On the estimation problem of mixing/pair formation matrices with applications to models for sexually-transmitted diseases

by

Carlos Castillo-Chavez,
Shwu-Fang Shyu,
Gail Rubin
and
Dave Umbach

BU-1144-MA

January 1992

ON THE ESTIMATION PROBLEM OF MIXING/PAIR FORMATION MATRICES WITH APPLICATIONS TO MODELS FOR SEXUALLY-TRANSMITTED DISEASES

Carlos Castillo-Chavez, Shwu-Fang Shyu, Gail Rubin and David Umbach

Abstract

A problem of considerable importance lying at the interface of social dynamics, demography, and epidemiology is determining and modeling who is mixing with whom. In this article we describe a general approach, using nonlinear mixing matrices, for modeling the process of pair-formation in heterogeneous populations. Determining who is mixing with whom is complicated by a variety of factors, including the problem of denominators, which is, in our context, equivalent to the nonexistence of closely interacting social/sexual networks. We describe the use of a mark-recapture model for estimating the sizes of the missing link, that is, the size of the population having sexual contact with a specified population and hence at risk for sexually-transmitted diseases. The need to estimate the size of the sexually-active subset before estimating the size of the population at risk introduces extra variability into the problem. An estimator of the variance of the estimated size of the population at risk that accounts for this extra variability and an expression for the bias of such an estimator have been derived. We illustrate our results with data collected from a population of university undergraduates, and make use of our axiomatic modeling approach for mixing/pair formation to compute specific mixing matrices. Complete details of this work will be published elsewhere.

1. Introduction

The importance of social and sexual interactions in the spread of sexually transmitted diseases, especially AIDS, has been well documented by sociologists, modelers, public health officials, etc. However, the development of methods for quantifying social dynamical processes in ways that allow different rates of mixing between subgroups in sexually-active populations has been quite difficult (but see Anderson *et al.*, 1986, 1989; Blythe *et al.*, 1991, 1992; Busenberg and Castillo-Chavez, 1991; Castillo-Chavez, 1989; Castillo-Chavez, ed., 1989; Castillo-Chavez and Busenberg, 1991; Castillo-Chavez *et al.*, 1991; Dietz, 1988; Dietz and Hadeler, 1988; Gupta *et al.*, 1989; Hethcote and Yorke, 1984; Hethcote and Van Ark, 1987, 1991; Hethcote *et al.*, 1991; Hyman and Stanley,

1988, 1989; Jacquez *et al.*, 1988; Rubin *et al.*, 1991; Sattenspiel and Castillo-Chavez, 1990; and references therein). Despite these efforts, work in this direction is still in its infancy in part because there is almost no adequate and/or sufficient data. Furthermore, the lack of substantial medical progress in dealing with HIV infections at the individual level has already had a strong impact on the dynamics of HIV at the population level. Dramatic changes of behavior have been observed in homosexually-active communities in San Francisco (Centers for Disease Control, 1985; McKusick et al., 1985a, 1985b; Shilts, 1987; Winkelstein, *et al.*, 1988), New York (Martin, 1986a; McFarland, 1972), and Boston (Saltzman *et al.*, 1987), demanding the development of dynamic models for the transmission of STDs that incorporate behavioral change. Unfortunately, epidemiological models that consider behavioral changes may not exhibit "typical" dynamics. In fact, multiple endemic equilibria may be quite common for models with state-dependent mixing/pair-formation processes, and control policies may have unpredictable results in these circumstances. For some recent efforts in this direction, the interested reader is referred to the work of Blythe *et al.* (1991), Castillo-Chavez (1989), Castillo-Chavez, ed. (1989), Huang *et al.* (1992) and Palmer *et al.* (1991).

This article is organized as follows: in Section 2, we describe a general axiomatic approach, using nonlinear mixing matrices for modeling the process of pair-formation in heterogeneous populations and explain their role in the transmission dynamics of STDs. Determining who is mixing with whom is complicated by a variety of factors including the problem of denominators which is, in our context, equivalent to the nonexistence of closely interacting social/sexual networks. In Section 3, we describe an axiomatic framework for modeling human interactions such as dating, mixing, or pair-formation (other interpretations are possible). In Section 4, we describe the use of a mark-recapture model for estimating the sizes of the missing link, that is, the size of the population having sexual contact with a specified population and hence at risk for sexually-transmitted diseases. One must estimate the size of the sexually-active subset before estimating the size of the population at risk, which introduces extra variability into the problem. An estimator of the variance of the estimated size of the population at risk that accounts for this extra variability and an expression for the bias of such an estimator have been derived (see Rubin *et al.*, 1991). In Section 5, we outline the possible use of these results with data collected from a population of university undergraduates, and by combining the results of Section 4 with our axiomatic modeling approach for mixing/pair formation we are able to compute specific mixing matrices (closed networks). In Section 6, we provide some conclusions and outline possible new directions.

2. Modeling of human epidemics

We begin with the description of a key component for a model of this type: the incidence rate (new cases of infection per unit time). To keep the level of discussion simple, we assume that we are dealing with a specific disease: gonorrhea. Therefore we have to consider, for a simple model, only two type of epidemiological classes: susceptibles and infecteds (here assumed infectious). The mixing probabilities, as well as other behavioral and epidemiological parameters, determine the rate at which new infections are generated. The incidence rate is given by a nonlinear function of the different interacting subpopulations and, in this context, we develop our approaches to estimating the social/sexual mixing structure of a population.

The data used in applying this modeling framework consists of a population of heterosexually-active college students and, consequently, we formulate our ideas in the context of two-sex heterosexually mixing populations (the description is considerably simplified when one deals with exclusively homosexually-active populations). Our heterosexually-active population is divided into classes or subpopulations which may be defined by sex, race, socio-economic background, average degree of sexual activity, etc. Models that incorporate factors such as chronological age, age of infection, variable infectivity, and partnership duration also have been formulated (see Busenberg and Castillo-Chavez, 1989, 1991).

For the purposes of this article, we consider only N-sexually active populations of females and L-sexually active populations of males, each divided into two epidemiological classes: $S_f^j(t)$ and $S_m^i(t)$ (susceptible females and males, i.e., uninfected and sexually active at time t); $I_f^j(t)$ and $I_m^i(t)$ (infected females and males at time t); for $j = 1,\cdots,N$ and $i = 1,\cdots,L$. Consequently, individuals at risk (sexually-active) of each sex and each subpopulation at time t are represented by $T_f^j(t) = S_f^j(t) + I_f^j(t)$ and $T_m^i(t) = S_m^i(t) + I_m^i(t)$. We obviously do not need to consider other individuals if our only concern is, as in this paper, the study of the dynamics of sexually-transmitted diseases.

Following the superscript notation, $B_f^j(t)$ and $B_m^i(t)$ denote the jth and ith incidence rates for females in group j and males in group i at time t, that is, the number of new infective cases in each subpopulation per unit time. As we shall see, $B_f^j(t)$ and $B_m^i(t)$ constitute a set of complicated expressions. In fact, they are functions of the frequency and type of sexual interactions that susceptible females of group j and susceptible males of group i have with all other sexually-active individuals (in this case, of the opposite sex, although this condition can be easily relaxed).

A dynamic model needs a source of new individuals; the modeling of this demographic process could be extremely complicated (see

Busenberg and Castillo-Chavez, 1989, 1991; Castillo-Chavez and Busenberg, 1991; Castillo-Chavez *et al.*, 1991). If one wishes, for example, to study the demographic consequences of a disease like HIV/AIDS, one needs to model carefully the "recruitment" of new individuals. Here we want to keep the demography simple and assume constant "recruitment" and constant per-capita mortality (removal from sexual activity) rates. Specifically, we let Λ_f^j and Λ_m^i denote the "recruitment" rates (assumed constant), μ_f^j and μ_m^i denote the (constant) per-capita removal rates from sexual activity, and γ_f^j and γ_m^i denote the (constant) per-capita recovery rates from gonorrhea infection. A simple model for the transmission dynamics of gonorrhea is given by the following set of differential equations:

$$\frac{dS_f^j(t)}{dt} = \Lambda_f^j - B_f^j(t) - \mu_f^j S_f^j(t) + \gamma_f^j I_f^j(t), \tag{1}$$

$$\frac{dI_f^j(t)}{dt} = B_f^j(t) - (\gamma_f^j + \mu_f^j) I_f^j(t), \tag{2}$$

$$\frac{dS_m^i(t)}{dt} = \Lambda_m^i - B_m^i(t) - \mu_m^i S_m^i(t) + \gamma_m^i I_m^i(t), \tag{3}$$

$$\frac{dI_m^i(t)}{dt} = B_m^i(t) - (\gamma_m^i + \mu_m^i) I_m^i(t), \tag{4}$$

$$i = 1,\cdots, L \text{ and } j = 1,\cdots,N.$$

Of course, this model is not fully specified until we provide explicit expressions for $B_f^j(t)$ and $B_m^i(t)$. The formulae are provided in two steps: first we provide expressions for the incidences in terms of a set of mixing probabilities $\{p_{ij}(t)$ and $q_{ji}(t)$: $i=1,\cdots,L$ and $j=1,\cdots,N\}$; and secondly, we describe these mixing probabilities (in the next section) in terms of an axiomatic system for social/sexual interactions. More definitions are needed:

$p_{ij}(t)$: fraction of partnerships of males in group i with females in group j at time t,

$q_{ji}(t)$: fraction of partnerships of females in group j with males in group i at time t,

$T_m^i(t)$: male population size of group i at time t,

$T_f^j(t)$: female population size of group j at time t,

c^i : average (constant) number of female partners per unit time of males in group i, or the ith-group rate of male pair-formation,

b^j : average (constant) number of male partners per unit time of females in group j, or the jth-group rate of female pair-formation,

β_m^i : transmission coefficient (constant) of males in group i,

β_f^j : transmission coefficient (constant) of females in group j.

The following expressions for the incidence rates are a direct consequence of these definitions:

$$B_m^i(t) = c^i S_m^i(t) \sum_{j=1}^{N} \beta_f^j p_{ij}(t) \frac{I_f^j(t)}{T_f^j(t)}, \tag{5}$$

and

$$B_f^j(t) = b^j S_f^j(t) \sum_{i=1}^{L} \beta_m^i q_{ji}(t) \frac{I_m^i(t)}{T_m^i(t)}. \tag{6}$$

The modeling of the mixing/pair-formation probabilities constitute the body of the next section.

3. Modeling of mixing/pair-formation probabilities

Solutions for one-sex mixing populations have been previously obtained by Anderson et al. (1989), Blythe and Castillo-Chavez (1989), Castillo-Chavez and Blythe (1989), Gupta *et al.* (1989), Hethcote and Yorke (1984), Hyman and Stanley (1988, 1989), Jacquez *et al.* (1988, 1989), Nold (1980), and many others. A representation theorem describing all mixing/pair-formation solutions as random perturbations of random (proportionate) mixing, based on the work of Blythe and Castillo-Chavez (op. cits.), was obtained by Busenberg and Castillo-Chavez (1989, 1991). Models that follow pairs of individuals (two-sex models) can be found (in a demographic context) in the works of Kendall (1949), Keyfitz (1949), Parlett (1972), and Pollard (1973). Formulations of the standard two-sex mixing pair-formation framework are found in the work of Fredrickson (1971) and Martin (1986b), while applications of the Fredrickson-McFarland framework to epidemiological models has been carried out by Castillo-Chavez (1989), Castillo-Chavez *et al.* (1991), Dietz (1988), Dietz and Hadeler (1988), Hadeler (1989a, 1989b), Hadeler and Ngoma (1990) and Waldstätter (1989). In this section we provide an alternative approach to modeling the process of pair-formation or social mixing. Like Fredrickson (1971), we use an axiomatic framework to describe the probabilities associated with possible interactions such as pair-formation, or social mixing (further details are found in Castillo-Chavez *et al.*, 1991, where some special solutions were given). Specifically, the set of mixing probabilities $\{p_{ij}(t)$ and $q_{ji}(t)$: $i = 1,\cdots,L$ and $j = 1,\cdots,N\}$ establishes the mixing/pair formation in a heterosexually-active population in agreement with the following (postulated) set of properties:

Def $(\mathcal{P}(t), \mathcal{Q}(t)) \equiv (p_{ij}(t), q_{ji}(t))$ is called a mixing/pair-formation matrix if and only if it satisfies the following properties at all times:

(A1) $0 \le p_{ij} \le 1, \qquad 0 \le q_{ji} \le 1,$

(A2) $\sum_{j=1}^{N} p_{ij} = 1 = \sum_{i=1}^{L} q_{ji}$,

(A3) $c^i T_m^i p_{ij} = b^j T_f^j q_{ji}, \qquad i = 1, \cdots, L, \quad j = 1, \cdots, N.$

(A4) If for some i, $1 \le i \le L$ and/or some j, $1 \le j \le N$, $c^i b^j T_m^i T_f^j = 0$, then we define $p_{ij} \equiv q_{ji} \equiv 0$.

Property (A3) can be interpreted as a conservation of partnerships law or a group reversibility property (applied to rates), while (A4) asserts, the obvious, that is, that the mixing of "non-existing" or non-sexually active subpopulations cannot be arbitrarily defined. A very useful solution is the Ross solution which corresponds to proportionate mixing when there are two clearly distinct sets of individuals who do not mix among themselves. Ross solutions naturally arise if we search for separable solutions.

Def A two-sex mixing/pair-formation function is called separable iff it is given by products of the form

$$p_{ij} = p_i \tilde{p}_j \quad \text{and} \quad q_{ji} = q_j \tilde{q}_i ,$$

where p_i, $\tilde{p}_j$, q_j, $\tilde{q}_i$ are arbitrary functions subject to the mixing constraints, $i = 1, \cdots, L$ and $j = 1, \cdots, N$.

Theorem 1: The only separable solution is Ross solution given by $(\bar{p}^j, \bar{q}^i)$, which are featured by superscripts and bars, and

$$\bar{p}^j = \frac{b^j T_f^j}{\sum_{i=1}^{L} c^i T_m^i}, \quad \bar{q}^i = \frac{c^i T_m^i}{\sum_{j=1}^{N} b^j T_f^j}; \quad j = 1, \cdots, N \text{ and } i = 1, \cdots, L . \quad (7)$$

Remark: From (A3) it follows that

$$\frac{p_{ij}}{q_{ji}} = \frac{b^j T_f^j}{c^i T_m^i} = \frac{\bar{p}^j}{\bar{q}^i}, \quad (8)$$

and hence (A4) implies that the support of any two-sex mixing function is contained in the support of $(\bar{p}^j, \bar{q}^i)$.

We now use (7) to generate all solutions to axioms (A1)-(A4). We begin by introducing some new terms. Let

$(\phi_{ij}^m) \equiv$ males' structural covariance matrix $(0 \le \phi_{ij}^m)$ denoting the

degree of preference (i.e., the deviation from random mixing) that group i-males have from group j-females, $j = 1,\cdots,N$, $i = 1,\cdots,L$.

$$\ell_m^i \equiv \sum_{k=1}^{N} \bar{p}^k \phi_{ik}^m \equiv \text{weighted average preference of group } i \text{ males}, \quad i = 1,\cdots,L.$$

$$R_m^i \equiv 1 - \ell_m^i, \qquad i = 1, \cdots, L. \tag{9}$$

We require that $R_m^i \geq 0$, and that

$$\sum_{i=1}^{L} \ell_m^i \bar{p}^i = \sum_{i=1}^{L} \sum_{k=1}^{N} \bar{p}^k \phi_{ik}^m \bar{p}^i < 1 . \tag{10}$$

Similarly, let

$(\phi_{ji}^f) \equiv$ females' structure covariance matrix ($0 \leq \phi_{ji}^f$) denoting the degree of preference (i.e., the deviation from random mixing) that group j-females have for group i-males, $j = 1,\cdots,N$, $i = 1,\cdots,L$.

$$\ell_f^j \equiv \sum_{k=1}^{L} \bar{q}^k \phi_{jk}^f \equiv \text{weighted average preference of group } j\text{-females}, \quad i = 1,\cdots,N .$$

$$R_f^j \equiv 1 - \ell_f^j, \qquad j = 1, \cdots, N . \tag{11}$$

Again, we require that $R_f^j \geq 0$, and that

$$\sum_{j=1}^{N} \ell_f^j \bar{q}^j = \sum_{j=1}^{N} \sum_{k=1}^{L} \bar{q}^k \phi_{jk}^f \bar{q}^j < 1 . \tag{12}$$

All solutions to axioms (A1) – (A4) are given (formally) by the following multiplicative perturbations to the separable mixing solution $(\bar{p}^j, \bar{q}^i)$:

$$p_{ij} = \bar{p}^j \left[\frac{R_f^j R_m^i}{\sum_{k=1}^{N} \bar{p}^k R_f^k} + \phi_{ij}^m \right], \qquad i = 1,\cdots,L; \quad j = 1,\cdots,N, \tag{13}$$

$$q_{ji} = \bar{q}^i \left[\frac{R_m^i R_f^j}{\sum_{k=1}^{L} \bar{q}^k R_m^k} + \phi_{ji}^f \right]. \tag{14}$$

The formal proof of this result can be found in Castillo-Chavez and Busenberg (1991). For future reference, we write down this theorem explicitly:

Theorem 2. Let $\{\phi^m_{ij}\}$ and $\{\phi^f_{ji}\}$ be two nonnegative matrices. Let $\ell^i_m \equiv \sum_{k=1}^{N} \bar{p}^k \phi^m_{ik}$ and $\ell^j_f \equiv \sum_{k=1}^{L} \bar{q}^k \phi^f_{jk}$, where $\left\{(\bar{p}^j, \bar{q}^i)\colon j = 1,\cdots,N \text{ and } i = 1,\cdots,L\right\}$ denotes the set composed of Ross solutions. We also let $R^i_m \equiv 1 - \ell^i_m$, $i = 1, \cdots, L$ and $R^j_f \equiv 1 - \ell^j_f$, $j = 1, \cdots, N$, and assume that ϕ^m_{ij} and ϕ^f_{ji} are chosen in such a way that R^i_m and R^j_f remain nonnegative for all time. We further assume that

$$\sum_{i=1}^{L} \ell^i_m \bar{p}^i = \sum_{i=1}^{L} \sum_{k=1}^{N} \bar{p}^k \phi^m_{ik} \bar{p}^i < 1,$$

and

$$\sum_{j=1}^{N} \ell^j_f \bar{q}^j = \sum_{j=1}^{N} \sum_{k=1}^{L} \bar{q}^k \phi^f_{jk} \bar{q}^j < 1 .$$

Then all the solutions to axioms (A1)-(A4) are given by Equations (13) and (14).

Remark: ϕ^m_{ij} and ϕ^f_{ji} can always be chosen in such a way that R^i_m and R^j_f remain nonnegative for all time (i.e., let them be in the interval [0,1]). However, there is no recipe for specifying necessary conditions for the nonnegativity of R^i_m and R^j_f because their values are intimately connected to the time-dependent values of Ross solutions and hence to the associated (and disease-dependent) dynamical system.

4. Estimation of sizes of mixing subpopulations

Our main purpose here is to compute *explicit examples* of mixing/pair-formation matrices from our data on mixing. These time-dependent matrices describe the network of interactions between groups of individuals (who is mixing with whom). Our examples (to be illustrated in Section 5), albeit for a single time, provide the first mixing matrices computed from data. Knowledge of these matrices over a period of time is essential to *any* type of long-term forecasting. Because our purposes are limited and our data is too specific, we do not need to use sophisticated approaches in the construction of these matrices. Mark-recapture methodology can be applied to survey data to estimate the number of different sexual partners from each of several groups that an individual has had in a fixed period of time, or to estimate the size of the population having sexual contact with members of a given group. Thus, one can apply this methodology to survey data to estimate the size of the population at risk for a sexually transmitted disease. Using data from our survey conducted at Cornell University in 1989 (see Crawford *et al.*, 1990), we use mark-recapture estimators to provide estimates of the size

of the population that has sexual contact with Cornell undergraduates but are not Cornell undergraduates. We use these estimates and our one- and two-sex mixing framework to construct explicit mixing matrices (see Section 5). In our situation, prior to sampling, the population contains both marked and unmarked individuals: contacts (i.e., sexual partners) are either Cornell undergraduates or not and, obviously, we only have access to information about Cornell and non-Cornell partners from the Cornell students surveyed. It is appropriate to think of the students surveyed as observers in mark-recapture bird studies in which "recapture" is done by sighting. Because the nature of our population, we need not worry about loss of marks or marks being overlooked, which is a problem in many applications of mark-recapture to bird and mammalian populations. For each student surveyed, the contacts reported are distinct sexual partners. However, any two Cornell students that were surveyed may share one or more sexual partners, either from the Cornell undergraduate pool, from the greater Ithaca area, or from the world. Thus, the combined number may contain multiple counts of the same sexual partner; we are sampling *with replacement* with respect to sexual contacts when we combine information across the students surveyed. Hence the closed population single mark release model, which is based on sampling with replacement (Bailey 1951), gives an appropriate *first* estimate of the population size. We believe that given the current availability of data on mixing, this approximation is entirely appropriate for our purposes. Let the subscript k denote sex (k = male, female). S_k denotes the number of undergraduates of sex k registered at Cornell, and T_k of those (S_k) are sexually active. The total contacts with individuals (Cornell undergraduates, or not Cornell undergraduates) of sex k per unit time (two months) from respondents of the opposite sex is denoted by y_k, and x_k of those are contacts with Cornell undergraduates. Then the Lincoln-Petersen estimator $\hat{N}_k$ is given by

$$\hat{N}_k = T_k(y_k+1) \Big/ (x_k+1) , \tag{15}$$

which is a nearly unbiased estimator of the number of individuals of sex k at risk (N_k) (see Bailey, 1951), and

$$\hat{\mathrm{Var}}(\hat{N}_k \mid y_k, T_k) = T_k^2(y_k+1)(y_k-x_k) \Big/ \left\{(x_k+1)^2(x_k+2)\right\} \tag{16}$$

is a nearly unbiased estimator of the variance of $\hat{N}_k$, when the number of sexually active students is known. Because a given sexual partner can be reported by more than one of the students surveyed, the total number of contacts (y_k) can be greater than N_k, thereby increasing the precision of the survey for N_k (see Seber, 1982).

However, we must estimate T_k from the survey data; T_k can be estimated with the maximum likelihood estimator under Bailey's approximate binomial model as

$$\hat{T}_k = S_k t_k \Big/ r_k \equiv S_k \hat{\pi}_k ,$$

where $\hat{\pi}_k$ estimates $\pi_k = T_k \,/\, S_k$, the probability of an individual of sex k in the surveyed population being sexually active, and r_k and t_k denote the number of Cornell undergraduates in the sample and the corresponding number that are sexually active. Hence, the corresponding estimator of N_k is

$$\tilde{N}_k = \hat{T}_k(y_k+1) \Big/ (x_k+1) , \tag{17}$$

which is a nearly unbiased estimator also, with proportional bias of order

$$\left[1-\pi_k+\pi_k \exp\left\{-y_k S_k \Big/ (N_k r_k)\right\}\right]^{r_k} \equiv \left\{B(N_k)\right\}^{r_k} . \tag{18}$$

An estimator of the variance of $\tilde{N}_k$ that takes into account the additional variability due to estimation of T_k is given by

$$\hat{\mathrm{Var}}(\tilde{N}_k \,|\, y_k) = A(\hat{\pi}_k) \Big/ C(\hat{\pi}_k) + \tilde{N}_k^2 \left[\left\{B(\tilde{N}_k/2)\right\}^{r_k} - \left\{B(\tilde{N}_k)\right\}^{2r_k}\right], \tag{19}$$

where

$$\begin{aligned} A(\hat{\pi}_k) = {} & S_k^3 y_k (y_k+1)^2 (\tilde{N}_k r_k^3)^{-1} r_k \hat{\pi}_k [\{1-a_k\} + \hat{\pi}_k (r_k-1)\{3-7a_k\} \\ & + \hat{\pi}_k^2 (r_k-1)(r_k-2)\{1-6a_k\} - \hat{\pi}_k^3 (r_k-1)(r_k-2)(r_k-3)a_k], \end{aligned} \tag{20}$$

with $a_k = S_k/(\tilde{N}_k r_k)$, and

$$\begin{aligned} C(\hat{\pi}_k) = {} & \hat{\pi}_k^4 r_k (r_k-1)(r_k-2)(r_k-3) y_k^4 a_k^4 + 2\hat{\pi}_k^3 r_k (r_k-1)(r_k-2) \\ & \times y_k^3 a_k^3 \left\{3y_k a_k + 2\right\} + \hat{\pi}_k^2 r_k (r_k-1) y_k^2 a_k^2 \left\{6(y_k a_k+1)^2 + 1\right\} \\ & + \hat{\pi}_k r_k y_k a_k (y_k a_k+2)\left\{2 + y_k a_k (y_k a_k+2)\right\} + 1 . \end{aligned} \tag{21}$$

The bias of this variance estimator is given in Rubin *et al.* (1991).

We wish to estimate the size of the population that has sexual contact with Cornell undergraduates but are not Cornell undergraduates, that is, $O_k = N_k - T_k$. An estimate of O_k is given by

$$\begin{aligned} \hat{O}_k = \tilde{N}_k - \hat{T}_k &= \left\{\hat{T}_k(y_k+1) \Big/ (x_k+1)\right\} - \hat{T}_k \\ &= \hat{T}_k \left[\left\{(y_k+1) \Big/ (x_k+1)\right\} - 1\right]. \end{aligned} \tag{22}$$

The estimated variance of $N_k - T_k$ conditional on y_k contacts, is equal to the variance given in (19) plus

$$\hat{\mathrm{Var}}(\hat{T}_k \mid y_k) = \hat{\pi}_k(1-\hat{\pi}_k)S_k^2 \Big/ (r_k - 1) \,. \qquad (23)$$

Mark-recapture estimators are design-based rather than model-based; they do not rely on a probabilistic model, such as exponential or Weibull, for the growth of the population whose size the researcher wishes to estimate. Therefore, mark-recapture population estimates can provide an independent benchmark against which to compare estimates based on different probabilistic models.

5. Mixing Matrices

We (Castillo-Chavez, Crawford, and Schwager, see Crawford *et al.*, 1990) found in our recent survey of social/sexual mixing among Cornell undergraduates (CUSSP) that over a period of two months a larger fraction of females (111/253) than males (21/249) reported sexual activity. Those males that reported sexual activity during this two-month period had an average of 2.5 sexual partners while females reported about 1.4 sexual-partners during the same period of time. Table 1 shows that female Cornell undergraduate respondents that were sexually active during September and October 1989 had about 50% of their sexual contacts with Cornell undergraduates, and the remaining 50% with "outsiders", which includes Cornell graduate students, staff, faculty (GSF), and indiviudals not affiliated with Cornell (non-CU). The diagonal elements in Table 1 are always larger; that is, we see a strong "like-with-like" component. Further, the upper triangular elements are larger than the lower triangular elements, that is, females mix more often with upperclassmen, usually older males. We further note that Table 1 is not a complete mixing matrix because the population is not closed. Estimates on the sizes and sexual activity of the external mixing populations are still required.

Using the mark-recapture methods described above in conjunction with our survey data and using the mixing axioms to input "missing" values, one can complete plausible mixing matrices (see Figures 1 and 2 below).

The two figures constitute a sample of the type of matrices that one may get as one closes the network with the help of the estimate of the size of the sexually-active population that has sexual or social contact with the Cornell undergraduate population and with the use of the axioms for mixing (that is we "force" the conservation of partners law). Of course, many other matrices are possible for the same data. However, the continued repetition of the above procedure yields the same qualitative picture if the number of groups is not too small or too large. The writing of this article had the purpose of describing the nature of the

mixing problem and an outline of a possible solution. The specific details involved in the construction of these mixing matrices will be published elsewhere.

Total number of female respondents	Female respondent – male partners								
	Male Cornell undergraduate partners					Other male partners			
	Freshman	Sophomore	Junior	Senior	*Subtotal* ($x_{mj.}$)	GSF	non-CU	*Subtotal* ($y_{mj.}-x_{mj.}$)	*Total Contacts* ($y_{mj.}$)
Freshman 20	5 (*17.2%*)	4 (*13.8%*)	4 (*13.8%*)	2 (*6.9%*)	15 (*51.7%*)	0 (*0%*)	14 (*48.3%*)	14 (*48.3%*)	29 (*100%*)
Sophomore 26	1 (*2.9%*)	12 (*34.3%*)	6 (*17.1%*)	2 (*5.7%*)	21 (*60.0%*)	5 (*14.3%*)	9 (*25.7%*)	14 (*40.0%*)	35 (*100%*)
Junior 36	1 (*2.4%*)	4 (*9.8%*)	11 (*26.8%*)	7 (*17.1%*)	23 (*56.1%*)	0 (*0%*)	18 (*43.9%*)	18 (*43.9%*)	41 (*100%*)
Senior 28	0 (*0%*)	0 (*0%*)	1 (*3.5%*)	7 (*24.1%*)	8 (*27.6%*)	9 (*31.0%*)	12 (*41.4%*)	21 (*72.4%*)	29 (*100%*)
Total 111	7 (*5.2%*)	20 (*15.0%*)	22 (*16.4%*)	18 (*13.4%*)	67 (*50.0%*)	14 (*10.4%*)	53 (*39.6%*)	67 (*50.0%*)	134 (*100%*)

Table 1. An extract from data derived from the sexual survey among Cornell undergraduates, Fall 1989. Data show the sexual mixing pattern of female respondents. All survey respondents classify their Cornell partners by college class. Therefore, three subscripts are required to tabulate Cornell sexual partners properly: k denotes the sex of the partner, j denotes the college class of the respondent and i denotes the college class of the partner. Above, $y_{kj.}$ denotes the total number of sexual contacts with individuals of sex k during the two month period reported by individuals of class j, and $x_{kj.}$ denotes the total number of sexual contacts with Cornell undergraduates of sex k reported by respondents of class j.

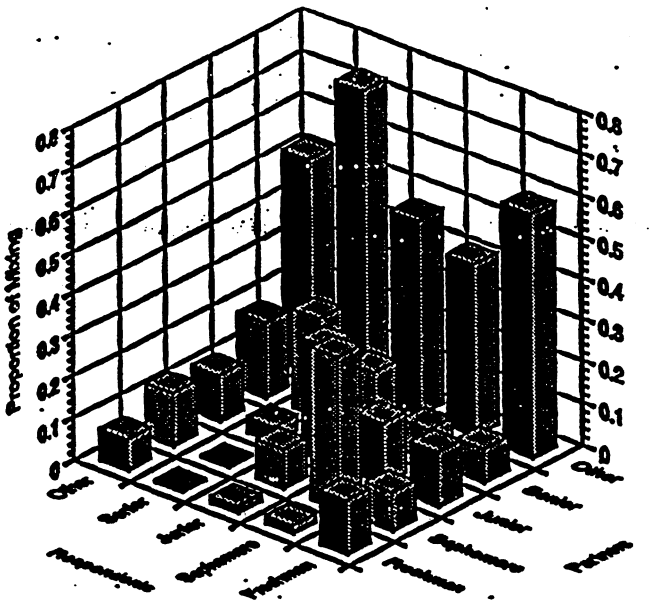

Fig. 1. Mixing matrix $\mathcal{Q}(0)$ from survey data.

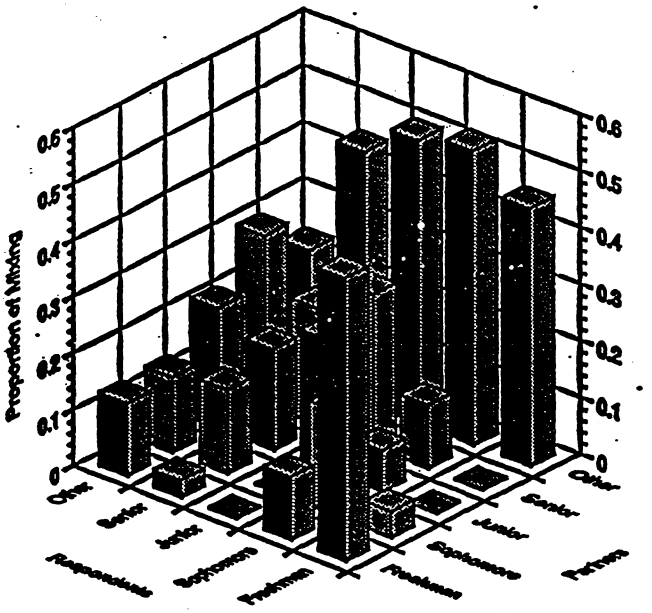

Fig. 2. Mixing matrix $\mathcal{P}(0)$ from survey data.

6. Conclusions

Several conclusions can be drawn from our preliminary efforts in estimating the contact structure of a heterosexually mixing population. A large number of groups will present very difficult estimation problems because of the large number of parameters involved, and because some of the mixing probabilities will be near zero. But a very small number of groups will not capture the level of heterogeneity needed to understand the consequences of extreme sexual behaviors. Given the difficulties involved in estimating mixing probabilities, levels of sexual activity, and group preferences or affinities from data, we see very little *practical* use of models that involve more than 8 groups, and strongly recommend the use of 4 to 6 groups for detailed epidemiological studies. Here the "correct" scale is determined by data. The selection of useful models for detailed epidemiological studies has to be guided by our clear understanding of the key features in HIV transmission. While very detailed epidemiological models may not be useful for specialized epidemiological investigations, their study is central to the theoretical understanding of the importance that several epidemiological and sociological factors – including long periods of incubation, variable infectivity, age-structure, and social mixing – may have on the dynamics of HIV/AIDS. Numerical and analytical studies (see Thieme and Castillo-Chavez, 1989, 1990) of detailed models provide the basis for the selection of the less detailed models that are required to address specific practical questions. In summary, theoretical studies in combination with data help us sort out the boundaries between practice and theory. Theoretical studies, through mathematical models, help us rank the importance of biological detail and guide us in choosing, *a priori*, the most appropriate scales at which to address specific biological questions.

7. Acknowledgements

This research has been partially supported by NSF grant DMS-8906580, NIAID Grant R01 A129178-02, and Hatch project grant NYC 151-409, USDA to CC-C. Also many thanks to Dr. K. Dietz for his valuable comments.

References

Anderson, R. M., May, R. M. and Medley, G. F. (1986). A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS. *IMA J. Math. Med. Biol.* **3**, 229-263.

Anderson, R. M., Blythe, S. P., Gupta, S. and Konings, E. (1989). The transmission dynamics of the Human Immunodeficiency Virus Type 1 in the male homosexual community in the United Kingdom: the influence

of changes in sexual behavior. *Phil. Trans. R. Soc. Lond.* **B 325**, 145-198.

Bailey, N. T. J. (1951). On estimating the size of mobile populations from recapture data. *Biometrika* **38**, 293-306.

Blythe, S. P. and Castillo-Chavez, C. (1989). Like-with-like preference and sexual mixing models. *Math. Biosci.* **96**, 221-238.

Blythe, S. P., Castillo-Chavez, C. and Casella, G. (1992). Empirical methods for the estimation of the mixing probabilities for socially-structured populations from a single survey sample. *Mathematical Population Studies* (in press).

Blythe, S. P., Castillo-Chavez, C., Palmer, J. and Cheng, M. (1991). Towards unified theory of mixing and pair formation. *Math. Biosci.* **107**: 379-405.

Blythe, S. P., Cooke, K. and Castillo-Chavez, C. (1991). Autonomous risk-behavior change, and non-linear incidence rate, in models of sexually transmitted diseases. *Biometrics Unit Technical Report* **BU-1048-M**, Cornell University, Ithaca, NY.

Busenberg, S. and Castillo-Chavez, C. (1989). Interaction, pair formation and force of infection terms in sexually-transmitted diseases. In *Mathematical and Statistical Approaches to AIDS Epidemiology,* C. Castillo-Chavez (ed.), Lecture Notes in Biomathematics 83, 289-300. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Busenberg, S. and Castillo-Chavez, C. (1991). A general solution of the problem of mixing subpopulations, and its application to risk- and age-structured epidemic models for the spread of AIDS. *IMA J. of Mathematics Applied in Med. and Biol.* **8**, 1-29.

Castillo-Chavez, C. (1989). Review of recent models of HIV/AIDS transmission. In *Applied Mathematical Ecology*, S. A. Levin, T. G. Hallam and L. J. Gross (eds.), Biomathematics 18, 253-262. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Castillo-Chavez, C. (ed.). (1989). *Mathematical and Statistical Approaches to* AIDS *Epidemiology*, Lecture Notes in Biomathematics 83. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Castillo-Chavez, C. and Blythe, S. P. (1989). Mixing framework for social/sexual behavior. In *Mathematical and statistical approaches to AIDS epidemiology,* C. Castillo-Chavez (ed.), Lecture Notes in Biomathematics 83, 275-288. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Castillo-Chavez, C. and Busenberg, S. (1991). On the solution of the two-sex moxong problem. In *Proceedings of the International Conference on Differential Equations and Applications to Biology and Population Dynamics*, S. Busenberg and M. Martelli (eds.), Lecture Notes in Biomathematics 92, 80-98. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong, Barcelona, Budapest: Springer-Verlag.

Castillo-Chavez, C., Busenberg, S. and Gerow, K. (1991). Pair formation in structured populations. In *Differential Equations with Applications in Biology, Physics and Engineering*, J. Goldstein, F. Kappel and W. Schappacher (eds.), 47-65. New York: Marcel Dekker.

Castillo-Chavez, C., Cooke, K. L., Huang, W. and Levin, S. A. (1989a). Results on the dynamics for models for the sexual transmission of the human immunodeficiency virus. *Applied Math. Letters* 2, 327-331.

Castillo-Chavez, C., Cooke, K. L., Huang, W. and Levin, S. A. (1989b). On the role of long incubation periods in the dynamics of HIV/AIDS, Part 2: Multiple group models. In *Mathematical and Statistical Approaches to* AIDS *Epidemiology*, C. Castillo-Chavez (ed.), Lecture Notes in Biomathematics 83, 200-217. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Centers for Disease Control. (1985). Self-reported behavioral change among gay and bisexual men, San Francisco. *MMWR* **34**, 613-615.

Crawford, C. M., Schwager, S. J. and Castillo-Chavez, C. (1990). A methodology for asking sensitive questions among college undergraduates. *Biometrics Unit Tech. Report* **BU-1105-M**, Cornell University, Ithaca, New York.

Dietz, K. (1988). On the transmission dynamics of HIV. *Math. Biosci.* **90**, 397-414.

Dictz, K. and Hadeler, K. P. (1988). Epidemiological models for sexually transmitted diseases. *J. Math. Biol.* **26**, 1-25.

Fredrickson, A. G. (1971). A mathematical theory of age structure in sexual populations: Random mating and monogamous marriage models. *Math. Biosci.* **20**, 117-143.

Gupta, S., Anderson, R. M. and May, R. M. (1989). Networks of sexual contacts: implications for the pattern of spread of HIV. *AIDS* **3**, 1-11.

Hadeler, K. P. (1989a). Pair formation in age-structured populations. *Acta Applicandae Mathematicae* **14**, 91-102.

Hadeler, K. P. (1989b). Modeling AIDS in structured populations. *47th Session of the International Statistical Institute, Paris, August /September.* Conference Proc. C1-2.1, 83-99.

Hadeler, K. P. and Ngoma, K. (1990). Homogeneous models for sexually

transmitted diseases. *Rocky Mountain Journal of Mathematics* **20**, 967-986.

Hethcote, H. W. and Yorke, J. A. (1984). *Gonorrhea transmission dynamics and control*, Lect. Notes Biomath. 56. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Hethcote, H. W. and Van Ark, J. W. (1987). Epidemiological models for heterogeneous populations: proportionate mixing, parameter estimation, and immunization programs. *Math. Biosci.* **84**, 85-111.

Hethcote, H. W., Van Ark, J. W. and Karon, J. M. (1991). A simulation model of AIDS in San Francisco, II. Simulations, therapy, and sensitivity analysis. *Math, Biosci.* **106**, 223-247.

Hethcote, H. W. and Van Ark, J. W. (1992). Weak linkage between HIV epidemics in homosexual men and intravenous drug users (in this volume).

Hyman, J. M. and Stanley, E. A. (1988). Using mathematical models to understand the AIDS epidemic. *Math. Biosci.* **90**, 415-473.

Hyman, J. M. and Stanley, E. A. (1989). The effect of social mixing patterns on the spread of AIDS. In *Mathematical Approaches to Problems in Resource Management and Epidemiology*, C. Castillo-Chavez, S. A. Levin and C. A. Shoemaker (eds.), Lect. Notes Biomath. 81, 190-219. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Huang, W., Cooke, K. and Castillo-Chavez, C. (1992). Stability and bifurcation for a multiple group model for the dynamics of HIV/AIDS transmission. *SIAM J. of Applied Math.* (in press).

Jacquez, J. A., Simon, C. P., Koopman, J., Sattenspiel, L. and Perry, T. (1988). Modeling and analyzing HIV transmission: the effect of contact patterns. *Math. Biosci.* **92**, 119-199.

Jacqez, J. A., Simon, C. P. and Koopman, J. (1989). Structured mixing: heterogeneous mixing by the definition of mixing groups. In *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez (ed.), Lecture Notes in Biomathematics 83, 301-315. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Kendall, D. G. (1949). Stochastic processes and population growth. *Roy. Statist. Soc., Ser.* **B 2**, 230-264.

Keyfitz, N. (1949). The mathematics of sex and marriage. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, Vol. IV: Biology and Health*, 89-108.

McFarland, D. D. (1972). Comparison of alternative marriage models. In

Population Dynamics, T. N. E. Greville (ed.), 89-106. New York, London: Academic Press.

Martin, J. L. (1986a). AIDS risk reduction recommendations and sexual behavior patterns among gay men: a multifactorial categorical approach to assessing change. *Health Educ. Qtly.* **13**, 347-358.

Martin, J. L. (1986b). The impact of AIDS in gay male sexual behavior patterns in New York City. *Am. J. of Pub. Health* **77**, 578-581.

McKusick, L., Horstman, W. and Coates, T. J. (1985a). AIDS and sexual behavior reported by gay men in San Francisco. *Public Health Reports* **75**, 493-496.

McKusick, L., Wiley, J. A., Coates, T. J., Stall, R., Saika, B., Morin, S., Horstman, C. K. and Conant, M. A. (1985b). Reported changes in the sexual behavior of men at risk for AIDS, San Francisco, 1983-1984: the AIDS behavioral research project. *Public Health Reports* **100**, 622-629.

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

Palmer, J. S., Castillo-Chavez, C. and Blythe, S. P. (1991). State-dependent mixing and state-dependent contact rates in epidemiological models. *Biometrics Unit Technical Report* **BU-1122-M**, Cornell University, Ithaca, NY.

Parlett, B. (1972). Can there be a marriage function?. In *Population Dynamics*, T. N. E. Greville (ed.), 107-135. New York, London: Academic Press.

Pollard, J. H. (1973). The two-sex problem. In *Mathematical Models for the Growth of Human Populations*, Chapter 7. Cambridge University Press.

Rubin, G., Umbach, D., Shyu, S-F. and Castillo-Chavez, C. (1991). Application of capture-recapture methodology to estimation of size of population at risk of AIDS and/or other sexually-transmitted diseases. *Biometrics Unit Technical Report* **BU-1112-M**, Cornell University, Ithaca, NY.

Saltzman, S. P., Stoddard, A. M., McCusker, J., Moon, M. W. and Mayer, K. H. (1987). Reliability of self-reported sexual behavior risk factors for HIV infection in homosexual men. *Public Health Reports* **102**, 692-697.

Sattenspiel, L. and Castillo-Chavez, C. (1990). Environmental context, social interactions, and the spread of HIV. *American Journal of Human Biology* **2**, 397-417.

Seber, G. A. F. (1082). *The estimation of animal abundance and related parameters.* New York: MacMillan.

Shilts, R. (1987). *And the band played on.* New York: St. Martin's Press.

Thieme, H. R. and Castillo-Chavez, C. (1989). On the role of variable infectivity in the dynamics of the human immunodeficiency virus epidemic. In *Mathematical and Statistical Approaches to* AIDS *Epidemiology*, C. Castillo-Chavez (ed.), Lecture Notes in Biomathematics 83, 157-176. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Thieme, H. R. and Castillo-Chavez, C. (1990). On the possible effects of infection-age-dependent infectivity in the dynamics of HIV/AIDS. *Biometrics Unit Technical Report* **BU-1102-M**, Cornell University, Ithaca, NY.

Waldstatter, R. (1989). Pair formation in sexually transmitted diseases. In *Mathematical and Statistical Approaches to AIDS Epidemiology*, C. Castillo-Chavez (ed.), Lecture Notes in Biomathematics 83, 260-274. Berlin, Heidelberg, New York, London, Paris, Tokyo, Hong Kong: Springer-Verlag.

Winkelstein, W. Jr., Wiley, J. A., Padian, N. S., *et al.* (1988). The San Francisco Men's Health Study, continued decline in HIV seroconversion rates among homosexual/bisexual men. *Am. J. Pub. Health* **78**, 1472-1474.