all elements equal to zero, since there is no movement between nonsocial and social or nonsocial and nonsocial subneighborhoods. Note that the matrix $\mathbf{M}$ represents only between-neighborhood transmission. Within-neighborhood transmission is represented by separate terms in Equations (1)–(4).

Using methods similar to those of Post et al. [24], a sufficient condition for the extinction of the disease in the population will be derived. The condition takes the following form:

$$N_{0i} < \frac{b_i + \gamma_i}{\beta(2\sigma_i)} \quad \text{and} \quad N_{si} < \frac{b_i + \gamma_i}{\beta\left[2\sigma_i + (\mathbf{M}\mathbf{M}^T)_{si}(\mathbf{1})\right]} \tag{7a}$$

for all $0i$ and si. $\mathbf{1}$ is a $2n \times 1$ vector with each element equal to one. If this condition holds, then the disease becomes extinct in the population.

There is a slightly different sufficient condition for the maintenance of the disease in a population. This condition takes the following form:

$$N_{0i} > \frac{b_i + \gamma_i}{\beta(\sigma_i)} \quad \text{or} \quad N_{si} > \frac{b_i + \gamma_i}{\beta\left[\sigma_i + (\mathbf{MM}^T)_{si}(\mathbf{1})\right]} \tag{7b}$$

for some $0i$ or si.

Let $\tau_{0i} = (b_i + \gamma_i)/\beta\sigma_i$ and let $\tau_{si} = (b_i + \gamma_i)/\beta[\sigma_i + (\mathbf{MM}^T)_{si}(\mathbf{1})]$. Then these conditions show that if $N_{0i} > \tau_{0i}$ or $N_{si} > \tau_{si}$ for some si or $0i$, then the disease remains endemic, while if $N_{0i} < \frac{1}{2}\tau_{0i}$ and $N_{si} < \frac{1}{2}\tau_{si}$ for all si and $0i$, the disease becomes extinct in the population. The true thresholds occur somewhere in between the numbers τ_{*i} and $\frac{1}{2}\tau_{*i}$.

The parameter τ has a simple biological interpretation. It is the ratio of the removal rate of infectives in the subneighborhood to the infection rate of susceptibles, and this can be called the relative removal rate of infectives. The sufficient condition for disease maintenance is then that the initial number of susceptibles in at least one neighborhood must be sufficiently large so that the infectives will not be removed before adequate contact has occurred between an infective and a susceptible for the disease to be transmitted to a susceptible. This condition is analogous to the classic threshold condition found for the general epidemic model of Kermack and McKendrick [35], which considers a homogeneous population with no vital dynamics, with the exception that when the population exceeds the threshold size the disease will be maintained rather than just temporarily increase in incidence. A proof of this result is presented below. A more detailed formulation of this model is given in [28].

MATHEMATICAL PROPERTIES OF THE MODEL

In this section we will verify the important mathematical properties of the system of $6n$ differential equations given by Equations (1)–(6). In particular, we will:

(1) Reduce this system to a system of $4n$ equations.

(2) Define a natural compact domain B of the new system and show that the domain is invariant under the system (1)–(6), i.e., that solutions which start in B stay in B.

(3) Show that $\{x_{0i} = N_{0i},\ x_{si} = N_{si},\ y_{0i} = y_{si} = z_{0i} = z_{si} = 0 \text{ for } i = 1,\dots,n\}$ is always an equilibrium. This equilibrium, which we call $\hat{\mathbf{N}}$, corresponds to the situation in which everyone is susceptible, with no infective or removed individuals.

(4) Show that there is a threshold level such that when parameters of the system determine group sizes that are below this threshold, $\hat{\mathbf{N}}$ is the only

equilibrium and it is *globally* asymptotically stable in B; *every* solution of Equations (1)–(6) tends to $\hat{\mathbf{N}}$.

(5) Show that when the parameters of the system determine group sizes that are above this threshold level, $\hat{\mathbf{N}}$ becomes an unstable equilibrium.

(6) Show that when $\hat{\mathbf{N}}$ becomes unstable, a unique new "endemic" equilibrium appears in the interior of B.

(7) Relate this threshold level to the criteria for maintenance and extinction of a disease in the population given by the inequalities (7a) and (7b) above.

1. *REDUCTION FROM 6n EQUATIONS TO 4n EQUATIONS*

By definition, $x_{0i}+y_{0i}+z_{0i}=N_{0i}$ and $x_{si}+y_{si}+z_{si}=N_{si}$ for all i, where the N's are constant. Furthermore, it follows from Equations (1)–(6) that

$$\frac{dx_{0i}}{dt}+\frac{dy_{0i}}{dt}+\frac{dz_{0i}}{dt}=b_i(N_{0i}-x_{0i}-y_{0i}-z_{0i}) \qquad (8a)$$

and

$$\frac{dx_{si}}{dt}+\frac{dy_{si}}{dt}+\frac{dz_{si}}{dt}=b_i(N_{si}-x_{si}-y_{si}-z_{si}). \qquad (8b)$$

Let $v_0(t)\equiv x_{0i}(t)+y_{0i}(t)+z_{0i}(t)-N_{0i}$. Define $v_s(t)$ similarly. Equation (8a) implies that $v_0(t)$ satisfies the initial value problem $(d/dt)v_0(t)=-b_iv_0(t)$ and $v_0(0)=0$. The only solution to this initial value problem is $v_0(t)\equiv 0$. Similar reasoning shows that $v_s(t)\equiv 0$. Therefore, $x_{0i}(t)+y_{0i}(t)+z_{0i}(t)=N_{0i}$ for all t, and $x_{si}(t)+y_{si}(t)+z_{si}(t)=N_{si}$ for all t. If we know x_{0i}, x_{si}, y_{0i}, and y_{si}, then z_{0i} and z_{si} will be uniquely determined. Consequently, we can drop the $2n$ equations given by (5) and (6) and work with the system of $4n$ equations given by Equations (1)–(4) alone.

2. *THE NATURAL DOMAIN OF THE SYSTEM (1)–(4) IS INVARIANT UNDER THE SYSTEM*

Clearly, for each i and for $*=0,s$ we want $x_{*i}\geqslant 0$, $y_{*i}\geqslant 0$, $z_{*i}\geqslant 0$, and $x_{*i}+y_{*i}+z_{*i}=N_{*i}$. These constraints can be summarized in xy space by letting the domain be

$$B=\{(x_{01},x_{s1},\dots,x_{0n},x_{sn},y_{01},y_{s1},\dots,y_{0n},y_{sn}):\\ 0\leqslant x_{0i},0\leqslant x_{si},0\leqslant y_{0i},0\leqslant y_{si},\ x_{0i}+y_{0i}\\ \leqslant N_{0i},x_{si}+y_{si}\leqslant N_{si}\text{ for }i=1,\dots,n\}$$

The compact convex set B is bounded by the $6n$ hyperplanes

$$0 = x_{0i}, \quad 0 = x_{si}, \quad 0 = y_{0i}, \quad 0 = y_{si}, \quad x_{0i} + y_{0i} = N_{0i}, \quad x_{si} + y_{si} = N_{si}$$

for $i = 1, \dots, n$. We want to show that any solution which starts on one of these $6n$ bounding hyperplanes moves into B. For example, from Equation (1), when $x_{0i} = 0$, $dx_{0i}/dt = b_i N_{0i} > 0$, i.e., $x_{0i}(t)$ is increasing and therefore moving into B. When $x_{si} = 0$, $dx_{si}/dt = b_i N_{si} > 0$, so $x_{si}(t)$ is moving into B. When $y_{0i} = 0$, $dy_{0i}/dt = \beta\sigma_i x_{0i} y_{si} \geqslant 0$, so $y_{0i}(t)$ does not move out of B. When $y_{si} = 0$, $dy_{si}/dt = \beta\sigma_i x_{si} y_{0i} + \beta x_{si} \Sigma_{j,k} m_{ij} m_{jk} y_{sk} \geqslant 0$, so $y_{si}(t)$ does not move out of B. Finally, when $x_{0i} + y_{0i} = N_{0i}$,

$$\begin{aligned}\frac{d}{dt}(x_{0i} + y_{0i}) &= b_i N_{0i} - b_i x_{0i} - \beta\sigma_i x_{0i}(y_{0i} + y_{si}) \\ &\quad + \beta\sigma_i x_{0i}(y_{0i} + y_{si}) - (\gamma_i + b_i) y_{0i} \\ &= -\gamma_i y_{0i} + b_i(N_{0i} - x_{0i} - y_{0i}) \\ &= -\gamma_i y_{0i} < 0.\end{aligned}$$

So solutions which start on the boundary $x_{0i} + y_{0i} = N_{0i}$ move into B. Similarly, one shows that when $x_{si} + y_{si} = N_{si}$, $(d/dt)(x_{si} + y_{si}) = -\gamma_i y_{si} < 0$. For more details of this proof technique, see Lemma 3.1 and Lemma 3.2 in [17].

3. $\hat{N} = (N_{01}, N_{s1}, \dots, N_{0n}, N_{sn}, 0, \dots, 0)$ *IS ALWAYS AN EQUILIBRIUM*

This observation follows immediately by inserting $x_{0i} = N_{0i}$, $x_{si} = N_{si}$, $y_{0i} = 0$, $y_{si} = 0$ into the right hand side of Equations (1)–(4) and noting that each expression becomes zero. At the point $\hat{\mathrm{N}}$, all the z_{0i}'s and z_{si}'s are also zero. There are no infective or removed individuals in the population; everyone is in the susceptible category.

4. *THERE IS A THRESHOLD LEVEL OF THE PARAMETERS BELOW WHICH* $\hat{\mathrm{N}}$ *IS A GLOBALLY ASYMPTOTICALLY STABLE EQUILIBRIUM FOR B*

First, we translate the equilibrium $\hat{\mathrm{N}}$ to the origin, using the following change of variables:

$$\begin{aligned} u_{0i} &= N_{0i} - x_{0i}, \qquad & y_{0i} &= y_{0i}, \\ u_{si} &= N_{si} - x_{si}, \qquad & y_{si} &= y_{si} \end{aligned}$$

for $i=1,\dots,n$. In terms of these new variables, the system (1)–(4) becomes

$$\frac{du_{0i}}{dt} = -b_i u_{0i} + \beta\sigma_i(N_{0i} - u_{0i})(y_{0i} + y_{si}), \tag{9}$$

$$\begin{aligned}\frac{du_{si}}{dt} &= -b_i u_{si} + \beta\sigma_i(N_{si} - u_{si})(y_{0i} + y_{si})\\ &\quad + \beta(N_{si} - u_{si})\sum_{j,k} m_{ij} m_{kj} y_{sk},\end{aligned} \tag{10}$$

$$\frac{dy_{0i}}{dt} = \beta\sigma_i(N_{0i} - u_{0i})(y_{0i} + y_{si}) - (\gamma_i + b_i)\, y_{0i}, \tag{11}$$

$$\begin{aligned}\frac{dy_{si}}{dt} &= \beta\sigma_i(N_{si} - u_{si})(y_{0i} + y_{si})\\ &\quad + \beta(N_{si} - u_{si})\sum_{j,k} m_{ij} m_{kj} y_{sk} - (\gamma_i + b_i)\, y_{si}.\end{aligned} \tag{12}$$

Let A be the $4n \times 4n$ matrix

$$\begin{aligned}\mathbf{A} &= \left[\begin{array}{cc|cc} -\mathbf{B} & 0 & \mathbf{S}_0 & \mathbf{S}_0 \\ 0 & -\mathbf{B} & \mathbf{S}_s & \mathbf{S}_s + \mathbf{M}_s \\ \hline 0 & 0 & \mathbf{S}_0 - \mathbf{G} - \mathbf{B} & \mathbf{S}_0 \\ 0 & 0 & \mathbf{S}_s & \mathbf{S}_s - \mathbf{G} - \mathbf{B} + \mathbf{M}_s \end{array}\right]\\ &= \left[\begin{array}{c|c} \mathbf{A}_1 & \mathbf{A}_2 \\ \hline 0 & \mathbf{A}_4 \end{array}\right]\end{aligned} \tag{13}$$

where $\mathbf{B}$ is an $n \times n$ diagonal matrix with diagonal elements b_i, $\mathbf{S}_0$ is an $n \times n$ diagonal matrix with diagonal elements $\beta\sigma_i N_{0i}$, $\mathbf{S}_s$ is an $n \times n$ diagonal matrix with diagonal elements $\beta\sigma_i \mathbf{N}_{si}$, $\mathbf{G}$ is an $n \times n$ diagonal matrix with diagonal elements γ_i, and $\mathbf{M}_s$ is an $n \times n$ matrix whose (i,k)th entry is $\sum_{j=1}^{n} m_{ij} m_{kj}$. All of the off diagonal entries in the matrix A come from the submatrices $\mathbf{S}_0$, $\mathbf{S}_\mathrm{s}$, and $\mathbf{M}_\mathrm{s}$, all of whose entries are nonnegative.

Let $\mathbf{N(u)}$ be the $4n \times 1$ column vector of quadratic functions

$$\mathbf{N(u)} = \begin{bmatrix} -\mathbf{F}_0 \\ -\mathbf{F}_s \\ -\mathbf{F}_0 \\ -\mathbf{F}_s \end{bmatrix},$$

where the ith entry of the $n \times 1$ column vector $\mathbf{F}_0$ is

$$\beta\sigma_i u_{0i}(y_{0i} + y_{si}) \geqslant 0$$

and the ith entry of the $n\times 1$ column vector $\mathbf{F}_s$ is

$$\beta\sigma_i u_{si}(y_{0i}+y_{si})+u_{si}\Sigma^n_{k,j=1}m_{ij}m_{kj}y_{sj}\geqslant 0.$$

Writing $\dot{u}$ and $\dot{y}$ for du/dt and dy/dt, we can abbreviate the system (9)–(12) as

$$\begin{bmatrix}\dot{\mathbf{u}}_0\\ \dot{\mathbf{u}}_s\\ \dot{\mathbf{y}}_0\\ \dot{\mathbf{y}}_s\end{bmatrix}=\mathbf{A}\begin{bmatrix}\mathbf{u}_0\\ \mathbf{u}_s\\ \mathbf{y}_0\\ \mathbf{y}_s\end{bmatrix}+\begin{bmatrix}-\mathbf{F}_0\\ -\mathbf{F}_s\\ -\mathbf{F}_0\\ -\mathbf{F}_s\end{bmatrix}$$

or

$$\begin{bmatrix}\dot{\mathbf{u}}\\ \dot{\mathbf{y}}\end{bmatrix}=\begin{bmatrix}\mathbf{A}_1 & \mathbf{A}_2\\ \mathbf{0} & \mathbf{A}_4\end{bmatrix}\begin{bmatrix}\mathbf{u}\\ \mathbf{y}\end{bmatrix}+\begin{bmatrix}-\mathbf{F}\\ -\mathbf{F}\end{bmatrix}. \tag{14}$$

Before analyzing this system further, we discuss some concepts of matrix algebra that will play a central role in our analysis. Let $\mathbf{P}$ be a matrix, like $\mathbf{A}$ in (13), in which all off diagonal entries are nonnegative. Such a matrix is sometimes called a *Metzler matrix*; see, for example, [18]. An $n\times n$ matrix $\mathbf{P}$ is *irreducible* if for any proper subset S of $\{1,\dots,k\}$ there exists an i in S and j in $\{1,\dots,k\}-S$ such that $p_{ij}\neq 0$. Two equivalent formulations of irreducibility are:

(a) some power of $\mathbf{P}$ has no zero entries, and
(b) for any i and j, there exist $k_1,\dots,k_m$ such that $p_{ik_1}p_{k_1k_2}\cdots p_{k_mj}\neq 0$.

For any $n\times n$ matrix $\mathbf{P}$, let $r_1,\dots,r_n$ be the eigenvalues of $\mathbf{P}$. Define the *stability modulus* of $\mathbf{P}$, $s(\mathbf{P})$, to be the maximum of the real parts of the r_i's for $i=1,\dots,n$. This notation reflects the fact that $\mathbf{0}$ is an asymptotically stable equilibrium of the linear system $\dot{\mathbf{x}}=\mathbf{P}\mathbf{x}$ if and only if $s(\mathbf{P})<0$. The following theorem indicates the connections between these concepts. For a proof, see [18], [17], or [4].

THEOREM 1

Suppose that $\mathbf{P}$ *is an* $n\times n$ *Metzler matrix* ($p_{ij}\geqslant 0$ *for* $i\neq j$). *Then*, $s(\mathbf{P})$ *is an eigenvalue of* $\mathbf{P}$, *and it has a corresponding eigenvector with only nonnegative components. If, furthermore,* $\mathbf{P}$ *is irreducible, then* $s(\mathbf{P})$ *is a simple eigenvalue (multiplicity one) of* $\mathbf{P}$, *and its eigenvector has only positive components. This eigenvector is the only eigenvector of* P *with all its components* $\geqslant 0$.

We now return to the system (14). Note that $\mathbf{A}$ is a Metzler matrix. However, because of the zeros in the lower left hand corner of $\mathbf{A}$, $\mathbf{A}$ cannot be irreducible. For this reason, we cannot directly apply Lajmanovich and Yorke's [17] theorem to analyze the equilibria of (14) and their stability. In fact, we will show later that some of the conclusions of this theorem do not hold for (14). We can, however, adapt the analyses in [17] and in [12] to derive the results we need for (14).

Note that the natural domain for the system (14) is

$$S=\{(u_{01},u_{s1},\dots,u_{0n},u_{sn},y_{01},\dots,y_{sn}):0\leqslant y_{0i}\leqslant u_{0i}\leqslant N_{0i}\\ \text{and } 0\leqslant y_{si}\leqslant u_{si}\leqslant N_{si} \text{ for } i=1,\dots,n\}.$$

The compact convex space S is just B in the new coordinates and is invariant under (14).

Assume that $\mathbf{M}_s$ is an irreducible matrix. In this case, $\mathbf{A}_4$ in (14) and $\mathbf{A}_4^T$ are irreducible $2n\times 2n$ Metzler matrices. Let $s=s(\mathbf{A}_4)$ be their common stability modulus. Let $\mathbf{w}$ be the $2n$-dimensional eigenvector of s for $\mathbf{A}_4^T$. All components of $\mathbf{w}$ are positive. The corresponding eigenvector of s for $\mathbf{A}^T$ is $(\mathbf{0}\ \ \mathbf{w})^T$.

The stability modulus $s(\mathbf{A}_4)$ turns out to be the desired threshold for our system (14) [or equivalently (1)–(6)]. In this section, we suppose that $s(\mathbf{A}_4)<0$.

Consider the linear function $V\colon S\to R$,

$$V(\mathbf{u},\mathbf{y})=\mathbf{w}\cdot\mathbf{y}.$$

Since $\mathbf{w}>\mathbf{0}$, $V(\mathbf{u},\mathbf{y})\geqslant 0$ on S. The derivative of V along orbits of (14) at $(\mathbf{u},\mathbf{y})$ is

$$\begin{aligned}\dot V(\mathbf{u},\mathbf{y})&=\nabla V(\mathbf{u},\mathbf{v})\cdot(\dot{\mathbf{u}},\dot{\mathbf{y}})\\ &=\begin{bmatrix}\mathbf{0}\\ \mathbf{w}\end{bmatrix}\cdot\left(\mathbf{A}\begin{bmatrix}\mathbf{u}\\ \mathbf{y}\end{bmatrix}-\mathbf{N}(\mathbf{u},\mathbf{y})\right)\\ &=\begin{bmatrix}\mathbf{0}\\ \mathbf{w}\end{bmatrix}\cdot\begin{bmatrix}\mathbf{A}_1\mathbf{u}+\mathbf{A}_2\mathbf{y}\\ \mathbf{A}_4\mathbf{y}\end{bmatrix}-\begin{bmatrix}\mathbf{0}\\ \mathbf{w}\end{bmatrix}\cdot\begin{bmatrix}\mathbf{F}(\mathbf{u},\mathbf{y})\\ \mathbf{F}(\mathbf{u},\mathbf{y})\end{bmatrix}\\ &=\mathbf{w}\cdot\mathbf{A}_4\mathbf{y}-\mathbf{w}\cdot\mathbf{F}(\mathbf{u},\mathbf{y})\\ &=\mathbf{A}_4^T\mathbf{w}\cdot\mathbf{y}-\mathbf{w}\cdot\mathbf{F}(\mathbf{u},\mathbf{y})\\ &=s\mathbf{w}\cdot\mathbf{y}-\mathbf{w}\cdot\mathbf{F}(\mathbf{u},\mathbf{y}).\end{aligned}$$

Since $\mathbf{w}>\mathbf{0}$, $\mathbf{y}\geqslant\mathbf{0}$, $\mathbf{F}(\mathbf{u},\mathbf{y})\geqslant\mathbf{0}$, we have $\dot V\leqslant 0$ on S. Furthermore, $\dot V=0$ if and only if $\mathbf{y}=\mathbf{0}$. When $\mathbf{y}=\mathbf{0}$, (14) reduces to $\dot u_i=-bu_i$, all of whose solutions tend to $\mathbf{0}$ as $t\to\infty$. Therefore, V is a Liapunov function for (14) on S, and $\mathbf{0}$ is a globally asymptotically stable equilibrium for (14), when $s<0$.

(See [16], [17], or [9] for a more complete discussion of the use of Liapunov functions.)

5. *WHEN THE GROUP SIZES DETERMINED BY THE PARAMETERS ARE ABOVE THE THRESHOLD LEVEL, THEN $\hat{\mathbf{N}}$ IS UNSTABLE. ALL ORBITS WHICH START INSIDE B WILL TEND AWAY FROM $\hat{\mathbf{N}}$*

Our threshold is the stability modulus $s(\mathbf{A}_4)$. Because $\mathbf{A}$ is a block matrix with $\mathbf{A}_3 = \mathbf{0}$ and $\mathbf{A}_1$ diagonal, the eigenvalues of $\mathbf{A}$ are the diagonal entries of $\mathbf{A}_1$ (the $-b_i$'s) and the eigenvalues of $\mathbf{A}_4$. By definition, when $s(\mathbf{A}_4) < 0$, all eigenvalues of $\mathbf{A}_4$ have negative real part and $\mathbf{0}$ is a locally asymptotically stable equilibrium. In fact, it was shown above that it is globally asymptotically stable.

When $s(\mathbf{A}_4) > 0$, $\mathbf{A}$ has an eigenvalue with positive real part and $\mathbf{0}$ is now an unstable (saddle) equilibrium. The eigenvector $(\mathbf{v},\mathbf{w})$ of $\mathbf{A}$ corresponding to eigenvalue $s = s(\mathbf{A}_4)$ satisfies $\mathbf{v} = (s\mathbf{I} - \mathbf{A}_1)^{-1}\mathbf{A}_2\mathbf{w}$. Since $s\mathbf{I} - \mathbf{A}_1$ is a diagonal matrix with positive entries on the diagonal, since all entries of $\mathbf{A}_2$ are nonnegative, and since $\mathbf{w}$ is a positive vector, all entries of $\mathbf{v}$ are nonnegative. If each row of $\mathbf{A}_2$ has a positive entry, $\mathbf{v}$ will be a strictly positive vector and $(\mathbf{v},\mathbf{w})$ will be a strictly positive eigenvector of $\mathbf{A}$. By the stable manifold theorem ([10] or [9]), there is an orbit leaving $\mathbf{0}$ which is tangent to this positive eigenvector $(\mathbf{v},\mathbf{w})$, and, in fact, most orbits move away from $\mathbf{0}$. We now prove this fact directly using the above Liapunov function.

Recall that $V(\mathbf{u},\mathbf{y}) = \mathbf{w}\cdot\mathbf{y}$ and $\dot{V}(\mathbf{u},\mathbf{y}) = s\mathbf{w}\cdot\mathbf{y} - \mathbf{w}\cdot\mathbf{F}(\mathbf{u},\mathbf{y})$. Since $(\mathbf{F}(\mathbf{u},\mathbf{y}))_i = u_i\Sigma_j c_{ij}y_j$,

$$\|\mathbf{F}(\mathbf{u},\mathbf{y})\| \leqslant C\|\mathbf{u}\|\,\|\mathbf{y}\| \quad \text{and} \quad |\mathbf{w}\cdot\mathbf{F}(\mathbf{u},\mathbf{y})| \leqslant C\|\mathbf{w}\|\,\|\mathbf{u}\|\,\|\mathbf{y}\|,$$

for some $C > 0$, where all the norms are on R^{2n}. Let $r_0 = \min\{\mathbf{w}\cdot\mathbf{z}|\mathbf{z} \geqslant \mathbf{0}$ and $\|\mathbf{z}\| = 1$ in $R^{2n}\}$, and let r_1 be the corresponding maximum. It follows that

$$r_0\|\mathbf{y}\| \leqslant \mathbf{w}\cdot\mathbf{y} \leqslant r_1\|\mathbf{y}\|$$

for all $\mathbf{y} \geqslant \mathbf{0}$ in R^{2n} and that

$$\dot{V}(\mathbf{u},\mathbf{y}) \geqslant sr_0\|\mathbf{y}\| - C\|\mathbf{w}\|\,\|\mathbf{u}\|\,\|\mathbf{y}\| = \|\mathbf{y}\|(s\mathbf{r}_0 - C\|\mathbf{w}\|\,\|\mathbf{u}\|).$$

Therefore, $\dot{V}(\mathbf{u},\mathbf{y}) > 0$ provided $y \neq 0$ and $\|\mathbf{u}\| \leqslant sr_0/C\|\mathbf{w}\|$. Consequently, any orbit $(\mathbf{u}(t),\mathbf{y}(t))$ which starts near zero with $\mathbf{y} \neq \mathbf{0}$ will continue to move away from $\mathbf{0}$, at least until $\|\mathbf{u}(t)\| = sr_0/C\|\mathbf{w}\|$. Notice, on the other hand, that on the boundary $\{\mathbf{y} = \mathbf{0}\}$ of S, the system (14) becomes $\dot{u}_i = -b_iu_i$, $\dot{y}_i = 0$, whose solutions are $(u_{i0}e^{-b_it},0)$—all of which tend to $\mathbf{0}$. This situation is in marked contrast to the SI epidemic model treated by the Lajmanovich-Yorke theorem. There, *every* solution which starts near $\mathbf{0}$ eventually moves $m > 0$ units away from $\mathbf{0}$.

6. *WHEN* $\hat{\mathbf{N}}$ *BECOMES UNSTABLE, A UNIQUE NEW "ENDEMIC" EQUILIBRIUM APPEARS IN THE INTERIOR OF* $\mathbf{B}$

We want to show the existence and uniqueness of an endemic equilibrium when $s(\mathbf{A}_4) > 0$. The uniqueness part will follow from arguments in [17] and [12]. However, neither of these two papers properly handles the existence part for the system (14). In Lajmanovich and Yorke's SI model, $V \geqslant \epsilon > 0$ is a compact, convex invariant set, which must therefore contain a nonzero equilibrium by standard Brouwer fixed point type arguments. For the system (14), $V \geqslant \epsilon$ is invariant only for small enough $\|\mathbf{u}\|$, not globally. A nontrivial adaptation of the Lajmanovich-Yorke technique is required for (14)—a point that has not been previously addressed for this problem.

Define

$$r_2 \equiv sr_0/C\|\mathbf{w}\|,$$
$$S_0 = \{(\mathbf{u},\mathbf{y}) \in S : \mathbf{y} = \mathbf{0} \text{ and } u_i = N_i \text{ for some } i\},$$
$$S_1 = \{(\mathbf{u},\mathbf{y}) \in S : \mathbf{y} = \mathbf{0} \text{ and } \|\mathbf{u}\| = r_2\}.$$

Let $\phi(t; u_0, y_0)$ denote the solution curve of (14) with initial condition $\phi(0; u_0, y_0) = (u_0, y_0)$. Since (14) becomes $\dot{u}_i = -b_i u_i$ on $\{\mathbf{y} = \mathbf{0}\}$,

$$\phi(t; u_0, 0) = \left(u_{01_0}e^{-b_1 t}, \dots, u_{sn_0}e^{-b_n t}, 0, \dots, 0\right),$$

which goes to $\mathbf{0}$ as $t \to \infty$. By the "flow box theorem" (a straightforward application of the implicit function theorem; see [16, pp. 242–244] or [9, pp. 43–46]), for each $(u_0, 0) \in S_0$, there is a unique $t_1 = t_1(u_0, 0)$ such that $\phi(t_1; u_0, 0) \in S_1$ and t_1 depends smoothly on $(u_0, 0)$. This process induces a diffeomorphism (smooth invertible map with a smooth inverse), $\Phi : S_0 \to S_1$, defined by $\Phi(u_0, 0) = \phi(t(u_0, 0); u_0, 0)$. (See Figure 2.)

By the same flow box theorem, Φ extends to a diffeomorphism $\hat{\Phi} : N_0 \to N_1$, where N_0 is a neighborhood of S_0 in $\{(\mathbf{u},\mathbf{y}) : u_i = N_i \text{ for some } i\}$, N_1 is a neighborhood of S_1 in $\{(\mathbf{u},\mathbf{y}) : \|\mathbf{u}\| = r_2\}$, and $\hat{\Phi}$ has the form

$$\hat{\Phi}(u_0, y_0) = \phi(t_1(u_0, y_0); u_0, y_0).$$

Geometrically, if $\mathbf{z}$ is on the $u_i = N_i$ boundary of S, one follows the orbit of (14) from $\mathbf{z}$ until it hits $\|\mathbf{u}\| = r_2$ at the point $\hat{\Phi}(\mathbf{z})$.

Choose $\epsilon > 0$ small enough so that $T_1 = \{(\mathbf{u},\mathbf{y}) \in S : V(\mathbf{u},\mathbf{y}) = \mathbf{w} \cdot \mathbf{y} = \epsilon$ and $\|\mathbf{u}\| = r_2\}$ lies in the neighborhood N_1 of S_1, i.e., in the image of $\hat{\Phi} : N_0 \to N_1$.

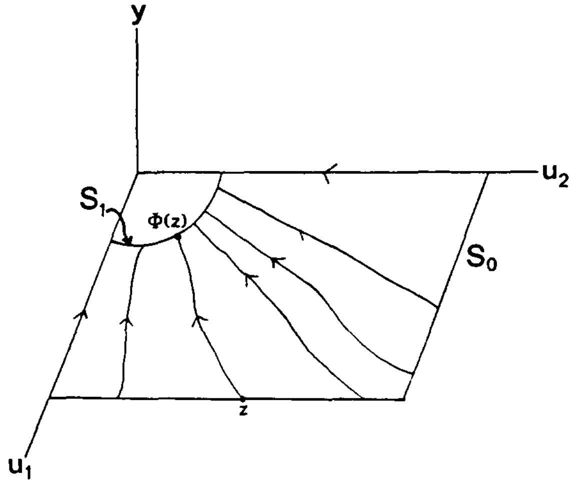

FIG. 2. The diffeomorphism $\Phi : S_0 \to S_1$.

Let $T_2 = \hat{\Phi}^{-1}(T_1)$ in N_0. Finally, let

$$S_3 = \{(\mathbf{u},\mathbf{y}) \in S : (\mathbf{u},\mathbf{y}) = \phi(t; u_0, y_0) \text{ for some } (u_0, y_0) \in T_2 \text{ and some } 0 \leqslant t \leqslant t(u_0, y_0)\}.$$

Then, S_3 is a "hypersurface" composed of orbits of (14) which connects the boundary $\{u_i = N_i \text{ for some } i\}$ to $\{(\mathbf{u},\mathbf{y}) : \mathbf{w}\cdot\mathbf{y} = \epsilon \text{ and } \|\mathbf{u}\| = r_2\}$. (See Figure 3.)

Now, let S_4 be the $4n$-dimensional space bounded by the manifolds $u_i = y_i$, $u_i = N_i$, $\{\mathbf{w}\cdot\mathbf{y} = \epsilon,\ \|\mathbf{u}\| \leqslant r_2\}$, and S_3. Figure 4 gives a schematic drawing of S_4 for a two-dimensional S. The set S_4 is invariant; orbits which start in S_4 stay in S_4. It is also compact, but is probably not convex. However, it is homeomorphic to the convex set S. By the usual Brouwer fixed point like arguments (see, for example, [17] or [16]), (14) has an equilibrium within S_4. Since $\mathbf{0}$ is not in S_4, this is our new "endemic" equilibrium.

To prove uniqueness, we basically follow the arguments in [17] and [12]. The equilibria are found by setting the right hand sides in the system (1)–(4) equal to zero. From $dx_{0i}/dt + dy_{0i}/dt = 0$ in (1) and (3), we find that at equilibrium

$$x_{0i} = N_{0i} - \frac{b_i + \gamma_i}{b_i} y_{0i}.$$

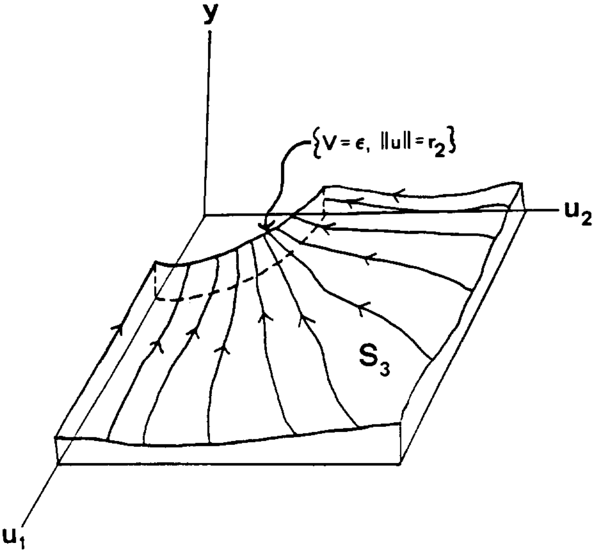

FIG. 3. The manifold of orbits S_3 joining $\{V=\epsilon,\ \|\mathbf{u}\|=r_2\}$ to the outer boundary of S.

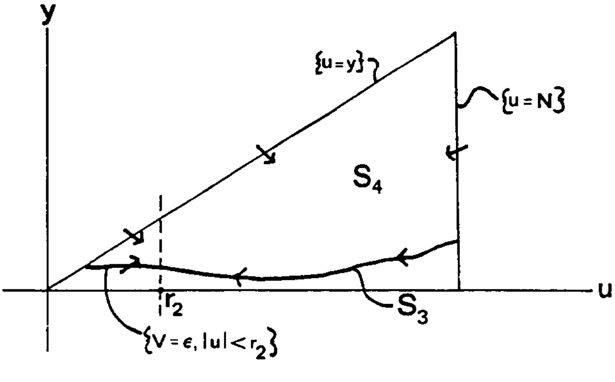

FIG. 4. The invariant set $S_4 \subset S$.

From $dx_{si}/dt + dy_{si}/dt = 0$, we find that at equilibrium

$$x_{si} = N_{si} - \frac{b_i + \gamma_i}{b_i} y_{si}.$$

Plug these relationships into Equations (1) and (2) to obtain the $2n$ quadratic equations in $2n$ unknowns:

$$-(\gamma_i + b_i) y_{0i} + (N_{0i} - \xi_i y_{0i})[\beta\sigma_i(y_{0i} + y_{si})] = 0,$$

$$-(\gamma_i + b_i) y_{si} + (N_{si} - \xi_i y_{si})\left[\beta\sigma_i(y_{0i} + y_{si}) + \beta \sum_{k,j} m_{ij} m_{kj} y_{sk}\right] = 0,$$

for $i = 1, \ldots, n$, where $\xi_i = (\gamma_i + b_i)/b_i$.

In order to simplify notation, rewrite these $2n$ equations as

$$-a_i y_i + (n_i - y_i) \sum_j \beta_{ij} y_j = 0, \tag{15}$$

where $(y_1, \ldots, y_{2n}) = (y_{01}, \ldots, y_{0n}, y_{s1}, \ldots, y_{sn})$, $a_i = (b_i + \gamma_i)/\xi_i$, $n_i = N_{0i}/\xi_i$, etc. Assume, following Lajmanovich and Yorke [17], that $\mathbf{y} = \mathbf{k}$ and $\mathbf{y} = \mathbf{h}$ are two constant nonzero solutions of (15). In fact, by what we know about the behavior of our system on the boundary of B, we can assume that every component of $\mathbf{h}$ and $\mathbf{k}$ is positive. If $\mathbf{h} \neq \mathbf{k}$, we can assume without loss of generality that $h_1 > k_1$ and $h_1/k_1 \geqslant h_i/k_i$ for all i. Then

$$-a_1 h_1 + (n_1 - h_1) \sum_j \beta_{1j} h_j = -a_1 k_1 + (n_1 - k_1) \sum_j \beta_{1j} k_j = 0.$$

Multiplying the term on the left by k_1/h_1 yields

$$-a_1 k_1 + (n_1 - h_1) \sum_j \beta_{1j} h_j k_1/h_1 = -a_1 k_1 + (n_1 - k_1) \sum_j \beta_{1j} k_j = 0,$$

or

$$(n_1 - h_1) \sum_j \beta_{1j} h_j k_1/h_1 = (n_1 - k_1) \sum_j \beta_{1j} k_j.$$

But this is a contradiction to

$$h_j k_1/h_1 \leqslant k_j \quad \text{and} \quad n_1 - h_1 < n_1 - k_1.$$

Therefore, there is only one constant solution of (1)–(6), other than $\mathbf{0}$.

It is extremely difficult to analyze the stability of this endemic equilibrium, as Hethcote [12], Post et al. [24], and Hethcote and Thieme [13] note in similar models. This is one of the major unsolved problems in the dynamic analysis of the spread of epidemics through a population. Hethcote and Thieme [13] do provide an argument to show that such an equilibrium is locally asymptotically stable.

7. *UPPER AND LOWER BOUNDS ON THE THRESHOLDS*

By the above discussion, the threshold level which separates the situation in which the disease dies out ($\hat{\mathbf{N}}$ a global attractor) from the situation in which the disease becomes endemic ($\hat{\mathbf{N}}$ locally unstable) is determined by the sign of the eigenvalue $s(\mathbf{A}_4)$. Our assumption that $\mathbf{A}_4$ is an irreducible Metzler matrix does allow us to derive some necessary and some sufficient conditions for $s(\mathbf{A}_4)$ to be negative. As Post et al. [24] point out, a matrix $\mathbf{B}$ for which $b_{ij} \leqslant 0$ for $i \neq j$ and all eigenvalues have positive real part is called an *M*-matrix. Therefore, $s(\mathbf{A}_4) < 0$ if and only if $-\mathbf{A}_4$ is an M-matrix. There are a number of equivalent conditions for a matrix with $b_{ij} \leqslant 0$ for $i \neq j$ to be an M-matrix:

(a) all principal minors of $\mathbf{B}$ are positive,
(b) all leading principal minors of $\mathbf{B}$ are positive,
(c) all eigenvalues of $\mathbf{B}$ have positive real part,
(d) there is a vector $\mathbf{u} > \mathbf{0}$ such that $\mathbf{Mu} > \mathbf{0}$,
(e) there is a vector $\mathbf{v} > \mathbf{0}$ such that $\mathbf{M}^T\mathbf{v} > \mathbf{0}$.

See [4] or [22] for complete proofs and discussion.

Write

$$-\mathbf{A}_4 = \begin{bmatrix} -\mathbf{S}_0 + \mathbf{G} + \mathbf{B} & -\mathbf{S}_0 \\ -\mathbf{S}_s & -\mathbf{S}_s + \mathbf{G} + \mathbf{B} - \mathbf{M}_s \end{bmatrix}.$$

Applying condition (d) to $\mathbf{u} = (1,1,\ldots,1)$ and condition (e) to $\mathbf{v} = (1,\ldots,1)$, we find that $s(\mathbf{A}_4) < 0$ and the disease dies out if either of the following conditions holds:

$$(\mathrm{A}_1) \qquad \frac{\gamma_i + b_i}{2\beta\sigma_i} > N_{0i} \qquad \text{for} \quad i = 1,\ldots,n$$

and

$$(\mathrm{A}_2) \qquad \frac{\gamma_i + b_i}{2\beta\sigma_i + \beta\Sigma_{j,k} m_{ij} m_{kj}} > N_{si} \qquad \text{for} \quad i = 1,\ldots,n,$$

or

$$(\mathrm{B}) \qquad \frac{\gamma_i + b_i}{\beta\sigma_i} > N_{0i} + N_{si} + \frac{1}{\sigma_i}\sum_{j,h} m_{hj} m_{ij} N_{sh} \qquad \text{for} \quad i = 1,\ldots,n.$$

On the other hand, if we apply condition (a) to the diagonal entries of $-\mathbf{A}_4$, we find that *any* of the following conditions will guarantee that $s(\mathbf{A}_4)>0$ and the disease becomes endemic:

$$\text{(C)}\qquad N_{0i} > \frac{\gamma_i + b_i}{\beta\sigma_i} \qquad \text{for any} \quad i=1,\dots,n,$$

$$\text{(D)}\qquad N_{si} > \frac{\gamma_i + b_i}{\beta\sigma_i + \beta\Sigma_j m_{ij}^2} \qquad \text{for any} \quad i=1,\dots,n,$$

$$\text{(E)}\qquad N_{si} > \frac{\gamma_i + b_i}{\beta\sigma_i} \qquad \text{for any} \quad i=1,\dots,n.$$

These estimates show that the actual threshold for N_{0i} lies between $(\gamma_i+b_i)/2\beta\sigma_i$ and $(\gamma_i+b_i)/\beta\sigma_i$, and that the actual threshold for N_{si} lies between $(\gamma_i+b_i)/(2\beta\sigma_i+\beta\Sigma_{j,k}m_{ji}m_{kj})$ and $(\gamma_i+b_i)/(\beta\sigma_i+\beta\Sigma_j m_{ij}^2)$.

For $n=2$, $\mathbf{A}_4$ is a 4×4 matrix. One can compute its four leading principal minors and then use condition (b) above to derive an exact criterion for $s(\mathbf{A}_4)$ to be negative.

THEOREM 2

Consider the system (1)–(6) *for* $n=2$. *Suppose that* $\mathbf{M}$ *is a* 2×2 *irreducible matrix. Let* $g_i \equiv (\gamma_i+b_i)/\beta$ *for* $i=1, 2$, *and* $\Sigma_{ij}\equiv\Sigma_h m_{hi}m_{hj}$, *the* (ij)*th entry of* $\mathbf{MM}^T$. *Then all solutions of* (1)–(6) *tend to* $\mathbf{0}$ *as* $t\to\infty$ *if and only if each of the following four numbers is positive*:

(1) $P_1 = g_1 - \sigma_1 N_{01}$,
(2) $P_2 = g_2 - \sigma_2 N_{02}$,
(3) $P_3 = (g_1-\sigma_1 N_{01})(g_1 - N_{s1}\Sigma_{11}) - g_1\sigma_1 N_{s1}$ *or*
$P_3^* = (g_2-\sigma_2 N_{02})(g_2 - N_{s2}\Sigma_{22}) - g_2\sigma_2 N_{s2}$,
(4) $P_4 = P_3 P_3^* - P_1 P_2 (N_{s1}N_{s2}\Sigma_{12}^2)$.

These four numbers are positive if and only if all of the eigenvalues of $\mathbf{A}_4$ have negative real part. Since P_4 is the determinant of $\mathbf{A}_4$, it is the product of the eigenvalues of $\mathbf{A}_4$. Therefore, it is the threshold which matters most, in that P_4 must change sign as the system moves from the extinction equilibrium to the endemic equilibrium.

PATTERNS OF CONTACT AND DISEASE ENDEMICITY

By using the migration matrix approach it is possible to evaluate the effects of heterogeneity in contact patterns on the conditions for disease maintenance in a population. To accomplish this task, we need to relate the threshold $s(\mathbf{A}_4)$ to properties of the migration matrices $\mathbf{M}$ and $\mathbf{MM}^T$. This relationship is fairly complex. One can achieve some insights by seeing how different patterns of movement affect the threshold bounds we have just derived. An alternative approach, which may be more useful, is to make some simplifying assumptions on the parameters of the problem and then to derive a relationship between $s(\mathbf{A}_4)$ and $s(\mathbf{MM}^T)$ for this special case.

Assume that $n = N_{01} = N_{s1} = N_{0j} = N_{sj}$ for all $j > 1$ and that $b_1 = b_j$, $\gamma_1 = \gamma_j$, $\sigma_1 = \sigma_j$ for all j. Then we can write

$$A_4 = \left[\begin{array}{c|c} -g\mathbf{I} & 0 \\ \hline 0 & -g\mathbf{I} \end{array}\right] + \left[\begin{array}{c|c} h\mathbf{I} & h\mathbf{I} \\ \hline h\mathbf{I} & h\mathbf{I} + n\mathbf{MM}^T \end{array}\right].$$

Then

$$\mathbf{A}_4 \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = \mathbf{r} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}$$

if and only if

$$\begin{bmatrix} h\mathbf{I} & h\mathbf{I} \\ h\mathbf{I} & h\mathbf{I} + n\mathbf{MM}^T \end{bmatrix} \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix} = (r+g) \begin{bmatrix} \mathbf{w}_1 \\ \mathbf{w}_2 \end{bmatrix}$$

if and only if

$$h\mathbf{w}_1 + h\mathbf{w}_2 = (r+g)\mathbf{w}_1 \quad \text{and} \quad h\mathbf{w}_1 + h\mathbf{w}_2 + n\mathbf{MM}^T\mathbf{w}_2 = (r+g)\mathbf{w}_2$$

if and only if

$$\mathbf{w}_1 = \frac{h}{r+g-h}\mathbf{w}_2 \quad \text{and} \quad \mathbf{MM}^T\mathbf{w}_2 = \frac{(r+g)(r+g-2h)}{n(r+g-h)}\mathbf{w}_2.$$

Here, $(r+g)(r+g-2h)/n(r+g-h)$ is $s(\mathbf{MM}^T)$, the Perron-Frobenius eigenvalue of $\mathbf{MM}^T$, if and only if $r = s(\mathbf{A}_4)$. Since $(r+g)(r+g-2h)/n(r+g-h)$ is an increasing function of r, the larger $s(\mathbf{MM}^T)$ is, the larger $S(\mathbf{A}_4)$ is. The question now becomes: what patterns in the entries of $\mathbf{M}$ yield larger values for $S(\mathbf{MM}^T)$?

In Figure 5 we present four idealized movement patterns. In pattern 1 all social groups mix randomly with all other social groups. In pattern 2 all social groups are isolated from each other. In pattern 3 there are two types of social groups: a small local cluster that sends individuals only to other groups within the cluster but receives individuals from all groups, and a randomly mixing cluster in which each group sends individuals to all other groups but receives individuals only from other groups within the randomly mixing cluster. In pattern 4, every individual in the population goes to a single social facility.

Note that if $\mathbf{M}$ is symmetric, so that $\mathbf{M}^T = \mathbf{M}$, then $\mathbf{MM}^T = \mathbf{M}^2$ is still a Markov matrix and $s(\mathbf{MM}^T) = 1$. The matrices for models 1 and 2 are

$$\begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \vdots & & \vdots \\ \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix}$$

pattern 1

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

pattern 2

$$\begin{array}{c} (r \times r) \quad (r \times n-r) \\ \left(\begin{array}{cccc|cccc} \frac{1}{r} & \cdots & \cdots & \frac{1}{r} & 0 & \cdots & \cdots & 0 \\ \frac{1}{r} & \cdots & \cdots & \frac{1}{r} & 0 & \cdots & \cdots & 0 \\ \hline \frac{1}{n} & \cdots & \cdots & \frac{1}{n} & \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \\ \frac{1}{n} & \cdots & \cdots & \frac{1}{n} & \frac{1}{n} & \cdots & \cdots & \frac{1}{n} \end{array}\right) \\ (n-r \times r)\,(n-r \times n-r) \end{array}$$

pattern 3

$$\begin{pmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \\ \vdots & & & \vdots \\ 1 & 0 & \cdots & 0 \end{pmatrix}$$

pattern 4

FIG. 5. The lower right hand block, $\mathbf{M}_s$, of the movement matrices $\mathbf{M}$ for the patterns discussed. The total number of social subneighborhoods in the population in n, and r is the number of social subneighborhoods in the local cluster.

symmetric. Hence, $s(\mathbf{MM}^T)=1$ for both these patterns. However, for model 3, when $r=n/2$, $s(\mathbf{MM}^T)=\frac{1}{4}(3+\sqrt{5})\simeq 1.3$, as one can check directly. The largest possible s, $s(\mathbf{MM}^T)=n$, occurs for the "least symmetric" case, that of pattern 4.

Let us consider more closely the $n=2$ case treated in Theorem 2. As noted there, as the system moves from the extinction equilibrium to the endemic equilibrium, $P_4=\det \mathbf{A}_4$ must change sign. One can rewrite P_4 as

$$\begin{aligned} P_4 = {} & (g_1-\sigma_1 N_{01})(g_2-\sigma_2 N_{02})N_{s1}N_{s2}(\Sigma_{11}\Sigma_{22}-\Sigma_{12}\Sigma_{21}) \\ & + g_1 g_2 (g_1-\sigma_1 N_{01})(g_2-\sigma_2 N_{02}) \\ & + g_1 g_2 \sigma_1 \sigma_2 N_{s1} N_{s2} - (g_1-\sigma_1 N_{01}) N_{s1}[g_1(g-\sigma_2 N_{02}) - g_2\sigma_2 N_{s2}]\Sigma_{11} \\ & -(g_2-\sigma_2 N_{02}) N_{s2}[g_2(g_1-\sigma_1 N_{01}) - g_1 \sigma_1 N_{s1}]\Sigma_{22} \\ & - g_1 g_2 \sigma_1 N_{s1}(g_2-\sigma_2 N_{02}) - g_1 g_2 \sigma_2 N_{s2}(g_1-\sigma_1 N_{01}). \end{aligned}$$

Since this is an unwieldy expression to work with, let us once again make the simplifying assumptions: $g_1=g_2=g$, $\sigma_1=\sigma_2=\sigma$, $N_{01}=N_{02}=N_{s1}=N_{s2}=N$. Now, P_4 simplifies to

$$\begin{aligned} P_4 = {} & (g-\sigma N)^2 N^2 \det(\mathbf{MM}^T) - (g-\sigma N)(g-2\sigma N) gN \operatorname{tr}(\mathbf{MM}^T) \\ & + g_2 (g-2\sigma N)^2. \end{aligned} \tag{16}$$

The first and last terms of (16) are positive. If the middle term is also positive, then P_4 is always positive and no bifurcation occurs. Therefore, we will assume that the middle term of (16) is negative, i.e., that $g-\sigma N > g-2\sigma N>0$. Recall that $g-\sigma N>0$ is a necessary condition for the disease to die out, while $g-2\sigma N>0$ is part of the sufficient condition.

One checks easily for a 2×2 Markov matrix $\mathbf{M}$ that

$$\operatorname{tr}(\mathbf{MM}^T)=\det(\mathbf{MM}^T)+1+(m_{12}-m_{21})^2.$$

It follows that the more asymmetric $\mathbf{M}$ is, the larger $(m_{12}-m_{21})^2$ is, the larger the value of $\operatorname{tr}(\mathbf{MM}^T)$ relative to the value of $\det(\mathbf{MM}^T)$ is, the smaller P_4 becomes in (16), and the more likely it is that $P_4 < 0$, and that the disease becomes endemic.

These observations show that the asymmetry of the mixing patterns is critical. Any pattern of mixing that is symmetric will have little effect on the condition for endemicity for different groups. On the other hand, groups that form a small local cluster have a larger effective population size because of the asymmetry of the mixing patterns, so that actual neighborhood sizes can be smaller and still support the disease. It is not clear whether this is a general phenomenon related to the symmetry itself, or whether it derives from the assumption of a stochastic movement matrix together with the symmetry of the matrices. This is an important point that remains to be explored, because the models of Rvachev and Longini [26] and Hethcote et al. [15] that explicitly incorporate a mixing matrix rely upon either symmetry or proportionate mixing. Because of these assumptions, it is likely that there will be no differential effects found among populations because of the patterns of contact between populations.

Actual patterns of movement among populations, however, are rarely, if ever, symmetric, so that the effects of this source of heterogeneity need to be examined more carefully. The data on hepatitis A in Albuquerque, New Mexico presented earlier illustrate the importance of this heterogeneity. Those centers that had close social or geographic links (analogous to the local cluster of pattern 3) did appear to be at higher risk for the infection than other centers within the city, although it is possible that other factors could account for the increased incidence in these centers.

CONCLUSIONS

The development of models to describe the effects of the various forms of heterogeneity in infection transmission is of great importance in understanding and predicting the patterns of spread of most infections. The use of a matrix approach promises to be of great value for such models, because this approach allows for the easy incorporation of multiple factors which work together to increase the risk of contact among individuals—factors such as geographic location, social networks, and specific behaviors such as drug use or sexual preference. Such factors are of critical importance in the maintenance of chains of transmission of infections. The matrix approach also makes it possible to explore the effects of different patterns of contact, not just variation in the frequency of contact among groups. This knowledge will increase the ability of models to capture the essential information from the real world and to maintain a sufficient degree of realism, so that meaningful predictions about ways to control infections are possible.

The research on which this manuscript is based was funded in part by grant no. BNS83-10491 from the National Science Foundation and by a grant from the Wenner-Gren Foundation for Anthropological Research.

The following individuals provided helpful comments during the preparation of this paper: Ted Bergstrom, Herbert Hethcote, James Koopman, Ira Longini, Jr., and Alan Rogers.

REFERENCES

1 N. T. J. Bailey, Spatial models in the epidemiology of infectious diseases, *Lecture Notes in Biomath.* 38:233–261 (1980).

2 N. G. Becker, The spread of an epidemic to fixed groups within the populations, *Biometrics* 24:1007–1014 (1968).

3 N. G. Becker, A stochastic model for two interacting populations, *J. Appl Probab.* 7:544–564 (1970).

4 A. Berman and R. Plemmons, *Nonnegative Matrices in the Mathematical Sciences*, Academic, New York, (1979).

5 F. L. Black and B. Singer, Elaboration versus simplification in refining mathematical models of infectious disease, *Ann Rev. Microbiol*, 41:677–701 (1987).

6 W. F. Bodmer and L. L. Cavalli-Sforza, A migration matrix model for the study of random genetic drift, *Genetics* 59:565–592 (1968).

7 K. Dietz and D. Schenzle, Mathematical models for infectious disease statistics, in *A Celebration of Statistics* (A. C. Atkinson and S. E. Feinberg, Eds.), Springer, New York, 1985, pp. 167–204.

8 V. Capasso, A stochastic model for epidemics in two interacting regions of a large population, *Boll. Un. Mat. Ital.* 5(13B):216–235 (1976).

9 J. Hale, *Ordinary Differential Equations*, Krieger, Huntington, N.Y., 1980.

10 P. Hartman, *Ordinary Differential Equations*, Wiley, New York, 1964.

11 H. W. Haskey, Stochastic cross-infection between two otherwise isolated groups, *Biometrika* 44:193–204 (1957).

12 H. W. Hethcote, An immunization model for a heterogeneous population, *Theoret. Population Biol.* 14:338–349 (1978).

13 H. W. Hethcote and H. R. Thieme, Stability of the endemic equilibrium in epidemic models with subpopulations, *Math. Biosci.* 75:205–277 (1985).

14 H. W. Hethcote and J. W. Van Ark, Epidemiological models for heterogeneous populations: Proportionate mixing, parameter estimation, and immunization programs, *Math. Biosci.*, 84:85–117 (1987).

15 H. W. Hethcote and J. A. Yorke, and A. Nold, Gonorrhea modeling: A comparison of control methods, *Math. Biosci.* 58:93–109 (1982).

16 M. Hirsch and S. Smale, *Differential Equations, Dynamical Systems, and Linear Algebra*, Academic, New York, 1974.

17 A. Lajmanovich and J. A. Yorke, A deterministic model for gonorrhea in a nonhomogeneous population, *Math. Biosci.* 28:221–236 (1976).

18 D. G. Luenberger, *Introduction to Dynamic Systems*, Wiley, New York, 1979.

19 R. M. May and R. M. Anderson, Spatial heterogeneity and the design of immunization programs, *Math Biosci.* 72:83–111 (1984).

20 R. M. May and R. M. Anderson, Transmission dynamics of HIV infection, *Nature* 326:137–142 (1987).

21 R. Nallaswamy and J. B. Shukla, Effects of dispersal on the stability of a gonorrhea endemic model, *Math. Biosci.* 61:63–72 (1982).

22 N. Nikaido, *Introduction to Sets and Mappings in Modern Economics*, American Elsevier, New York, 1970.

23 A. Nold, Heterogeneity in disease-transmission modeling, *Math. Biosci.* 52:227–240 (1980).

24 W. M. Post, D. L. DeAngelis, and C. C. Travis, Endemic disease in environments with spatially heterogeneous host populations, *Math. Biosci.* 63:289–302 (1983).

25 S. Rushton and A. J. Mautner, The deterministic model of a simple epidemic for more than one community, *Biometrika* 42:126–132 (1955).

26 L. A. Rvachev and I. M. Longini, A mathematical model for the global spread of influenza, *Math. Biosci.* 75:3–22 (1985).

27 L. Sattenspiel, "The Spread of Disease in Subdivided Populations" Ph.D. Dissertation, Univ. of New Mexico, 1984.

28 L. Sattenspiel, Population structure and the spread of disease, *Human Biol.* 59:411–438 (1987).

29 C. A. B. Smith, Local fluctuations in gene frequenceies, *Ann. Human Genetics* 32:251–260 (1969).

30 C. C. Travis and S. M. Lenhart, Eradication of infectious diseases in heterogeneous populations, *Math Biosci.* 83:191–198 (1987).

31 P. Waltman, A deterministic model of the spread of an infection between two populations, in *Delay and Functional Differential Equations and their Applications* (K. Schmitt, Ed.), Academic, New York, 1972, pp. 281–291.

32 P. Waltman, A threshold criterion for the spread of an infection in a two population model, *Math. Biosci.* 21:119–125 (1974).

33 R. K. Watson, On an epidemic in a stratified population, *J. Appl Probab.* 9:659–666 (1972).

34 E. B. Wilson and J. Worcester, The spread of an epidemic, *Proc. Nat. Acad. Sci. U.S.A.* 31:327–333 (1945).

35 W. O. Kermack and A. G. McKendrick, Contributions to the mathematical theory of epidemics, part I. *Proc. R Soc., A* 115:700–721 (1927).