tions, among males, of the four classes should be 74%, 13%, 6%, and 7%. The correctness of these proportions, compared to, for example, the results from a population study of sexual behavior, would then speak in favor of the proportional mixing hypothesis. In a recent study of sexual behavior of young Swedes [1], the sexual history during one year was investigated. The results are presented in such a fashion that it is only possible to calculate that, conditional on having at least one partner, the proportion of young men having exactly one partner during a year was 67%, while the remaining 33% had more than one partner. Considering that the one-year figure for exactly one partner should be lower than a half-year figure, it seems to us that the model-derived estimate and the observed population based estimate are in agreement.

This study was supported by the AFA Insurance Company, Stockholm.

REFERENCES

1 M. Göthberg, J. Giesecke, G. Scalia-Tomba, and P. Tüll, Sexual contact patterns among young Swedes—a population study on Gotland 1988 (in Swedish); Technical Report 12/89, Depart. of Infectious Disease Control (MME), Karolinska Hospital, Stockholm, Sweden.

2 H. H. Handsfield, L. L. Jasman, P. L. Roberts, V. W. Hanson, R. L. Kothenbeutel, and W. E. Stamm, Criteria for selecting screening for *Chlamydia trachomatis* infection in women attending family planning clinics, *JAMA* 255:1730–1734 (1986).

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

4 K. Ramstedt, L. Forssman, J. Giesecke, and G. Johannisson, Epidemiological characteristics of two different populations of women with *Chlamydia trachomatis* infection and their male partners, accepted for publication in *Sex. Transm. Dis.* 1991.

5 K. Ramstedt, L. Forssman, and F. Granath, Risk factors for *Chlamydia trachomatis* infection in 6,810 young women attending family planning clinics for contraceptive advice in Gothenburg in 1988, accepted for publication in *Int. J. STD. AIDS.* 1991.

A Log-Linear Modeling Framework for Selective Mixing

MARTINA MORRIS
Department of Sociology, Columbia University, New York, New York 10027

Received 12 November 1990; revised 30 May 1991

ABSTRACT

Nonrandom mixing can significantly alter the diffusion path of an infectious disease such as AIDS that requires intimate contact. Recent attempts to model this effect have sought a general framework capable of representing both simple and arbitrarily complicated mixing structures, and of solving the balancing problem in a nonequilibrium multigroup population. Log-linear models are proposed here as a general framework for solving the first problem. This approach offers several additional benefits: The parameters used to govern the mixing have a simple, intuitive interpretation, the framework provides a statistically sound basis for the estimation of these parameters from mixing-matrix data, and the resulting estimates are easily integrated into compartmental models for diffusion. A modified selection model is proposed to solve the second problem of generalizing the selection process to nonequilibrium populations. The distribution of contacts under this model is derived and is found to satisfy the assumptions of statistical inference for log-linear models. Together these techniques provide an integrated and flexible framework for modeling the role of selective mixing in the spread of disease.

INTRODUCTION

The role of selective mixing in the transmission dynamics of infection has become the focus of a sustained modeling effort since the emergence of AIDS. Most of these developments have been based on integrating a contact or mixing matrix into compartmental models of transmission via the infection rate term. There are several reviews of the developments in this field [17, 19, 28]. In general, the development has been from fairly simple representations of mixing structures, such as proportional [4, 27] or preferred mixing [3, 18, 29], to more general representations of arbitrarily complicated structures [7, 17, 25]. In moving from simple to general formulations, two problems arise. The first is complexity. For an $n \times n$ contact matrix, there are potentially n^2 parameters that must be identified and integrated into the infection rate term. When the subpopulation sizes are not stable, or when activity levels are allowed to change, more than n^2

©Elsevier Science Publishing Co., Inc., 1991
655 Avenue of the Americas, New York, NY 10010 0025-5564/91/$03.50

parameters are needed to account for the process of selection in a nonequilibrium structure. Thus, in some of the general models, the total number of parameters used to model the process is as much as four times the number of cells in the matrix.

This raises the second problem, that of selecting and justifying the parameter settings. It would seem hard to argue against basing these settings on empirical data, though the lack of such data often makes this an unattainable ideal. To base the parameters on data, however, a framework for estimation is required. A statistically sound estimation framework is noticeably lacking from most of the papers that propose selective mixing models.

There is no unique representation (or model) for a given contact matrix. Some formulations are inherently limited, like proportional or preferred mixing, but even these can be parametrized in different ways. Arbitrary mixing functions, which have proportional or preferred mixing as special cases, can also be represented in many ways. With enough manipulation, the parameters used in one model can always be expressed in terms of the parameters used in another. Though generality should be considered a minimum criterion for an acceptable modeling framework, a particular representation cannot claim superiority solely on the basis of being general. Within the class of general models, however, a claim to superiority can be based on a model's ability to solve the two problems identified above: manageable complexity and a framework for estimation. Log-linear models can make this claim.

LOG-LINEAR MODELS OF MIXING MATRICES

Log-linear modeling is a multivariate technique developed for the statistical analysis of discrete data [5, 12, 22]. In this approach, the contact matrix is treated as a cross-classified table of counts. Under a variety of sampling assumptions, the cell frequencies can be represented by a large and very flexible class of models. Although this paper focuses on the power of log-linear models to summarize the structure of selective mixing in a contact matrix, it should be emphasized that the models are not simply tools for describing aggregate outcomes. They are instead causal models of individual behavior, fully stochastic representations in which each individual contributes a term to the overall likelihood function for the matrix. The underlying behavioral model must satisfy certain distributional assumptions in order for statistical inference to be valid. Within this constraint, however, there is still a great deal of flexibility available to model the individual and social forces that generate selective mixing. An example is derived in the appendix.

Take a very simple contact matrix such as the 2×2 table in Figure 1. This table could represent friendship or sexual pairing combinations. If

	Partner		
Subject	Male	Female	Total
Male	x_{11}	x_{12}	x_{1+}
Female	x_{21}	x_{22}	ix_{2+}
Total	x_{+1}	x_{+2}	T

FIG. 1. A schematic representation of a simple 2×2 contact or mixing matrix. The cell counts, x_{ij}, are the number of couples with subject from group i and partner from group j.

there were no bias in partner selection, that is, random or proportional mixing, the expected value of the cell entries would be a function only of the row and column frequencies. This is traditionally modeled as

$$m_{ij} = \frac{x_{i+}}{T}\left(\frac{x_{+j}}{T}\right)T, \qquad i = 1,2;\ j = 1,2, \tag{1}$$

where m_{ij} is the expected cell count under this model. In statistical terms, this is known as the model of independence. The basic intuition that underlies log-linear methods is that this multiplicative relation can be transformed into a linear relation on the log scale, making it more tractable and giving it a functional form that bears a close resemblance to traditional analysis of variance notation:

$$\log m_{ij} = \log T + \log\left(\frac{x_{i+}}{T}\right) + \log\left(\frac{x_{+j}}{T}\right). \tag{2}$$

In more general notation, m_{ij} can be expressed in the form

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)}. \tag{3}$$

The u terms on the right-hand side are usually reparametrized in order to increase flexibility, ease of estimation, and interpretability. The notation here is taken from Bishop et al. [5]. Other versions can be found in Goodman [12], Agresti [1], and McCullagh and Nelder [22].)

As the form of the model suggests, this technique provides a decompositional approach to the analysis of matrix structure: The fitted cell counts are decomposed into the contributions of several factors. The u term is a reference category, and the remaining components are expressed in terms of deviations from this reference category. The category that serves as the reference level depends on the parametrization chosen [e.g., the grand mean, a particular cell, or, in Equation (2), the total number]. A common

choice of parametrization, sometimes called "symmetric constraints" [22, p. 50], is analogous to the usual ANOVA model; the grand mean is the reference category, and the marginal terms are deviations from this mean. In this case, u is the mean of the logarithms of the expected counts,

$$u = \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J}\log m_{ij}; \tag{4}$$

$u_{1(i)}$ is the departure of the mean of the logarithms of the expected counts in the cells at the ith level of the first variable from the grand mean,

$$u_{1(i)} = \frac{1}{J}\sum_{j=1}^{J}\log m_{ij} - u; \tag{5}$$

and equivalently for the columns,

$$u_{2(j)} = \frac{1}{I}\sum_{i=1}^{I}\log m_{ij} - u. \tag{6}$$

Because $u_{1(i)}$ and $u_{2(j)}$ represent deviations from the grand mean, the constraints for this model are

$$\sum_{i=1}^{I} u_{1(i)} = \sum_{j=1}^{J} u_{2(j)} = 0, \tag{7}$$

which is the origin of the name "symmetric constraints."

Another common choice of parametrization simply uses the fitted value of the (1, 1) cell, $\log m_{11}$, as the reference category, with differences between the $\log m_{i1}$ and $\log m_{1j}$ cells to represent row and column effects, respectively. The constraints in this version set level 1 of the row and column effects to zero. This will be referred to as the *first-level constraints model*.

If the cell counts x_{ij} of this table are sufficiently different from the fitted values under the model of independence, this implies an interaction between the gender of one partner and the gender of the other—for example, that males are more likely to pair with females, and females with males. In the 2×2 table, there is one degree of freedom to model this interaction once the row and column margins have been fitted. The interaction is a function of

the cross product or odds ratio α, using the symmetric parametrization

$$u_{12(ij)} = \frac{1}{IJ}\log \alpha_{ij}, \tag{8}$$

where, in a 2×2 table,

$$\alpha_{11} = \alpha_{22} = \alpha_{12}^{-1} = \alpha_{21}^{-1} = \frac{m_{11}m_{22}}{m_{12}m_{21}}. \tag{9}$$

The constraints implied by this equality are that interaction terms in the model sum to zero across rows and columns. This interaction term is added to the right-hand side of Equation (3) to produce the expected cell counts. Note that there are now a total of nine parameters in the full model, whereas there are only four cells in the table. Some of the parameters are redundant, and the constraints (there are five of them here) eliminate this redundancy (cf. [22, pp. 49–51] for a lucid discussion of this issue). Each parametrization is defined by a unique set of constraints, so choosing a parametrization is equivalent to choosing a set of constraints.

In this simple 2×2 example, the interpretation of the interaction term is quite straightforward: It measures the relative likelihood of counts falling in the diagonal cells, and this is true for both the symmetric and the first-cell parametrizations. If there is a heterosexual bias in pairing, α_{11} is less than 1. With symmetric constraints, the $u_{12(ij)}$ term would be negative for the diagonal cells and positive for the off-diagonal cells. The value of this effect is also easy to interpret: On the frequency scale, that is, $e^{u_{12(ij)}}$, it represents the multiplicative increment of the cell count that is produced by the interaction. When first-level constraints are used, the interaction term is not symmetrically assigned, but it remains a function of the odds ratio. Here, $u_{11(ij)} = u_{12(ij)} = u_{21(ij)} = 0$ and $u_{22(ij)} = \log \alpha$.

Both here and in larger tables the odds ratio provides a convenient measure of the relative propensity toward self-selection for any two groups. It has some desirable properties, foremost of which is that it is invariant to changes in the margins of the matrix. This makes it a good measure for comparing group selection patterns over time, as it is not confounded with changes in group sizes or activity levels.

In larger tables, interaction terms can be used to provide tremendous flexibility for modeling the structure of the contact matrix. With the row and column margins fixed, there are $(n-1)^2$ degrees of freedom left for specifying interaction parameters. The standard set of interaction terms uses all of these degrees of freedom, producing a saturated model in which the cell counts are perfectly reproduced. The framework permits a much more parsimonious representation, however, using generalized interaction terms

similar to contrasts in the ANOVA format. Generalized interaction terms involve the specification of factor design matrices to assign specific cells to different levels of the hypothesized factor. The factor level can specify as few as one cell, for example, a contrast between white–white pairings and all other racial pairs, or it can specify groups of cells, for example, a contrast between older men–younger women pairs and younger men–older women pairs (a contrast between the upper and lower triangle of an age-mixing matrix), or between same-group pairs and mixed-group pairs (diagonal vs. nondiagonal cells in the matrix). The only requirement is that every cell in the matrix be assigned to one and only one level of the factor. When generalized interaction terms are used, the estimates of the cell frequencies will fit the generalized marginal quantities exactly; that is, for any factor level S,

$$\sum_{i \in S} \hat{m}_{ij} = \sum_{i \in S} x_{ij},$$

or in other words, the sum of the residuals for any factor level will be zero. If departures from random mixing in a table are systematic and regular, a small number of interaction parameters can be used to summarize the pattern. Where a particular cell has a disproportionately high or low count, a single interaction parameter can be used to isolate the effect. The generalized interaction parameters are thus responsive to complex and interesting hypotheses regarding the structure of the data.

These models can be generalized to multidimensional tables (e.g., three- and higher way cross-tabulations) and polytomous factor levels. Techniques have also been developed to summarize ordinal factor levels [14], and combinations of nominal and ordinal factors, for example, race and age, can be used simultaneously. Thus, a wide class of log-linear models exists for the multivariate analysis of matrices.

Techniques of estimation and model assessment have also been developed. Note that in Equations (4)–(9) the parameters are functions of the expected counts, not of the data directly. Estimation of these parameters is accomplished using iterative routines. The maximum likelihood estimates (MLEs) for the expected cell counts are the same under several different sampling assumptions, including Poisson, multinomial (total T fixed), and product multinomial (row margins fixed.)[1] The MLEs of the u's can be obtained by iterative proportional fitting routines, and the fit of the model to

[1]The distributional assumptions are an important issue. They are dependent on the behavioral model that is assumed to generate the observations in the mixing matrix. This issue is discussed in the Appendix.

the data can be assessed using the likelihood ratio statistic

$$G^2 = 2 \sum_{ij} x_{ij} \log \frac{x_{ij}}{m_{ij}}. \tag{10}$$

This statistic is distributed asymptotically as χ^2 with the degrees of freedom appropriate for the model. It is generally preferred to the χ^2 statistic because it can be decomposed into components due to each term in the model. As a result, nested models can be compared using the conditional likelihood ratio statistic

$$G^2(2 \mid 1) = G^2(2) - G^2(1), \tag{11}$$

where $G^2(1)$ is the deviance for a baseline model with an adequate fit (e.g., a saturated model) and $G^2(2)$ is the deviance for a model with a subset of the u terms contained in model 1. The conditional $G^2(2 \mid 1)$ is distributed as χ^2 with degrees of freedom equal to the difference in degrees of freedom between models 2 and 1 [5, pp. 125–126]. The trade-off between accuracy and parsimony can therefore be systematically assessed. There are several computer programs available (e.g., GLIM, SAS, SPSS, BMDP) that can be used to calculate the estimates and the goodness-of-fit statistics for these models. The validity of the statistical inference depends on sampling and on the underlying behavioral model. These are discussed in more detail below and in the appendix.

Log-linear models have been used to analyze various kinds of matrix data in the sociological literature. For example, the 1985 General Social Survey data on friendship networks have been arrayed in a matrix defined by respondent characteristics (rows) and the reported characteristics of their friends (columns) and analyzed for the degree of homogeneity (homophily) in friendship choices [6, 24]. Log-linear models have also been used extensively in the analysis of social mobility tables, square matrices of fathers' and sons' occupations [11, 13, 15, 30, 31]. In the mixing matrix context, the parameters estimated by the models are interpretable and are sensitive to sociological hypotheses regarding the patterns of differential association.

EXAMPLES

The social characteristics that produce selective mixing can be divided into two types for log-linear modeling purposes: nominal variables, such as sex, race, or marital status, and ordinal variables such as age and socioeconomic status. The patterns of selective mixing that might arise from these two types of variables could well be expected to be different. Nominal

variables such as gender or race might produce sharp, discontinuous distinctions among potential partner groups. Ordinal variables such as age or socioeconomic status are more likely to produce graduated, distance-like distinctions and potential edge effects. The log-linear models that one would use for nominal and ordinal variables reflect this difference, providing ANOVA-like parameters for nominal variables and regression-like parameters for ordinal variables.

NOMINAL PARAMETERS

Nominal characteristics such as gender, race, marital status, and religion, clearly play an important role in differential association of all kinds, from acquaintance and friendship [10, 33] to cohabitation and marriage [2]. There can be no doubt that such characteristics also play a leading role in defining the appropriateness and acceptability of sexual partners.

For some nominal characteristics, like race or religion, some form of preferred mixing seems plausible. Preferred mixing predicts greater density on the diagonals of the matrix, that is, positive assortative mating or homophily. As it is usually modeled [18], preferred mixing takes an additive form on the frequency scale, $\rho_i + (1-\rho_i)C_i$ for the probability of an ingroup partnership and $(1-\rho_i)C_j$ for the probability of an outgroup partnership, where $\rho_{(\cdot)}$ is the group-specific fraction of contacts reserved for ingroup partners and $C_{(\cdot)}$ is the group-specific fraction of all unreserved contacts. The analogous approach in the log-linear framework is a multiplicative model on the frequency scale, where the ingroup bias multiplies the probability of a contact on the diagonal. If the within-group selection bias is the same for each group, it can be modeled with a single generalized interaction parameter. This model, which could be called *uniform homophily*, takes the form

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \theta_k, \tag{12}$$

where, under the first-level constraints with the off-diagonal cells as level 1 of the generalized interaction factor,

$$\theta_k = \begin{cases} 0, & i \neq j, \\ \theta, & i = j. \end{cases}$$

Under this model, each group shares a common tendency toward self-selection, θ, and does not otherwise discriminate among potential partners outside its own group. As in the simple proportion reserved model, the off-diagonal cells are filled by proportional mixing, net of the contacts reserved for diagonal cells. The estimated homophily parameter $\hat{\theta}$ will

equal the common odds ratio

$$\hat{\theta} = \hat{m}_{kk}\hat{m}_{ij}/\hat{m}_{ik}\hat{m}_{kj}, \qquad i, j \neq k;\ i \neq j, \tag{13}$$

and the odds ratio formed by allowing $i = j$ will be double this value. The value of θ ranges from $-\infty$ to ∞, and this range describes the continuum from perfect negative assortative mating to perfect positive assortative mating ("restricted mixing"). When $\theta = 0$, a hypothesis that can be tested statistically, contacts are governed by proportional or random mixing. This model uses only one degree of freedom to capture the pattern of within-group selection bias. It is equivalent to the "constant inbreeding model" used by Marsden [24] to describe friendship networks, and to Goodman's "uniform inheritance model" for social mobility tables [11].

If the propensity for within-group selection varies among groups, one can relax the assumption that θ_k is constant for each diagonal cell. In this case, a model we might call *differential homophily*,

$$\theta_k = \begin{cases} 0, & i \neq j, \\ \theta_i, & i = j. \end{cases} \tag{14}$$

This allows the strength of the within-group preference to vary across groups but continues to distribute all remaining outgroup contacts proportionately; that is, it introduces no selective preferences for different outgroups. Given the proportional allocation across outgroups, this model uses n degrees of freedom for estimating the homophily parameters. If some groups display similar homophily biases, then fewer parameters may be needed. The alternative hypotheses of uniform, complete differential, or partial differential homophily can be statistically tested using a combination of the standard errors of the estimated coefficients $\hat{\theta}_i$ and the conditional likelihood ratio statistic.

If the proportional allocation across outgroups is relaxed, there remain $(n-2)(n-1)/2$ degrees of freedom for estimating outgroup preferences of various sorts. This flexibility is likely to be very useful in the modeling of selection patterns by gender and sexual preference. These patterns can be expected to be quite strong, though not systematic in the same way as uniform and differential homophily [26].

For nominal parameters, then, log-linear models provide a large and flexible set of descriptive hypotheses regarding the patterns of partner selection in a contact matrix. The models represent the proportional to restrictive mixing continuum in a simple and intuitive way. They also

generalize to arbitrary mixing functions, so that ingroup and outgroup preferences of all kinds can be examined. The conditional likelihood ratio statistic provides a basis for choosing among competing models for any particular mixing matrix. One can therefore systematically identify the most parsimonious model needed to describe the structure of empirical mixing.

As an example, consider the heterosexual race and ethnicity mixing matrix of Figure 2. The data for this matrix come from the AIDS in Multiethnic Neighborhood Study (AMEN). The rows of the matrix represent males in each of four ctegories: black, latino, white, and other. The columns represent the categories of their female partners. The numbers in the cells represent the total number of partnerships between men in that row and women in that column, as reported by the men. A strong degree of preferred mixing is evident, but one could ask whether it is a uniform or differential preference, and whether the remaining outgroup contacts are randomly distributed.

The results of three log-linear models for this matrix are presented in Table 1. The first-level fitting constraints are used here, so level 1 of each factor equals zero. The first model is proportional mixing [i.e., Eq. (3)], the second is uniform homophily [Eq. (12)], and the third is differential homophily [Eq. (14)]. The row and column effect estimates are not reported in this table. A quick glance at the deviance for each model makes it clear that proportional mixing fits the matrix very poorly, uniform homophily fits somewhat better but leaves a significant amount of unexplained variation, and differential homophily fits the matrix reasonably well. Thus, at the general level, it would appear that, net of differential homophily, outgroup contacts do appear to be randomly distributed. At the same time, it is worth noting that only 27% of all contacts are made with partners outside the group.

Under the model of uniform homophily, the within-group selection bias is estimated to be $e^{2.01} = 7.5$; in other words, it is 7.5 times as likely for a contact to be made between members of the same racial or ethnic group as

	Female partner				
Male subject	Black	Latino	White	Other	Total
Black	275	23	48	25	371
Latino	21	262	85	36	404
White	23	33	497	59	612
Other	3	11	36	33	83
Total	322	329	666	153	1470

FIG. 2. These data on race and ethnicity mixing in sexual partnerships are the basis for the analysis reported in Table 1. There is clearly a strong pattern of homophily, but one would like to test whether it is uniform or differential, and whether the off-diagonal contacts are selectively or proportionally allocated. (Source: AMEN study, courtesy of Dr. James A. Wiley.)

TABLE 1

Log-Linear Models for the Race and Ethnicity Mixing-Matrix

Parameter	Main effects only	Uniform homophily	Differential homophily
Reference category	4.398	3.489	2.384
Main effects:			
Males			
Latino	0.085*	0.117*	0.468
White	0.501	0.051*	0.752
Other	−1.497	−1.473	−0.488
Females			
Latino	0.022*	−0.056	0.503
White	0.727	0.693	1.579
Other	−0.744	−0.043*	0.852
Interactions:			
Uniform		2.01	
Differntial: 2			3.233
3			2.213
4			1.493
5			.748
G^2	1284	79.3	7.1
df	(9)	(8)	(5)

*$p > 0.05$, coefficient not statistically significant.

between members of different groups. The table of residuals (not shown here) reveals that this model overestimates the number of within-group contacts for whites and others and underestimates the number for blacks and latinos.

The model of differential homophily reveals large differences in the degree of within-group selection bias among the four groups. For blacks, the bias is $e^{3.23} = 25.3$, for latinos 9.0, for whites 4.5, and for other 2.1. Each of these coefficients differs significantly from zero and from the others. The strongest bias, however, is clearly among blacks, which suggests that color establishes a fairly strong sexual segregation among heterosexuals. The selective mixing structure of this matrix is well described by the model of complete differential homophily.

MODELS FOR ORDINAL PARAMETERS

Ordinal characteristics such as age, education, and SES also play an important role in differential association. It is in this area that concepts of social distance can be taken most literally. Friendship and association

patterns have been found to display marked ordinal variation by occupational status [21] and by age and education [24]. Marriage patterns have shown similar strong selective patterns [8, 16]. As with nominal variables, there can be no doubt that such characteristics also play a role in defining appropriate and acceptable sexual partners.

Models for ordinal variables take advantage of the rank ordering of preferences, generally focusing on location relative to the main diagonal. For this reason, a natural baseline model is either the uniform or differential homophily model, which posits no selective preference off the main diagonal. The ordinal parameters then reflect departures from this pattern and represent the "distance" of the partner from the respondent.

There are two types of models that one can use for ordinal variables. The first are within the log-linear class and have become known as *diagonals parameter* models [13]. These models posit a uniform distance effect for each group of cells the same distance from the diagonal. They take the form

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \delta_k, \qquad k = |i - j| \tag{15}$$

with appropriate fitting constraints. In this form, a diagonals parameter model assumes uniform homophily and uses an additional $n-1$ parameters over the proportional mixing model. For differential homophily, the diagonal cells can be fitted exactly by allowing δ_0, the diagonal coefficient, to vary across rows.

The off-diagonal parameter estimates from these models assume that outgroup selection preferences operate uniformly across respondent categories; for example, a respondent in age category 1 is as likely to choose a partner from age category 2 as a respondent in category 3 is to choose a partner from category 4. The incremental differences across columns, between δ_1 and δ_2, need not be uniform or linear. Thus, under this model it is possible that a respondent from age category 1 may be somewhat less likely to choose a partner from age category 2 but several times less likely to choose a partner from age category 3.

The allowable variability in distance estimates makes these models quite flexible. They would be good candidates for modeling selective effects that are likely to be nonlinear on the log scale. In other cases, the distance may be uniform for each unit off the diagonal. Fewer parameters are needed to describe this kind of pattern, and the class of log-multiplicative models may be more useful.

Log-multiplicative models, a second type of models for ordinal data, express the cell counts in terms of parameters that are multiplicative on the logarithmic scale. The models in this class provide a fairly wide range of options for modeling selection patterns with regression-like parameters. The

basic form of these models is

$$\log m_{ij} = u + u_{1(i)} + u_{2(j)} + \gamma\delta_{ij}. \tag{16}$$

A single degree of freedom is used to estimate the γ coefficient. It captures the pattern of selection defined by δ_{ij}, a term that can be modeled to represent distance in many ways. In a simple model, δ_{ij} could represent the distance of a cell from the diagonal using the difference between the row and column indices,

$$\delta_{ij} = |i - j|. \tag{17}$$

Positive values of γ indicate that preference increases as one moves off the diagonal, and negative values indicate that preference decreases (the more likely pattern). When the δ levels are equally spaced, as they are here, the model assumes a linear multiplicative effect on the log scale, which becomes a simple exponential effect on the frequency scale. Net of row and column effects, observations one step off the diagonal are e^{γ} times as likely as observations on the diagonal, those two steps off the diagonal are $e^{2\gamma}$ times as likely, and those $|i - j|$ steps off are $e^{|i-j|\gamma}$ as likely.

Another convention is to set δ_{ij} equal to the product of the normalized row and column indices

$$\delta_{ij} = r_i c_j, \tag{18}$$

where

$$r_i = i - N/2, \qquad c_j = j - N/2.$$

This is known as the *uniform association* model [13]. The γ coefficient in this model can be interpreted as the common odds ratio for any two adjacent pairs of cells. Negative values for γ indicate clustering around the main diagonal; positive values indicate clustering on the opposite diagonal. If the observed selection pattern is a relatively linear function of the ordinal variable (on the log scale), γ will be close to 1. In contrast to the diagonals parameter model, here both the row and column effects are assumed to be uniform and equal. This assumption can be relaxed if the r_i and c_j "scores" are not given their usual equidistant values. But even with variable scores, γ implies a constant increment for each unit score difference.

If the ordinal categories are not assumed to be equidistant, and the true scores for the rows and columns are not known, this model can be generalized to produce the *row and column effects*, or RC, model. In the RC model the row and column scores are estimated from the data [13; pp. 139–142]. The RC model can be used to search for "edge effects,"

systematic patterns of ingroup and outgroup selection that lead to greater homophily toward the ends of an ordinal scale. Such effects have been reported in a number of areas, from status crystalization studies [20] and friendship association patterns [21], to patterns of occupational mobility [15]. Edge effects lead to greater isolation in the upper and lower reaches of the scale and more undifferentiated association in the middle. One could imagine that this would affect transmission dynamics in several ways. Undifferentiated mixing makes the populations in the middle more vulnerable to spread if infection is introduced, while sharp boundaries on the ends can either prevent spread to these populations or, if infection is introduced there, enhance the rate of transmission within the group and reduce the spread to other groups.

As an example, consider the age-mixing matrix in Figure 3. These data represent the age matching among married couples and come from 1987 Census CPR Household and Family Characteristics series [32]. The numbers represent thousands of married couples. A very strong pattern of homophily is clear in these data. However, the pattern of outgroup contacts is far from random. Outgroup marriages display two fairly simple social distance rules: First, the number of marriages declines as the difference in age between partners increases, and second, there is a stronger propensity for older men to marry younger women than for older women to marry younger men.

Table 2 presents a summary of the results of a log-linear analysis of this matrix.[2] As in the previous example, the first-level fitting constraints are used. Three generalized interaction parameters are examined, alone and in combination: homophily, distance [using Eq. (17)], and asymmetry. The asymmetry factor takes level 1 for the lower triangle and level 2 for the upper triangle.

[2]Treating the cell entries as single couples rather than thousands of couples clearly affects the significance tests. If Table 2 were analyzed using the actual number of couples, the number of observations would be 34 million rather than 34,000, and it is virtually certain that any null hypothesis would be rejected using the standard significance levels. The observations then would not represent a sample, but the universe, making the question of inference moot. The more important issue is that the power of the tests increases dramatically as an artifact of the dramatic increase in the number of observations. This is a defect of hypothesis testing in general, not simply of the method presented here. One standard approach would be to reduce the size of the test (the α level) to compensate for the increase in power, but the issue is a controversial one and has no simple answer. In practice, even a table of 34,000 couples would be a real luxury; the few data sets available with full network information usually comprise a couple of thousand couples at best. For data sets of this size, the standard levels of significance are a useful guide.

	Husband's age				
Wife's age	18–24	25–34	35–44	45–54	Total
18–24	1,902	1,982	124	10	4,018
25–34	298	9,226	3,524	317	13,365
35–44	19	674	7,854	2,894	11,441
45–54	1	21	394	5,209	5,624
Total	2,219	11,903	11,896	8,430	34,448

FIG. 3. These data on age mixing in marriage are the basis for the analysis reported in Table 2. Note the strong patterns of homophily, distance from the diagonal, and asymmetry in the matrix. (Source: U.S. Census, Household and Family Characteristics, 1987; numbers represent thousands of couples [30].)

For the purpose of comparison with the race and ethnicity mixing example, models of proportional mixing and uniform and differential homophily were examined. Again focusing only on the deviance statistics, all of these models fit quite poorly; even the differential homophily model leaves a great deal of variance unexplained in this case. The distance effect, when used as the only interaction term, does substantially better than any other single effect, but it too leaves significant residual variation. Adding distance to either of the homophily models dramatically improves the fit, but it is only with the addition of the asymmetry factor that the deviance falls below standard levels of significance. At this point, however, some of the homophily terms are no longer significant. In the model with uniform homophily, distance, and asymmetry, the homophily coefficient hovers around the 0.05 level of significance. In the model with differential homophily, only the coefficient for the second age groups is significantly different from zero. This suggests that a more parsimonious homophily factor may be appropriate.

Turning to the magnitude of the coefficients, an interesting pattern can be observed. The homophily coefficients are strong and positive when they are the only effects in the model. When distance effects are added, the homophily coefficients change sign and become weak. The change of sign does not imply lack of robustness or disassortative mating. It can be interpreted only with reference to the other factors in the model, in this case, the distance effect. To fit the steep decline in partnerships off the diagonal, the estimated distance coefficient is very strongly negative. Estimated values of -2.66 to -3.22 imply that each step off the diagonal results in more than a 90% reduction in the probability of a pairing. As in the regression context, we have fit the distance effect assuming a linear relation, here linear on the log scale. This constraint forces the distance coefficient to overestimate the number of observations on the diagonal;

TABLE 2

Log-Linear Models for the Age-Mixing Matrix

Parameter	Main effects only	Uniform homophily	Diff. homophily	Distance	UH + dist.	DH + dist.	Dist + asym.	UH + dist + asym.	DH + dist + asym.	PDH + dist + asym.
Reference Category	5.557	5.318	3.959	7.566	9.040	8.955	7.558	7.869	7.780	7.552
Main Effects:										
Female: 2	1.202	0.608	0.974	−0.068	−0.140	0.017*	0.604	0.458	0.623	0.698
3	1.046	0.340	1.155	−0.832	−1.045	−1.031	0.441	0.150*	0.176*	0.407
4	0.336	−0.410	−1.401	−1.954	−2.197	−2.135	0.031*	−0.405*	−0.332*	−0.091*
Male: 2	1.679	1.301	2.448	1.678	1.720	1.853	0.975	1.121	1.243	1.118
3	1.679	1.478	2.996	2.300	2.451	2.412	0.965	1.256	1.198	1.009
4	1.334	1.539	2.198	2.941	3.190	3.266	0.962	1.398	1.451	1.099
Interactions										
Uniform		1.942			−1.483			−0.311*		
Differential: 2			3.591			−1.405			−0.229*	
3			1.749			−1.696			−0.516	−0.239
4			0.869			−1.368			−0.185*	
5			3.806			−1.528			−0.340*	
Distance				−1.905	−3.197	−3.224	−2.517	−2.663	−2.685	−2.555
Asymmetry							1.551	1.241	1.252	1.468
G^2	37,929	11,142	8,001	1,329	62.1	46.2	25.6	21.5	5.0	8.7
df	(9)	(8)	(5)	(8)	(7)	(4)	(7)	(6)	(3)	(6)

*$p > 0.05$, coefficient not statistically significant.

UH, uniform homophily; DH, differential homophily; PDH, partial differential homophily; Dist., distance; Asym., asymmetry.

thus, the homophily terms in this model compensate for the overestimation by becoming negative. The net probability of pairing on the diagonal is still strongly positive. For example, using the model with uniform homophily and distance, diagonal pairings are $e^{-1.48+3.20} = 5.6$ times as likely as pairings one step off the diagonal. But pairings one step off the diagonal are $e^{3.20} = 24$ times as likely as pairings two steps off the diagonal.

The most parsimonious model with an acceptable level of fit is the last one in the summary table: partial differential homophily, distance, and asymmetry. In this model, the within-group pairing bias as for age groups 1, 3, and 4 are constrained to be equal, and group 2 is allowed to be different. The factor sets groups 1, 3, and 4 at level 1, so the degree of within-group selection bias for these groups is completely determined by the distance parameter. For these three groups, pairings on the diagonal are $e^{-2.55} = 12.8$ times more likely than pairings one step off. For age group 2, by contrast, diagonal pairings are $e^{2.55-0.24} = 10.1$ times as likely as pairings one step off. For all groups, pairing $n+1$ steps off the diagonal are 12.8 times as likely as pairings $n+2$ steps off. In addition, older men are $e^{1.47} = 4.3$ times as likely to choose a younger woman as older women are to choose a younger man. In this case, the structure of selective mixing in the 16-cell table is adequately described with only three generalized interaction terms.

Assessing the significance of the fit statistics is compromised in this case by the extent of data dredging. The purpose of this example is more to demonstrate the way in which the parameters can be used and interpreted from a substantative, rather than a statistical, perspective. Given the virtual absence of a research tradition on the structure of sexual networks, data analysis is going to take an exploratory form for some time to come. The issue of statistical significance should be approached with this in mind.

Between the nominal and ordinal log-linear models available, every pattern of selective mixing can be described, and a wide range can be summarized. When the patterns of mixing are simple, the corresponding model will be parsimonious. When these patterns are complex and unsystematic, the corresponding model will be similarly complicated. The conditional likelihood ratio statistic can be used to monitor the accuracy–parsimony trade-off, and the size and standard error of the parameter estimates can be used to identify the major contributors to differential association.

INTEGRATING SELECTIVE MIXING INTO THE COMPARTMENTAL MODEL

To integrate selective mixing into a compartmental modeling framework, the infection rate term must be modified to reflect the subgroups in the

population and the mixing structure among them. This involves two things: establishing the connection between the mixing matrix and the infection rate term, and specifying the rules that will govern mixing patterns as the population structure changes.

In relatively standard notation, the infection rate term for a compartmental model is given as βSI, where β is the effective contact rate, a product of the number of contacts per individual and the probability of transmission per contact scaled by the number of persons in the active population, and S and I refer to the number of susceptibles and infecteds, respectively. A model for selective mixing replaces βSI with the term $\sum_{ij} \beta_{ij} S_i I_j$, where the subscripts i and j represent the subgroups in the population (e.g., gender, age categories) and β_{ij} is the effective contact rate between them. The link between the effective contact rate and the contact matrix can be made clear by expanding β_{ij} into its components:

$$\beta_{ij} = c_i \pi_{ij} \tau_{ij} / T_j, \tag{19}$$

where c_i is the average number of partners per unit time for members of group i; π_{ij}, the conditional probability of a partner from group j given a subject from group i; τ_{ij}, the per-partner probability of disease transmission for ij dyads; and T_j, the total number in group j.

Clearly, more terms could be partialled out in this expression, to reflect, for example, the role of particular forms of contact (e.g., anal vs. vaginal intercourse). The approach taken here shares the goal of partitioning the effective contact rate into substantively meaningful components. In order to focus more clearly on the separation of social from behavioral parameters, other factors are ignored, but this does not imply that they are irrelevant.

Because the cell counts of the contact matrix can be expressed as function of the row total and the conditional row probabilities,

$$x_{ij} = T_i c_i \pi_{ij}, \tag{20}$$

the effective contact rate β_{ij} can now be expressed in terms of the contact matrix cell counts,

$$\beta_{ij} = x_{ij} \tau_{ij} / T_i T_j. \tag{21}$$

The contact matrix cell counts, in turn, can be expressed in terms of an appropriate log-linear model, thus linking the parameters of the mixing structure to the infection rate.

However, the mixing model must be made responsive to changes in population sizes. This is a variant of the familiar "two-sex problem" in

demography. The question can be stated very simply: What do people do when their preferred partners become more (or less) available? For many reasons, including demographic cycles, changes in behavior, and differential mortality rates from the disease, the relative sizes of subgroups in the population are not likely to be stable over time. The model must therefore specify a set of rules that governs how people alter their mixing patterns in response to changes in their opportunity structure.

This is an interesting substantive question, and despite the absence of empirical data it is possible to define the range of possible answers. As a preferred partner group becomes less available, an individual may substitute with partners from other, less preferred groups or reduce the number of partners she or he has, a choice that is in some part constrained by the behavior of other groups. Whatever solution is adopted, it must obey something we might call a contact consistency constraint: The number of partnerships between i's and j's must equal the number of partnerships between j's and i's. This blatant tautology actually contains several insights.

The contact consistency constraint implies symmetry in the full contact matrix: x_{ij} must equal x_{ji}.[3] Given Equation (2), this means

$$T_i(t) c_i \pi_{ij} = T_j(t) c_j \pi_{ji}. \tag{22}$$

As noted earlier, the groups' sizes $[T_{(\cdot)}(t)]$ may change independently over time. In order to ensure that this equation remains satisfied, corresponding changes must be made to either or both of the remaining parameters, $c_{(\cdot)}$ and $\pi_{(\cdot)}$. In substantive terms, as the relative availability of groups changes over time, subjects must adjust either their contact rates or their selection patterns, or some combination of both.

As the parametrization of the constraint suggests, the possible behavioral responses can be arrayed along a continuum. At one end, the contact rates can be fixed, with all compensation operating through selection probabilities. At the other end, the selection probabilities can be fixed, with all compensation operating through contact rates. In between, some mix of changes in both contact rates and selection probabilities can be used to offset

[3] Note that the matrices in Figures 1 and 2 are partial matrices representing one of the heterosexual quadrants of the full matrix. The symmetry constraint does not apply within these partial matrices. In addition, empirical contact matrices will probably display some departure from symmetry owing to a combination of sampling variability and reporting error. In theory, however, the constraint must hold, and empirical matrices will have to be adjusted accordingly. This is an important modeling issue, but it does not interfere with the general framework developed here.

the changes in group size. It is worth examining each of these approaches in a bit more detail in order to clarify their underlying behavioral assumptions.

The first extreme can be seen as a "pure drive" model of sexual behavior. Here the assumption is made that contact rates are fixed as if each individual had an inherent drive or quota to achieve. As one group becomes less available, individuals simply substitute partners from more available groups to fill their quota of sexual contacts. It is possible to specify a limited number of selection rules for filling the quota (equal to the number of degrees of freedom of a square, symmetric fixed-margin table). These selection rules can be specified, for example, as a set of $(n-1)^2$ odds ratios, but not as a full set of n^2 conditional probabilities, $\{\pi_{ij}\}$. Under the pure drive model, the table margins and selection probabilities ($\pi_{(\cdot)}$'s) change over time as a function of group size; however, the contact rates ($c_{(\cdot)}$) and (if desired) odds ratios remain fixed.

The second extreme can be seen as a "pure selection" model of sexual behavior and might reflect some notion of sexual identity. Here the assumption is that people choose sexual partners in fixed ratios; as a group's size changes, its contact rates are adjusted inversely to maintain its relative share of partnerships with others. Under the pure selection model the table margins and selection probabilities are constant over time.

In between these two extremes lie what might be called "modified selection" models of sexual behavior. In these models, individuals accommodate changes in relative group size by adjusting both contact rates and selection probabilities. The way in which this mix is obtained is open to modeling, and in contrast to the fairly simplistic dynamics of the pure drive and pure selection models, the modified selection approach permits more reasonable sets of assumptions to be made about the nature of behavioral responses to demographic change. This flexibility entails a loss of direct control over the basic behavioral components, $c_{(\cdot)}$ and $\pi_{(\cdot)}$.

One class of modified selection models takes the following form. Decompose the contact matrix cell counts into a function of population structure, which changes over time, and mixing preferences, which may be treated as either stable or changing [26].

$$x_{ij}(t) = \frac{T_i(t)T_j(t)}{T(t)}\alpha_{ij} = K_{ij}(t)\alpha_{ij}, \tag{23}$$

where, for example, the mixing preference term, α_{ij}, could reflect a mutual signaling dynamic,

$$\alpha_{ij} = cs_{ij}s_{ji}.$$

Here, c can be taken as the number of contact opportunities, and the s_{ij} terms can be interpreted as the probability that a member of group i signals

yes to a member of group j, a function of activity level preference and selection preferences for members of group i. The signal parameters s_{ij} and s_{ji} are not constrained to be equal. Behavior, reflected in the cell counts defined using Equation (23), can now vary over time in response to relative group sizes while continuing to satisfy the contact consistency constraint. The components of this model can be estimated from data.[4] Note that elements of both personal selection preference and structural opportunity are included in this model. The framework does not constrain the analyst to focus exclusively on either individual or social forces; it permits both.

The α_{ij} term, which regulates mixing and activity level preferences, can be treated as either constant or time-dependent. The assumption could be verified empirically using longitudinal contact matrix data, if such data were available. In the absence of empirical guidance, it would seem reasonable to assume, as a first approximation, that such preferences are constant.[5] Using this assumption, one can then estimate the α_{ij} term from a contact matrix at any time point using log-linear models as follows. Denote the expected value of the contact matrix cell counts by m_{ij}.

$$E\left[x_{ij}(0)\right] = \exp\left\{u + u_{1(i)} + u_{2(j)} + \sum_{1}^{k} u_{12(k)}\right\} = m_{ij}(0). \tag{24}$$

These $m_{ij}(0)$ can be estimated from an empirical contact matrix, and by Equation (23), with the number of persons in each group (T_i), assumed known, one can use the resulting estimates to solve for the preferences, α_{ij}:

$$\hat{m}_{ij}(0) = K_{ij}(0)\hat{\alpha}_{ij}, \qquad \frac{\hat{m}_{ij}(0)}{K_{ij}(0)} = \hat{\alpha}_{ij}. \tag{25}$$

Then the cell counts can be updated as the group sizes change by substitution

$$\hat{x}_{ij}(t) = K_{ij}(t)\hat{\alpha}_{ij}. \tag{26}$$

[4] This modified selection model is overparametrized, so the c and s_{ij} terms cannot be separately identified and estimated. However, the product α_{ij} can be estimated.

[5] Some point to market research for evidence that preferences change continually. I suspect, however, that sexual preferences are likely to be less volatile than brand loyalties.

(It is also possible to model the α_{ij} terms directly by treating K_{ij} as an offset in the estimation process [22, p. 138]. This will not change the fitted values $[m_{ij}(0)]$, but it will change the value and interpretation of the u-parameter estimates. When modeled via Equations (24)–(26), the estimates represent the effects of both population structure and preferences. When modeled with K_{ij} as an offset, they represent only the effects of preferences.)

Finally, (26) can be substituted back into the infection rate term in Equation (21), completing the updating algorithm and establishing the final link in the integrated modeling framework:

$$\beta_{ij}(t) = \frac{\hat{x}_{ij}(t)\tau_{ij}}{T_i(t)T_j(t)} = \frac{\hat{m}_{ij}(0)\tau_{ij}}{K_{ij}(0)T(t)}. \tag{27}$$

In short, log-linear model parameter estimates can be used to construct fitted values, $\hat{m}_{ij}$, that can be substituted into the infection rate equation for $x_{ij}(t)$. These fitted values, with their underlying mixing structure, will then be used to update the contact patterns and thus the infection rate $\beta_{ij}(t)$ as the population profile changes over time.

In an earlier paper [26], this modified selection framework was used to examine the impact of different mixing structures on the transmission dynamics of a disease like AIDS. The impact was shown to be potentially quite dramatic. When the attributes that define partner selection rules, such as gender, race, or sexual preference, are relatively stable over an individual's lifetime, stable mixing groups are formed. The potential for spread between two groups that are not directly connected by sexual interaction, for example, homosexual males and heterosexual females, then depends on the existence and size of a bridge population, in this case bisexuals. The simulations in this paper demonstrated that a bridge is necessary but not sufficient for the spread of the epidemic between unconnected groups. With a small bridge and stable attributes that permanently define mixing group membership. the infection may remain isolated, even under fairly extreme conditions. Contact rates of over 80 partners per year in the seeded group and 5–10 per year in the bridge population were not enough to sustain transmission to other population groups in this example.

When the attributes that define partner selection rules, such as age, change over an individual's lifetime, more fluid mixing groups are formed because people can change their group membership, and the potential for spread is much higher. Here the mixing structure is characterized by a dual transmission regime, with the group-specific infection rate a function of both infections passed by sexual contact and the rate at which people change

group membership. That is, some infections will be passed between groups as people become members of a new group, for example, by aging or by changing marital status. This kind of mixing structure makes the epidemic much more likely to spread, even with contact rates of two partners per year or less.

These examples demonstrated the exceptionally strong effects mixing structures can have on the spread of AIDS. Clearly, if one could reduce the probability of transmission given contact to zero, the transmission would stop, but mixing structures have equally dramatic effects. By exploring the effects of mixing structures on transmission dynamics in a systematic fashion, it may be possible to anticipate and partially control the spread of this disease.

The framework presented here provides a simple way to model a wide variety of mixing patterns and incorporate a mixing structure into dynamic models of transmission. Log-linear models offer a mechanism for systematically exploring sociologically informed hypotheses regarding the nature of selective mixing. They make it possible to parsimoniously describe a mixing structure and generate parameters that have clear and relatively intuitive interpretations, and they provide a sound statistical basis for estimating these parameters from data. The modified selection model and the updating algorithm make it possible to channel the transmission through this mixing structure even as the population profile changes. It is, in short, an integrated and general framework for exploring the role of social structure in the dynamics of social transmission.

I thank Dr. James A. Wiley for providing the data in Figure 2. These data were collected as a part of the AIDS in Multiethnic Neighborhood Survey (AMEN), which is supported by National Institute of Mental Health Center grant No. MH42459 to the Center for AIDS Prevention Studies, University of California, San Francisco.

APPENDIX: AN EXAMPLE BEHAVIORAL MODEL FROM THE MODIFIED SELECTION CLASS

The statistical framework on which log-linear models are based relies on specific distributional assumptions. The distribution of sexual partnerships in the mixing matrix must be either Poisson, full multinomial, or product multinomial. Whether the distribution satisfies these assumptions or not depends on the underlying behavioral model that generates the observations. The derivation below uses the modified selection model described by Equation (23). The derivation proceeds in two steps; first deriving the distribution of contacts in the single-group, random mixing case, then generalizing to the multigroup selective mixing case.

Let the data consist of a random sample of T subjects. Each subject k has an observed number of contacts X_k, and we wish to estimate the expected value of X_k. If we make the simplifying assumptions that each subject is independent of other subjects in the sample (i.e., does not name another member of the sample as a partner) and that the number of contact opportunities, c, is identically distributed across subjects such that

$$c \sim \text{Poisson}(\lambda),$$

then the derivation of the distribution of X_k is straightforward. For every opportunity, a contact is either made or not made, so that, conditional on the number of opportunities, the number of successful contacts is binomially distributed. Letting the number of successful contacts for subject k be denoted by X_k, then

$$X_k = x \mid C = c \sim \text{binomial}(c, \phi),$$

and the unconditional probability is given by

$$\begin{aligned} p(X_k = x) &= \sum_{c=x}^{\infty} p(X_k = x \mid C = c)\, p(C = c) \\ &= \sum_{c=x}^{\infty} \binom{c}{x} \phi^x (1-\phi)^{c-x} \frac{e^{-\lambda} \lambda^c}{c!}, \end{aligned}$$

which can be shown to be distributed Poisson with parameter $(\lambda\phi)$ [8, p. 287]. Therefore,

$$X_k \sim \text{Poisson}(\lambda\phi).$$

The total number of contacts is simply the sum over all observations of X_k. Under the assumption that all subjects in the sample are independent, their contacts are identically distributed, and conditional on a fixed sample total T, this sum is also distributed Poisson,

$$X\left(= \sum_{k=1}^{T} X_k \right) \sim \text{Poisson}(\lambda\phi T),$$

so that

$$E(X) = \lambda\phi T$$

and

$$\text{Var}(X) = E(X).$$

Here λ represents the number of contact opportunities per subject, ϕ the probability that the opportunity results in a sexual partnership, and T the number of subjects.

The assumption that the subjects are independent is probably legitimate in a sample like that used for the General Social Survey, a small (2000), nationally representative sample of households. It is reasonable to assume that in such a sample the probability of a subject mentioning a sexual partner who is also in the sample is very small. This assumption would not be legitimate, however, for most of the community-based volunteer and convenience samples often used in AIDS-related research, such as, the MACS studies and the San Francisco City Clinic Cohort. The treatment of nonindependent sample units is an interesting issue but will not be addressed here.

The assumption that subjects have identically distributed numbers of contacts is a strong one but can easily be relaxed. In the context of sexual behavior, for example, it is very likely that there are substantial differences among individuals in the expected number of sexual contacts they will make, for example, variation by age, gender, or marital status. To the extent that these differences reflect measured characteristics of the subjects, they present no problem in terms of estimation and inference here; the variation in contact frequencies can be partitioned into a systematic component attributable to differences in characteristics and a residual component that can be referred to the appropriate reference distribution.

Let the data again consist of a random sample of T subjects. Each subject k now has a vector of attributes (e.g., race, age, gender) and a set of X_k partner vectors with the corresponding attributes of each partner. If there are N categories formed by the full cross-tabulation of the attributes, let i, $i = 1, \ldots, N$, index the category of the subject and j, $j = 1, \ldots, N$, index the categories of the subject's partners. For any subject k in group i, the partner vectors can be summed to give X_{kij}, the frequency of a subject's contacts with members of group j. For any subject group i, the X_{kij} can be summed over all its members to give X_{ij}, the frequency of contacts between subjects in group i and partners in group j. We wish to estimate the X_{ij}, which are the cell entries of the contact matrix. The distribution of these frequencies can be derived, as above, as a function of the opportunities for contact and the probability that an opportunity results in a successful contact.

As before, if we make some simplifying assumptions, the derivation of the distribution of X_{ij} is fairly straightforward. In particular, three assumptions are useful here:

(1) The opportunity for contact is distributed identically across the categories of i. This does not mean that all groups will have the same average number of contacts, only that there are no group-specific advantages or obstacles in the opportunity structure.

(2) The probability of encountering a member of group j is identical for all categories of i. Again, this does not mean that all groups will have the same propensity to select someone from group j, just that all groups are equally likely to encounter them.

(3) The contacts for each individual are independent and identically distributed; that is, each separate contact a subject makes is governed by the same probability process, like repeated throws of a die, and is unaffected by the outcome of prior contacts.

Under these assumptions, for any subject k in group i the number of contact opportunities with group j will be given by c_{kj}, where, conditional on the relative proportion of j's, T_j/T, and on assumptions 1 and 2 above,

$$c_{kj} \sim \text{Poisson}\left(\frac{T_j}{T}\lambda\right).$$

Conditional on the number of opportunities and assumption 3 above, the number of successful contacts, X_{kij}, is binomial,

$$X_{kij} = x \mid c_{kj} \sim \text{binomial}\left(c_{kj}, \phi_{ij}\right),$$

and using the same derivation as for random mixing above, the unconditional distribution is given by

$$p\left(X_{kij} = x\right) \sim \text{Poisson}\left(\frac{T_j}{T}\lambda\phi_{ij}\right).$$

Let $G_i = \{k\colon k \in \text{group } i\}$, and let T_i be the total number of observations in G_i; then the total number of contacts between members of group i and members of group j is the sum over G_i of the X_{kij}. Conditional on the number of observations in each group i,

$$X_{ij}\left(=\sum_{k\in G_i} X_{kij}\right) \sim \text{Poisson}\left(\frac{T_iT_j}{T}\lambda\phi_{ij}\right).$$

Apart from clustering and overdispersion considerations,[6] the modified selection model for the X_{ij} (the cell counts in the contact matrix) generates observations from a Poisson distribution with a multiplicative model for the mean. This establishes the formal requirements for the use of log-linear models.

REFERENCES

1 A. Agresti, *The Analysis of Ordinal Cross-Classified Data*, Wiley, New York, 1984.

2 R. D. Alba and R. C. Kessler, Patterns of interethnic marriage among American catholics, *Soc. Forces* 57:1124–1140 (1979).

3 R. M. Anderson, S. Gupta, and W. Ng, The significance of sexual partner contact networks for the transmission dynamics of HIV. *J. AIDS* 3:417–429 (1990).

4 A. Barbour, Macdonald's model and the transmission of bilharzia, *Trans. Roy. Soc. Trop. Med. Hyg.* 72:6–15 (1978).

5 Y. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete Multivariate Analysis: Theory and Practice*, MIT Press, Cambridge, Mass., 1975.

6 R. S. Burt, Measuring age as a structural concept, *Social Networks* (in press) (1990).

7 S. Busenberg and C. Castillo-Chavez, Interaction, pair formation and force of infection terms in sexually transmitted diseases, in *Mathematical and Statistical Approaches to AIDS Epidemiology*, pp. 289–300 in C. Castillo-Chavez, Ed., Springer-Verlag, Berlin, 1989.

8 R. Centers, Marital selection and occupational strata, *Am. J. Sociol.* 54:530–535 (1949).

9 W. Feller, *An Introduction to Probability Theory and Its Applications*, Vol. 1, Wiley, New York, 1950.

[6] The issue of heterogeneity or overdispersion is present in this context. There is, in fact, good a priori reason to expect that the type of data that will be encountered in this field will display some heterogeneity. There are intensive data requirements for a fully cross-tabulated subject index. Even with five simple characteristics like gender, race, sexual preference, marital status, and age, the cross-tabulated index would conservatively generate $2\times3\times3\times2\times3 = 108$ categories, resulting in a mixing matrix of over 10,000 cells. Given budgetary constraints on sample size, it is likely that the index categories would have to be collapsed to ensure adequate numbers in cells. Collapsing, of course, introduces some amount of heterogeneity into the resulting categories.

On the other hand, the clustered sampling of contacts could introduce some correlation among contacts within persons. This would result, for example, if respondents filled rigid quotas of partner types; the type of partner for the later contacts then would not be independent of the type of partner for the earlier contacts. Given the low average number of partners (per year) for most of the population, this problem may not be a serious one. Yamaguchi [34] presents methods that can be used to assess these assumptions.

10 H. H. Garrison, Education and friendship choice in urban Zambia, *Social Forces* 57:1310–1324 (1979).

11 L. A. Goodman, On the statistical analysis of mobility tables, *Am. J. Sociol.* 70:564–585 (1965).

12 L. A. Goodman, *Analyzing Qualitative/Categorical Data*, Abt Books, Cambridge, Mass., 1978.

13 L. A. Goodman, Multiplicative models for the analysis of occupational mobility tables and other kinds of cross-classification tables, *Am. J. Sociol.* 84:804–819 (1979).

14 L. A. Goodman, *The Analysis of Cross-Classified Data Having Ordered Categories*, Harvard Univ., Cambridge, Mass., 1984.

15 R. M. Hauser, Some exploratory methods for modeling mobility tables and other cross-classified data, pp. 413–458 in *Sociological Methodology*, K. F. Schuessler, Ed., Jossey-Bass, San Francisco, 1979.

16 T. C. Hunt, Occupational status and marriage selection, *Am. Sociol. Rev.* 5:495–504 (1940).

17 J. A. Jacquez, C. P. Simon, and J. Koopman, Structured mixing: heterogeneous mixing by the definition of activity groups, in *Mathematical and Statistical Approaches to AIDS Epidemiology*, pp. 301–315 in C. Castillo-Chavez, Ed., Springer-Verlag, Berlin, 1989.

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

19 J. Koopman, C. P. Simon, J. A. Jacquez, and T. S. Pans, Selective contact within structured mixing with an application to HIV transmission risk from oral and anal sex, in *Mathematical and Statistical Approaches to AIDS Epidemiology*, pp 316–348 in C. Castillo-Chavez, Ed., Springer-Verlag, Berlin, 1989.

20 W. S. Landecker, Class boundaries, *Am. Sociol. Rev.* 25:868–877 (1960).

21 E. O. Laumann, *Prestige and Association in an Urban Community*, Bobbs-Merrill, Indianapolis, 1966.

22 P. McCullagh and J. Nelder, *Generalized Linear Models*, Chapman & Hall, London, 1983.

23 P. V. Marsden, Models and methods for characterizing the structural parameters of groups, *Social Networks* 3:1–27 (1981).

24 P. V. Marsden, Homogeneity in confiding relations, *Social Network* 10:57–76 (1987).

25 M. Morris, Networks and diffusion: An application of log-linear models to the population dynamics of disease, Ph.D. thesis, Univ. Chicago, 1989.

26 M. Morris, Networks and diffusion: modeling the effects of selective mixing on the spread of disease, *Am. J. Soc.* (under review) (1990).

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

28 L. Sattenspiel, Modeling the spread of infectious disease in human populations, *Yearbook of Physical Anthropology* 33:245–276 (1990).

29 L. Sattenspiel and C. P. Simon, The spread and persistence of infectious diseases in structured populations, *Math. Biosci.* 90:341–366 (1988).

30 M. E. Sobel, Structural mobility, circulation mobility, and the analysis of occupational mobility, *Am. Rev. Sociol.* 48:721–727 (1983).

31 M. E. Sobel, M. Hout, and O. D. Duncan, Exchange, structure and symmetry in occupational mobility, *Am. J. Sociol.* 91:359–372 (1985).

32 U.S. Bureau of the Census, Current Population Reports, Series P-20, No. 424, *Household and Family Characteristics: March 1987*, U.S. Govt. Printing Office, Washington, D.C., 1987.

33 L. M. Verbrugge, The structure of adult friendship choices, *Social Forces* 56:1286–1309 (1977).

34 K. Yamaguchi, Homophily and social distance in the choice of multiple friends, *J. Am. Stat. Assoc.* 85:356–366 (1990).