

## Comments

---

Course: Soc551 / HPA 594

Laumann, Edward. O. and Yoosik Youm. 1999. "Race/ethnic group differences in the prevalence of sexually transmitted diseases in the United States: A Network Explanation." *Sexually Transmitted Diseases*. 26(5): 250-61.

# Racial/Ethnic Group Differences in the Prevalence of Sexually Transmitted Diseases in the United States: A Network Explanation

EDWARD O. LAUMANN, PH.D, AND YOOSIK YOUM, MA

**Background:** Many studies have observed that African Americans have comparatively high rates of selected STDs, often 10 to 20 times higher than whites and other racial/ethnic groups, but without convincing explanation.

**Goal:** This study attempts to solve this puzzle using data from a nationally representative probability sample and a network approach.

**Study design:** The National Health and Social Life Survey (NHSLs) is a nationally representative probability sample of 1,511 men and 1,921 women in the United States. Logistic regression analysis of these data permit a multivariate analysis of the individual risk factors associated with STDs. Using loglinear analysis and a simulation, we also identify the effects of sexual network patterns within and between racial/ethnic groups.

**Results:** Logistic regression analysis of the NHSLs revealed, even after controlling for all the appropriate individual-level risk factors, that African Americans are almost five times more likely to be infected by bacterial diseases than the other racial/ethnic groups.

**Conclusions:** African Americans' higher infection rate for bacterial diseases can be explained by the patterns of sexual networks within and between different racial/ethnic groups. First, infections are more widespread in the African American population at large because partner choice is more highly assortative—meaning that “peripheral” African Americans (who have had only one partner in the past year) are five times more likely to choose “core” African Americans (who have had four or more partners in the past year) than “peripheral” whites are to choose “core” whites. Secondly, sexually trans-

*From The Department of Sociology, University of Chicago,  
Chicago, Illinois*

mitted infections stay within the African American population because their partner choices are more segregated (assortative mating) than other groups. The likelihood of African Americans having a sexually transmitted infection is 1.3 times greater than it is for whites because of this factor alone.

AFRICAN AMERICANS HAVE substantially higher rates of sexually transmitted diseases (STDs) than other ethnic and racial groups in the United States.<sup>1-5</sup> Hispanics also manifest higher rates of primary and secondary syphilis and HIV, although not as disproportionately as African Americans.<sup>6</sup> Both groups are overrepresented among new human immunodeficiency virus cases.<sup>7</sup> There is some debate about why this is the case. In a report published in 1996, the Centers for Disease Control and Prevention (CDC) observed: “There are no known biological reasons to explain why racial or ethnic factors alone should alter the risk for STDs. Rather, race and ethnicity in the United States are markers that correlate with other more fundamental determinants of health status such as poverty, access to quality health care, health care seeking behavior, illicit drug use, and living in communities with high prevalence of STDs.”<sup>2</sup>

Three limitations of the most commonly used methods for collecting and analyzing data on STDs have made it difficult to account for these discrepancies. The first two limitations are inherent in the way the data have been collected and the last derives from the analytic approaches that are conventionally used. First, the most commonly used data collecting methods are usually case-based instead of person-based, including the “contact tracing method,”<sup>8</sup> “residential area studies,”<sup>9,10</sup> and “Center for Disease Control reports,”<sup>11</sup> which are based primarily on passive surveillance reports from diverse institutions. These methods have serious weaknesses, especially with respect to their systematic bias against inclusion of cases from people of higher socio-

A version of this paper was originally presented at a plenary session of the Conference on the Social Determinants of Sexually Transmitted Diseases held at the University of Alabama, Birmingham, Alabama, on May 26, 1996, and subsequently revised.

The authors thank James Koopman, MD, Alberto Palloni, John Potterat, Kazuo Yamaguchi, and Dingxin Zhao for their helpful comments and suggestions.

The authors acknowledge the financial support of the Ford Foundation (Grant No. 940-1417-2) and the National Institute for Child Health and Human Development (5 RO1 HD28356-03) that made this work possible.

Correspondence: Edward O. Laumann, Department of Sociology, University of Chicago, Chicago, IL, 60637.

Received for publication May 22, 1998, revised September 21, 1998, and accepted September 22, 1998.

economic status. To insure privacy, people of higher socioeconomic status often turn to private doctors who are likely to place a lower priority on reporting STDs to government agencies.<sup>11</sup> Therefore, their prevalence rates are underestimated.<sup>12,13</sup> Studies have found that private physicians may report only 3% to 60% of gonorrhea cases they treat and that the majority of cases reported in the national surveillance system are from public providers.<sup>14</sup> For a recent example, according to a 1995 cross-section survey of residents in Chicago and Cook County, more than 62% of the self-reported cases of STDs in that 12-month period were diagnosed and treated by private doctors, whereas only 5.2% were diagnosed and treated in publicly funded STD clinics; but 60% to 80% of the STD cases reported to the CDC from Chicago in that year were reported by the publicly funded STD clinics.<sup>15</sup>

Second, most of the data about STDs include information about only infected people. Using data lacking relevant information about the noninfected makes it impossible to examine several critical questions. For example, do the infected have more sexual partners than the noninfected? Do African Americans have higher infection rates, even after controlling for the differential numbers of partners across ethnic or racial groups? In principle, researchers have lacked appropriate tools to examine in detail the causes for the higher infection rates among African Americans. They can only speculate whether it is related to urban settings, poverty level, etc.<sup>1,2</sup>

The third limitation arises from deficiencies in the analytic approach conventionally used. Most studies of the incidence or prevalence of STDs focus on attributes of individuals, such as the number of partners, use of prostitutes, or frequency of condom use. Such an approach fails to recognize two fundamental network aspects of STD dynamics. The first concerns the potential infection status of the partners (or second-order network connectivity). For example, Person A may have a higher likelihood of being infected than Person B despite their having the same number of sex partners because Person A's sex partners have had more sex partners than Person B's sex partners. In determining who is more vulnerable to infection, then, we have to consider not only an individual's risk factors but also the risk factors contributed by the other partners constituting the sexual network. In other words, partner selection is a critical part of the risk equation. The second network feature critical to STD dynamics concerns the existence of bridges between socially distinct subpopulations. Even though Person A may have a higher probability of being infected than Person B, Person B can be a much more efficient (powerful) transmitter of infection if he or she has sex partners in socially distinct subpopulations, thereby providing a specific link or bridge for infection to spread between the two subpopulations. In this case, without Person B being infected, it is

impossible for the one group to transmit infection to the other.

This study will try to address these limitations by using a national probability-based sample that is person-based rather than case-based and includes infected and noninfected people in an unbiased selection process and by adopting a network analytic strategy.

## Methods

### Sample

The National Health and Social Life Survey (NHSL), conducted in 1992, is a nationally representative probability sample of 1,511 men and 1,921 women between the ages of 18 and 59 years living throughout the United States. It represents approximately 97% of the population in that age group, roughly 150 million Americans. The sample completion rate was greater than 79%. Checks with high-quality samples (such as the Census Bureau's Current Population Survey) suggest that the NHSL succeeded in getting a truly representative sample of the population. Each person was surveyed in person by experienced interviewers, who matched respondents on various social attributes, for an interview averaging 90 minutes in duration.<sup>11</sup>

## Measures

### Logistic Regression

After excluding the oversamples of African Americans and Hispanics, we constructed a set of independent and dependent variables as described in Table 1.

People who had no sexual partners during their lifetime were omitted because many risk factors for them are undefined. Three dependent variables are constructed from the question, "There are several diseases or infections that can be transmitted during sex. These are sometimes called venereal diseases or VD. We will be using the term sexually transmitted diseases or STDs to refer to them." After giving the respondent a hand card listing each STD, sometimes including vernacular terms (for example, "clap" or "drip" for gonorrhea), interviewers asked, "As I read each STD, tell me whether you have ever been told by a doctor that you had it."<sup>11</sup> Variables for each STD are coded as 1 for "ever infected" and 0 for "never infected."

### For Network Analysis

To examine the network effects, we constructed a contact matrix (who partners with whom) to reveal the network patterns between different population subgroups. First, we omitted the people who only had same-sex intercourse during the past 12 months from the data set that was used in the logistic regression analysis. We excluded them not be-

TABLE 1. Variable Measures

		Dependent Variables		
Variables	Included STDs	Mean	SD	N
Any STDs	Any STDs, including bacterial and viral	0.18	0.39	2878
Bacterial STDs	Gonorrhea, syphilis, chlamydia, nongonococcal urethritis for men, pelvic inflammatory disease for women	0.11	0.32	2896
Viral STDs	Genital warts, herpes, hepatitis, HIV (AIDS)	0.08	0.27	2929
		Independent Variables		
Variables	Description	Mean	SD	N
Number of partners	Continuous variable	10.75	40.50	2933
Female	Female = 1, Male = 0	0.56	0.50	2963
Big city*	Big city = 1, Others = 0	0.22	0.41	2963
Pay for sex	Ever = 1, Never = 0	0.09	0.29	2745
Being paid for sex	Ever = 1, Never = 0	0.03	0.18	2732
Anal sex	Ever = 1, Never = 0	0.25	0.44	2766
Never married	Never = 1, Ever = 0	0.26	0.44	2926
Military service	Ever = 1, Never = 0	0.14	0.34	2962
Drug injection†	Ever = 1, Never = 0	0.02	0.15	2742
Group sex	Ever = 1, Never = 0	0.09	0.29	2787
Concurrent partners‡	Ever = 1, Never = 0	0.29	0.45	2963
Prepubertal intercourse§	Ever = 1, Never = 0	0.03	0.16	2958
Age	Continuous variable	36.79	10.86	2963
Race/ethnic groups¶	White = 1, African Americans = 2, Hispanic = 3			2963
Education	Five categories			2945
Religious attendance*	Almost weekly = 1, Sometimes = 2, Rarely = 3			2956
Thinking of sex	Three categories**			2952
Sexual value orientation	Four categories††			2824

All variables are for lifetime duration except 'big city,' 'education,' 'religious attendance,' 'thinking of sex,' and 'sexual value orientation.' They refer to current status.

\*'Big city' is defined as either the 'central city of 12 largest standard metropolitan statistical areas (SMSAs)' or the 'central city of 100 largest SMSAs'.

†The question is 'have you ever injected drugs, that is taken drugs using a needle, that weren't prescribed by a doctor?'

‡'Concurrent partners' means having several sex partners overlap in a given time period.

§We defined 'pubertal age' as the age at which the first menstrual period occurred for females and the age at which the voice changed or pubic hair appeared for males. If they did not remember this age, we used '12 or 13 years of age.'

¶White (79%), African Americans (13%), and Hispanics (8%).

||Less than high school (14%), high school (29%), vocational school or 2-year degree (33%), finished college (17%), and Master's or advanced degree (7%).

\*\*Frequency of attending religious services. Almost weekly = weekly or nearly every week, Sometimes = 'several times a year' to '2 or 3 times a month,' Rarely = 'never' or at most 'about once or twice a year.' Almost weekly (29%), Sometimes (32%), and Rarely (40%).

††How often do you think about sex? Never or less than once a month (8%), Few times a month/year (56%), and Every day or several times a day (36%).

‡‡Most conservative (6%), Conservative (50%), Liberal (42%), and Most liberal (2%). These are the rounded average to nine questions about sexual values. These include premarital sex, teen sex, extramarital sex, same-gender sex, laws against pornography, sex without love, abortion, consensual sex (any kind of sexual activity between adults is okay as long as both persons freely agree to it), and religion-guided sexual behavior (my religious beliefs have shaped and guided my sexual behavior). See reference 11, chapter 14, for details.

STD = sexually transmitted disease.

cause they are unimportant in STD transmission dynamics but because our mathematical model requires the assumption that only people engaging in opposite-sex intercourse are included.

Second, we divided the entire sample into three groups according to their level of sexual activity: (1) "peripherals" are those who have had only one sexual partner in the past 12 months and who are therefore believed to be safe from infection; (2) "adjacents" are those who have had two or three sexual partners in the same period; and (3) "core group members" are those who have had at least four sexual

partners in the same 12-month period and are therefore considered to be primarily responsible for the existence of sexually transmitted diseases in the population over time. Definitions in mathematical epidemiology have distinguished only between core and noncore members of a population on the basis of their epidemiologic function in maintaining STDs in that population.<sup>16,17</sup> Prior empirical definitions<sup>8,9</sup> have distinguished among core, adjacent, and peripheral on the basis of geographic areas treated as proxy indicators for the underlying constructs. Our distinctions are operationalizations of these theoretical constructs and em-

TABLE 2. Contact Matrix (Number of Partnerships for the Last 12 Months)\*

	WP	WA	WC	AP	AA	AC	HP	HA	HC
WP	1463.02								
WA	78.44	199.99							
WC	37.39	160.65	175.98						
AP	12.25	1.53	0.86	172.16					
AA	0.48	3.01	2.16	18.93	67.02				
AC	1.19	5.61	3.91	16.64	59.88	44.93			
HP	33.75	1.91	2.07	2.24	0.39	0.93	82.32		
HA	3.96	4.75	8.25	0.14	1.00	3.68	4.86	9.67	
HC	0.29	4.73	7.79	0.23	2.59	4.32	3.14	13.80	10.41

\*This table shows the estimated number of partnerships between persons in row (chooser) and column (chosen) groups, calculated according to the procedures described in Appendix A. Because the estimated number of contacts for an ordered row/column combination is the same as for its reversed order, the matrix is symmetric, and the upper right-hand entries can thus be omitted to avoid redundancy. WP = white periphery; WA = white adjacent; WC = white core; AP = African American periphery; AA = African American adjacent; AC = African American core; HP = Hispanic periphery; HA = Hispanic adjacent; HC = Hispanic core.

pirical proxies based on individual-level data. We specified the sexual activity groups within each of the three racial/ethnic groups to obtain a contact matrix between nine subgroups. Table 2 presents this contact matrix. (See Appendix A for the technical details about constructing this table from the raw data using certain assumptions.)

### Analyses

#### Logistic Regression

Rather than trying to identify the minimum set of significant factors needed to predict infection status, we used the backward stepwise selection method, a more exploratory procedure that focuses on revealing possible significant factors. The forward selection method is less likely to uncover relationships. The backward step procedure starts with all of the variables in the model. At each successive step, each variable is evaluated for entry or removal. This process continues until no variables meet entry or removal criteria. We adopted the less stringent criterion of a 20% significance level for entry or removal to not dismiss potential significant factors. Logistic regressions for each dependent variable (any STDs, bacterial STDs, and viral STDs) were computed.<sup>18</sup>

#### Log-Linear Analysis for Intra-racial Network Effect

We explore two kinds of network effects. One comes from the sexual network patterns within each racial/ethnic group and the other comes from the sexual network patterns between each racial/ethnic group. The first effect (intra-racial effect) can be revealed by examining how sexual matches between the three sexual activity groups within each racial group differ across these racial/ethnic groups.

Log-linear analysis was used to estimate the population parameters of odds-ratios reflecting racial/ethnic differences in these matches. To find the best fitting log-linear model to the data displayed in Table 2, we examined 12 possible

models. The results are in Table 3. The analysis assumes that each instance of partner choice is an independent event. However, this is clearly not the case for persons who report multiple partners, because their choices, being made by the same persons, are dependent. Given the low average number of partners for most of the population, this problem may not be a serious one.<sup>19</sup>

This method is similar to "multidimensional scaling" in statistics and diverse measures of "clique detection" in network analysis. However, there are important differences. First, the log-linear analysis has a substantively meaningful built-in measure of relationship, the odds-ratio. Second, those odds-ratios come from a log-linear model that passes a goodness of fit test to check if it actually fits the data. Third, it is derived from parameter estimates, not just from sample statistics.

#### A Simulation for the Interracial Network Effect

In addition to the intra-racial network effect noted above, we also found an interracial network effect. This comes from the sexual matching patterns between racial/ethnic groups rather than the patterns within racial/ethnic groups. What happens if there are more (or fewer) sexual partnerships between Hispanic core and African American core members? Do these change the infection rates for African Americans, Hispanics, and whites? We examined this question by simulating different matching patterns between the racial/ethnic groups.

To capture the pure interracial network effect, we must separate this effect from all the other confounding and intermingled effects; two effects are especially worth discussing. First, we must take into account the differential initial infection rates for each group (for example, 4% of the whites and 5% of the Hispanics have ever been infected with gonorrhea, but 24% of the African Americans have been so infected, according to the NHSLs). Second, we must equalize the number of partners of each group because

TABLE 3. Possible Log-Linear Models for the Contact Matrix

Models	Log-likelihood Chi-square	Pearson Chi-square	Degrees of Freedom*	P-value
Independence <sup>†</sup>	4878.3	7839.9	36	0.0
Quasi-independence <sup>‡</sup>	1242.9	1754.3	27	0.0
1-dimensional RC <sup>§</sup>	1104.3	3141.8	28	0.0
2-dimensional RC <sup>  </sup>	236.8	285.9	21	0.0
3-dimensional RC	8.3	7.8	15	0.932
4-dimensional RC <sup>¶</sup>	2.0	2.0	10	0.996
5-dimensional RC	0.8	0.7	6	0.994
6-dimensional RC	0.2	0.2	3	0.978
1-dimensional quasi-RC	298.6	361.1	19	0.0
2-dimensional quasi-RC	95.8	127.5	12	0.0
3-dimensional quasi-RC	2.4	2.2	6	0.9
4-dimensional quasi-RC	0.0	0.0	1	1.0

\*Degrees of freedom are corrected from those suggested in the standard statistical packages (SPSS and CDAS) to reflect the fact that the data set and models are in fact symmetric.

<sup>†</sup>The 'independence' model assumes independence (no relationship) between rows and columns in the contact matrix. People choose partners regardless of race/ethnicity and activity level (periphery, adjacent, and core)—i.e., random mixing. Observing an almost zero p-value for this model tells us this assumption is wrong.

<sup>‡</sup>The 'quasi-' means diagonal cells are dropped in the model. The 'quasi-independence' model thus assumes independence between rows and columns except for the diagonal cells. This model is true if random mixing happens except the diagonal cells (self-selection). Again, the very small p-value tells us this model does not fit the data (contact matrix).

<sup>§</sup>The '1-dimensional RC' model assumes the log-odds ratio of 'i versus j' equals to  $(\mu_i - \mu_j)^2$ , where  $\mu_i$  and  $\mu_j$  is a parameter for i and j, respectively. Thus, the log-odds ratio depends only on one pair of (1-dimensional) parameters for row ( $\mu_i$ ) and column ( $\mu_j$ ), (RC).

<sup>||</sup>The '2-dimensional RC' model assumes log-odds ratio of 'i versus j' equals to  $(\mu_{i1} - \mu_{j1})^2 + (\mu_{i2} - \mu_{j2})^2$ . Thus, the log-odds ratio depends only on two pairs of row and column parameters.

<sup>¶</sup>Even though '4-dimensional RC' has the best goodness of fit (highest p-value), '3-dimensional RC' is the better model, because it requires fewer parameters for estimation (more degrees of freedom) for almost the same goodness of fit. The best model is determined by comparing two models by a chi-square significance test using 'log-likelihood chi-square' and 'degrees of freedom' for all the possible pairs of models.

RC = row-column effect model.

we know that this risk factor strongly affects the infection rates as well. For example, because the African American core group has a higher mean number of partners than the other racial/ethnic core groups, it contributes to the higher infection rate for African Americans.

From a series of matrix manipulations of the contact matrix, we can reveal the effects of different matching patterns on the infection rates for each racial/ethnic group, with all the other confounding effects (including the two discussed above) removed (see Appendix B). The result is the pure network effect regardless of initial infection rates, number of sexual partners, age at initial intercourse, etc.

## Results

### Logistic Regressions for Risk Factors

Table 4 is a summary of the logistic regression analyses produced by the backward selection method for the people who have had at least one sexual partner in their lifetime. The table reports only the statistically significant odds-ratios, indicating that the row variables were significantly associated with the respondents' reports of an STD (as captioned in the column head). If there is no entry in a

row/column intersection, then the factor in question has no significant relationship to having an STD when all the factors in the list are taken into account.

Two highly salient points deserve attention. First, African Americans have the highest infection rate for bacterial diseases, whereas whites, and more educated people, have the highest infection rate for viral diseases. A higher infection rate of viral STDs for whites is readily evident in our data at the zero-order level (8.4% for whites, 5.4% for African Americans, and 4% for Hispanics). This finding needs cautious interpretation, as we indicate in the discussion section.

Second, even after controlling for all the appropriate risk factors, African Americans have the highest infection rate for bacterial diseases. They are five times more likely to be infected than whites or Hispanics after controlling all the risk factors.

### Log-Linear Analysis for the Intra-racial Network Effect

From Table 3, we can identify the three-dimensional row-column (RC) effect model as the best model with a high goodness of fit ( $p = 0.93$ ). The three-dimensional RC model for symmetric data can be represented as follows.

TABLE 4. Logistic Regressions for STDs Ever for Lifetime

	Any	Bacterial	Viral
# of partners	1.01	1.00	1.02
(# of partners)*(# of partners)	(1.00)-		(1.00)-
Female	3.01	2.37	4.07
Ethnicity			
African American	2.16	4.55	0.52
Hispanics	0.62	0.90†	0.49
Whites	1.00	1.00	1.00
Big city	1.32	1.30	
Pay for sex	2.58	2.76	1.67
Being paid for sex	1.56	1.68	
Anal	1.42	1.47	1.47
Never married			1.42
Military service	1.37	1.63	
Drug injection	1.73	2.00	
Group sex			
Concurrent partners	2.33	2.27	2.55
Prepubertal sex	1.68	2.17	
Age	1.10		1.12
(age)*(age)	(1.00)-		(1.00)-
Education			
Less than high school			1.07†
High school grad.			0.72
Vocational school or 2-year degree			1.00
Finished college, 4 to 5 year degree			1.29†
Master's degree or advanced			1.68
Religious attendance	*	*	*
Almost	0.59	0.48	0.62
Sometimes	0.80	0.83†	0.54
Rarely	1.00	1.00	1.00
Thinking of sex			
Never or less than a month	0.73†	0.64	
Few times a month/week	1.00	1.00	
Every day or several times a day	1.19	1.41	
Sexual value orientation			
Most conservative	0.51		
Conservative	1.00		
Liberal	1.17†		
Most liberal	1.71		
Total number of cases	2402	2417	2444
STD (%)	18.3%	11.3%	7.9%
Pseudo R <sup>2</sup>	0.14	0.17	0.13

All values are odds-ratios (alpha = 0.2). Reference categories are italicized. (1.00)-: Rounded odds-ratio is 1.00, but actual odds-ratio is less than 1. \*These are significant but a single odds-ratio cannot be calculated. Instead, each category has an odds-ratio. †Not significantly different from the reference category at alpha = 0.2, although the category as a whole is significant at alpha = 0.2.

$$\ln(F_{ij}) = \gamma + \gamma_i + \gamma_j + \sum_{k=1}^3 \phi_k \mu_{ik} \mu_{jk}$$

where  $F_{ij}$  is an expected frequency for the cell of the contact matrix ( $i = 1, 2, 3, \dots, 9$  and  $j = 1, 2, 3, \dots, 9$ ) in Table 2.

So, the log odds-ratio of 'i vs j' becomes

$$\ln(\theta) = \ln\left(\frac{F_{ij}F_{ji}}{F_{ij}F_{ji}}\right) = \sum_{k=1}^3 \phi_k (\mu_{ik} - \mu_{jk})^2$$

The odds-ratio is obtained by taking the exponential of this value. Based on this best fitting log-linear model, Table 5 presents the odds-ratios between the nine subgroups. It reveals that there is a huge racial/ethnic difference for the odds ratios between "periphery versus core" groups. In the African American population, there are many more sexual matches between periphery and core. For whites, if they are peripheral (instead of being core), they are 180 times more likely to have white peripheral partners (rather than white core partners). However, if they are African American peripherals, they are only 33 times more likely to have African American peripheral partners (rather than African American core partners).

This is an intraracial network effect in the sense that the network effect comes only from the network patterns between different activity groups inside each racial/ethnic group. However, there is another network effect working purely between racial/ethnic groups.

*A Simulation for the Interracial Network Effect*

Using a simulation strategy, we can derive estimates of the "relative infection prevalence" for each racial/ethnic group. "Relative infection prevalence" is calculated by dividing the proportion of infected people in a particular group by the proportion of that group in the population as a whole. For example, if it is less than 1 for whites, it means that whites are less infected with respect to their population size. Figure 1 shows the "relative infection prevalence" for each racial/ethnic group.

The Z-axis represents "relative infection prevalence." The X-axis shows the amount of contacts from the African American core to the Hispanic core and the Y-axis represents the amount of contacts between the white core and the Hispanic core. The intersection of the X and Y coordinates at level '3' depicts the actual (current) matching patterns, according to our data. Larger values than 3 means that there are more interracial contacts between the African American core and the Hispanic core (with respect to the X-axis) or between the white core and the Hispanic core (with respect to the Y-axis). A unit increase means an 1.5% increase.

**Discussion**

Before discussing the findings, we should note several important limitations in the data generated in the NHSLs. First, self-reports of STD status are subject to underreporting biases arising from personal concerns about social stigmatization, failures of recall (especially for disease episodes

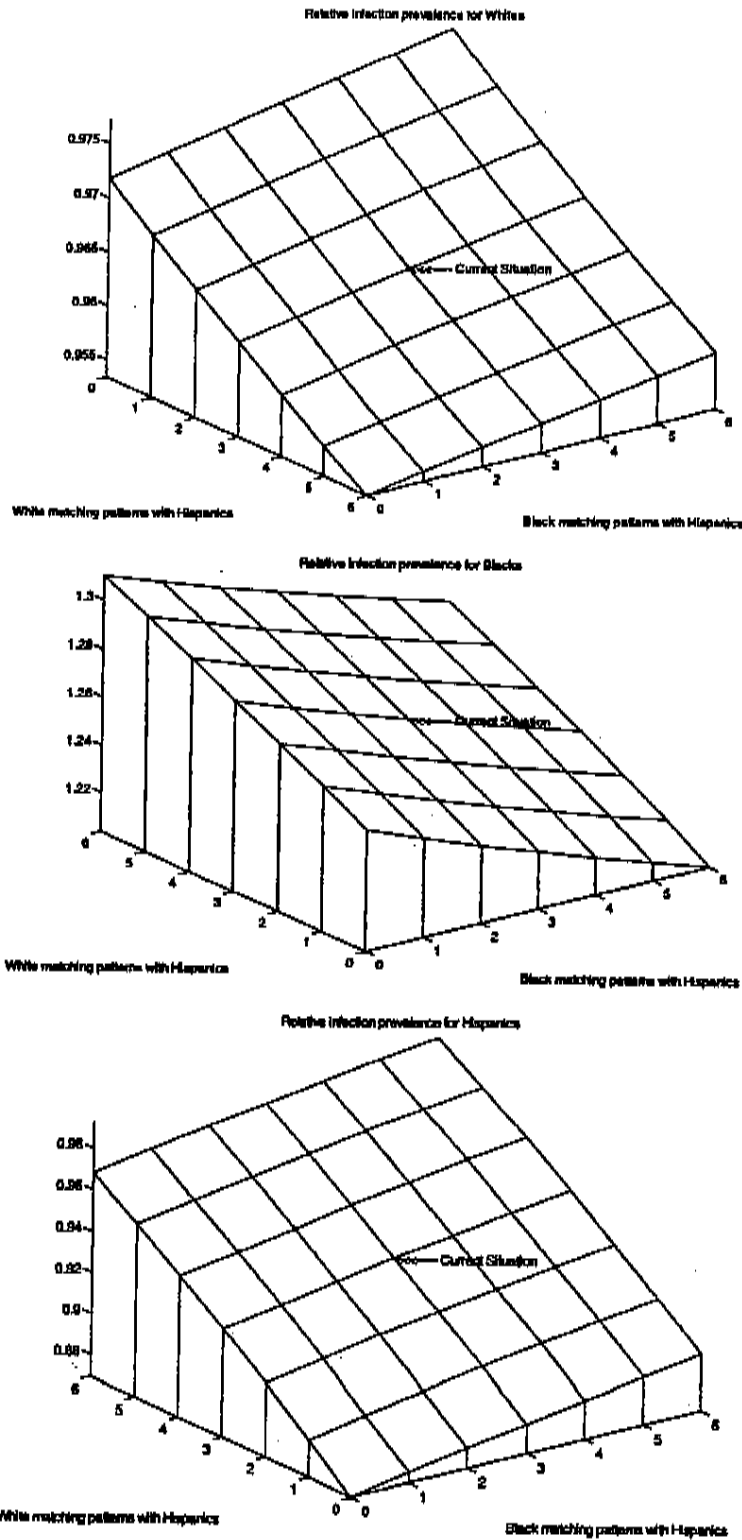


Fig. 1. Relative infection prevalence. For each x-axis and y-axis, '3' shows the current matching amount. Larger than '3' means more matches with Hispanic core group than the current. Smaller than '3' means less matches with Hispanic core group than the current. Z-axis represents "relative infection prevalence."

TABLE 5. Odds-Ratios Between Nine Subgroups\*

	WP	WA	WC	AP	AA	AC	HP	HA	HC
WP									
WA	47.8								
WC	180.4	1.3							
AP	1751.7	9365.0	75280.4						
AA	278951.1	1543.4	2644.0	34.2					
AC	41459.2	369.4	528.3	32.5	1.1				
HP	103.4	4986.2	4446.4	2889.3	32945.3	4660.0			
HA	2281.1	80.1	26.2	17493.3	325.5	66.0	29.8		
HC	18005.7	144.2	39.3	32880.7	154.5	38.1	134.6	1.2	

\*This table displays the odds-ratios between row (chooser) and column (chosen) groups. For example, look at the odds-ratio between WC and WP. White core members are 180 times more likely to have white core partners than they are to have white periphery partners. Or, in the case of HP and WA, Hispanic peripherals are 4,986 times more likely to have Hispanic periphery partners than white adjacent partners. Because the odds-ratio for an ordered row/column combination is the same as for its reversed order, the matrix is symmetric and the upper right-hand entries can thus be omitted to avoid redundancy. Diagonal cell odds-ratios are also omitted, because they are undefined. WP = white periphery; WA = white adjacent; WC = white core; AP = African American periphery; AA = African American adjacent; AC = African American core; HP = Hispanic periphery; HA = Hispanic adjacent; HC = Hispanic core.

more remote in time), and even lack of knowledge that the respondent had a particular sexually transmitted infection because it was clinically asymptomatic or was never diagnosed. Considerable effort was expended in minimizing underreporting by devising an interview protocol that gave a maximum sense of privacy and confidentiality, persuaded the respondent of full disclosure for public health reasons, and provided memory aids to facilitate respondent recall. Even if these procedures were especially effective, and we believe they were, we must still acknowledge that the self-reports understate the prevalence of STDs to a substantial but unknown extent, as shown in a comparison study of self-reported survey data with surveillance data.<sup>14</sup>

Second, there may be systematic biases in underreporting related to particular attributes of the respondents. For example, we might expect that better-educated people may be more knowledgeable and comfortable about disease labels and can recall them more accurately than less-educated persons, who may have less facility and familiarity in talking about such matters. However, countering this advantage in knowledge and recall may be a greater sensitivity to the social stigma associated with reporting a sexually transmitted infection on the part of better-educated persons aspiring to middle class standing.

Third, to construct the contact matrix, we estimated each partner's activity level by averaging two partners' reported number of partners, as specified in Appendix A. We did not have complete information about all the respondents' partnering activity because time limitations prevented direct questions from being asked about more than two partners reported by the respondent in the past year. Moreover, in relying entirely on the respondent's report of his/her own number of partners and the partner(s)' other partners, we necessarily introduce an unknown level of inaccuracy in estimating the contact matrix.

### Logistic Regressions for Risk Factors

First, we found that whites are more likely to be infected with viral STDs. This pattern contrasts sharply with the image of STD prevalence suggested by figures reported by the CDC that imply higher prevalence of certain viral diseases among lower socioeconomic groups.<sup>2,5,12</sup> This result implies the possibility of systematic bias in the reported data to the CDC. However, there are several other possible interpretations.

First, given that tests for viral STDs tend to be more costly than tests for bacterial STDs and viral STDs are less asymptomatic than bacterial STDs, we should expect some bias in the likelihood of diagnosing viral STDs in the direction of persons of higher socioeconomic status. If this were the case, self-reported data would indicate that whites have a higher infection rate than African Americans when, in fact, African Americans have the higher rate. One study using nationally based biomarker data from the National Health and Nutrition Surveys,<sup>20</sup> for example, has shown that African Americans are more likely than whites to be infected by herpes simplex virus type 2.

Second, this racial/ethnic difference might arise from African Americans engaging in more risky sex (for example, sex without condoms). We did not include a "sex without condom" variable in the logistic regression because of the lack of relevant data. However, we observe no racial or ethnic difference in this variable, at least for the past 12 months. Table 6 shows this data.

Third, a number of variables in the logistic regression are highly associated with reporting an STD but are not conventionally interpreted as STD risk factors—such as gender, concurrent partners, educational attainment, religious attendance, and thinking of sex. In other words, how can one explain how "thinking about sex" frequently or being a college graduate might themselves increase the probability

TABLE 6. Mean Frequency of Unprotected Sexual Events (No Condom) With a Nonspouse or Noncohabitant

	White	African American	Hispanic
Peripheral	8.6 (39.3)	12.5 (43.4)	5.8 (25.9)
Adjacent	54.7 (102.2)	34.4 (71.3)	27.6 (66.2)
Core	57.3 (112.3)	19.6 (48.3)	59.1 (153.7)

Standard deviations are in parenthesis.

of being infected with an STD? Our basic conjecture is that people with a higher probability of being infected in these "non-risk factor" categories have partners who are more sexually active than the sex partners of the people in the lower "non-risk factor" categories. For example, the reason women have higher odds than men (with the same number of partners) for being infected is that, in general, women's male partners have higher numbers of partners than men's female partners. It is therefore possible that women have a higher infection rate than men with the same number of sex partners. (This gender difference is also surely rooted in biologic considerations as well, such as differential duration of exposure to the pathogen in question.<sup>21</sup>) In the next section, we examine the case for race/ethnicity.

#### *Log-Linear Analysis for the Intraracial Network Effect*

Here we can find one clear explanation for why African Americans have higher infection rates, even after controlling for most risk factors. We found that, even though African American peripheral people have, by definition, only one partner, the chance that their partners are in the core is five times higher than it is for white peripheral people ( $180/33 \cong 5$ ) and four times higher than for Hispanic peripheral people ( $133/33 \cong 4$ ). Therefore, even after controlling for the number of partners as a major risk factor, African Americans necessarily have higher infection rates than whites or Hispanics. For African Americans, then, infections are not limited to the core but overflow to the periphery because of this strong assortative mating pattern. This is apparent if we consider the fact that infection is concentrated in the core and the core group is such a powerful transmitter because of its high partner turnover rate. This can be seen as a special case of the general proposition that a high number of sexual linkages between subgroups will lead to extensive dissemination through the population rather than to confined infections within the subgroups. This proposition has been tested and supported by many papers.<sup>22-24</sup>

There is a tradeoff for this. If the outflow from the African American core is too high, the African American core cannot maintain STDs over time because the transmission power of their partners (the number of partners of their partners) is too weak (the amount of second-order transmission is too low). However, the current amount of outflow

from the African American core is not sufficiently large to eliminate its role as a core group. We can confirm this proposition by testing whether the African American core is actually functioning as a core group in the sense that they are the primary actors in sustaining disease transmission. If we calculate the basic reproduction rates of gonorrhea as an example, with appropriate parameters concerning transmission probabilities and recovery rates, after the case in Garnett and Anderson,<sup>8</sup> we find that the core groups in our analysis are real cores in the sense that their basic reproduction rates are all greater than 1 and the basic reproduction rate for the whole population is also greater than 1, according to the model proposed by Jacquez, Simon, and Koopman.<sup>17</sup>

#### *A Simulation for the Interracial Network Effect*

Another network reason why African Americans have a higher infection rate can be found in this interracial effect. Figure 1 shows that, as interracial contacts between the African American core and the Hispanic core increase, the relative infection prevalence of the African American population decreases, given the white matching pattern. Also, more contacts of the white core with the Hispanic core increase the relative infection prevalence for the African Americans, given the African American matching pattern. In general, if a given non-Hispanic group has more sexual matches with an Hispanic group, its infection rate decreases whereas the other two racial groups' infection rates increase. We can interpret this result as follows: more matches with Hispanics provide an efficient way to transmit infections to the other racial/ethnic groups. If a group has fewer matches with the Hispanic core, then the group becomes more segregated from the other racial/ethnic groups, and this increases the infection rate for that group and decreases the infection rate for the other groups. Otherwise, if there are more matches, the group has an effective channel to relay infection to the other racial/ethnic groups.

Whites have relatively more sexual contacts with Hispanics than African Americans have, as can be observed in Table 5. Every odds-ratio between whites and Hispanics is smaller than or almost equal to the corresponding one between African Americans and Hispanics. Note these odds-ratios are obtained after controlling population size. This network pattern makes the white infection rate low and the African American infection rate high. Once infected, whites can spread infection to the other racial/ethnic groups more effectively than African Americans can. This is why the relative infection prevalence of African Americans is 1.26 whereas it is only 0.97 for the whites for the current matching pattern in which Hispanics are closer to whites. Therefore, the likelihood of African Americans having a sexually transmitted infection is 1.3 ( $\cong 1.26 / .97$ ) times greater than it is for whites, because of this factor alone.

In summary, relatively high sexual contacts between the African American core and the African American periphery (dissortative mating within the African American population) has a network effect that makes infections overflow into the entire African American population from the African American core—an intraracial network effect. At the same time, African Americans are somewhat distant from the other racial/ethnic groups (assortative mating of African Americans in the whole population), so infections stay inside the African American population—an interracial effect.

**Conclusion**

With respect to our most important results, we found, first, that, in addition to the usual behavioral risk factors, one can identify a variety of social and attitudinal items that provide stable and consistent predictors of STD prevalence and incidence that could be usefully combined with the STD surveillance system to provide more comprehensive, effective, and targeted intervention strategies for population subgroups at risk.

Secondly, we demonstrated the critical role social network patterns play in accounting for the known differentials in rates of infections across racial and ethnic groups. In particular, we found that the relatively high sexual contacts between the African American core and its periphery facilitate the spread of infection overflow into the entire African American population: the so-called intraracial network effect. These infections remain inside the African American population because African Americans are highly segregated from the other racial and ethnic groups: the so-called interracial network effect. These network effects cannot be detected in regressions that include only individual-level risk factors.

**References**

1. Aral SO. The social context of syphilis persistence in the Southeastern United States. *Sex Transm Dis* 1996; 23:9-15.
2. Centers for Disease Control and Prevention, Division of STD Prevention. Sexually Transmitted Disease Surveillance, 1995. Atlanta: U.S. Department of Health and Human Services, Public Health Service; September 1996.
3. Louis F, Aral SO. Untangling the persistence of syphilis in the South. *Sex Transm Dis* 1996; 23:1-4.
4. Nakashima AK, Rolfs RT, Flock ML, Kilmarx P, Greenspan JR. Epidemiology of Syphilis in the United States, 1941-1993. *Sex Transm Dis* 1996; 23:16-23.
5. Philipson JT, Posner AR. Private choices and public health: the AIDS epidemic in an economic perspective. Cambridge, MA: Harvard University Press, 1993.
6. Sabogal F, Faigles B, Catania J. Multiple sexual partners among hispanics in high-risk cities. *Fam Plann Perspect* 1993; 25:257-262.
7. Institute of Medicine. Hidden Epidemic. Thomas RE, William TB, eds. Washington, DC: National Academy Press, 1997.
8. Garnett G, Anderson R. Contact Tracing and the estimation of sexual mixing patterns: the epidemiology of gonococcal infections. *Sex Transm Dis* 1993; 20:181-191.

9. Rothenberg RB. The geography of gonorrhea. *Am J Epidemiol* 1983; 117:688-694.
10. Potterat J, Rothenberg RB, Woodhouse DE, Muth JB, Pratts CI, Fogle JS. Gonorrhea as a social disease. *Sex Transm Dis* 1985; 12:25-32.
11. Laumann EO, Gagnon JH, Michael RT, Michaels S. The social organization of sexuality: sexual practices in the United States. Chicago: The University of Chicago Press, 1994.
12. Aral SO, Holmes KK. Epidemiology of sexual behavior and sexually transmitted diseases. In: Holmes KK, Mardh PA, Weisner P, Cates W, Lemon SM, Stamm WE, eds. Sexually Transmitted Diseases, 2nd. ed. New York: McGraw-Hill, 1990:19-36.
13. Centers for Disease Control and Prevention, Division of STD Prevention. Sexually Transmitted Disease Surveillance, 1996. Atlanta: U.S. Department of Health and Human Services, September 1997.
14. Anderson JE, McCormick L, Fichtner R. Factors associated with self-reported STDs: data from a national survey. *Sex Transm Dis* 1994; 21:303-308.
15. Harris School of Public Policy. Monitoring Sexually transmitted disease: a comparative study of five Midwestern cities. Unpublished research, 1994.
16. Yorke JA, Hethcote HW, Nold A. Dynamics and control of the transmission of gonorrhea. *Sex Transm Dis* 1978; 5:51-56.
17. Jacquez J, Simon C, Koopman J. Core groups and the ROs for subgroups in heterogeneous SIS and SI models. In: Mollison D, ed. Epidemic Models: Their Structure and Relation to Change. Cambridge, UK: Cambridge University Press, 1995:279-301.
18. Menard S. Applied Logistic Regression. Thousand Oaks, CA: Sage, 1995.
19. Morris M. Epidemiology and social networks: modeling structural diffusion. *Sociol Meth Res* 1993; 22:99-126.
20. Fleming DT, McQuillan GM, Johnson RE, et al. Herpes simplex virus type II in the United States, 1976 to 1994. *N Engl J Med* 1997; 337:1105-1111.
21. Jones RB, Wasserheit JN. Introduction to the Biology and Natural History of sexually transmitted diseases. In: Wasserheit JN, Aral SO, Holmes KK, Hitchcock PI, eds. Research Issues in Human Behavior and Sexually Transmitted Diseases in the AIDS Era. Washington, DC: American Society for Microbiology, 1991; 11-37.
22. Wasserheit JN, Aral SO. The dynamic topology of sexually transmitted disease epidemic: implications for prevention strategies. *J Infect Dis* 1996; 174(suppl):201-213.
23. Rothenberg RB, Potterat JJ, Woodhouse DE. Personal risk taking and the spread of disease: beyond core groups. *J Infect Dis* 1996; 174(suppl):144-149.
24. Rothenberg RB, Potterat JJ. Temporal and social aspects of gonorrhea transmission: the force of infectivity. *Sex Transm Dis* 1988; 15:88-92.

**APPENDIX A: Constructing the Contact Matrix**

First, we construct the contact matrix for the past twelve months for each gender from the raw data, as follows.

**MATRIX A. Male Respondents to Female Partners (Row to Column): Number of Partnerships**

		White	African Americans	Hispanic
Peripheral	White	659	3	9
	African American	10	55	0
	Hispanic	15	0	34
Adjacent	White	261	2	6
	African American	2	68	3
	Hispanic	20	0	22
Core	White	287	0	10
	African American	14	133	12
	Hispanic	15	1	6

**MATRIX B. Female Respondents to Male Partners (Row to Column): Number of Partners**

		White	African Americans	Hispanic
Peripheral	White	882	12	16
	African American	1	135	4
	Hispanic	33	1	56
Adjacent	White	182	9	5
	African American	0	60	0
	Hispanic	3	1	15
Core	White	121	9	1
	African American	0	44	0
	Hispanic	2	9	13

**MATRIX D. Female Respondents to Male Partners (Row to Column): Row %**

		Peripheral	Adjacent	Core
White: P	P	94.73	3.29	1.98
	A	24.26	44.85	30.88
	C	15.23	43.15	41.62
African American: P	P	89.29	7.14	3.57
	A	16.28	51.16	32.56
	C	22.58	50	27.42
Hispanic: P	P	90	3.33	6.67
	A	0	40	60
	C	3.57	46.43	50

These two matrices lack one critical piece of information. Neither table includes any information about the level of sexual activity for partners. We do not know, for example, how many partnerships were formed between, say, white core members and white adjacent members. To remedy this deficiency, we constructed the following two mixing matrices that show the percentages of partnerships across activity-level subgroups.

Unfortunately we do not have complete information about the partners' level of sexual activity because limitations prevented direct questions from being asked about all partners reported by the respondent in the past year; it allowed questions about only the primary and secondary partner. Averaging these two partners' reported number of partners, excluding the respondent, we assigned all the partners to the periphery if the average was zero, to the adjacent sub-population if the average was between zero and two (including two), and to the core if the average was greater than two. Even though this assumption seems strong, the resultant tables are consistent with James Koopman's observation of a campus contact matrix that manifested strong racial differences in the patterns of matching (J. Koopman, written communication, November 1996). Second, we can now construct the following mixing matrix using this assumption.

**MATRIX C. Male Respondents to Female Partners (Row to Column): Row %**

		Peripheral	Adjacent	Core
White: P	P	95.22	2.69	2.09
	A	25.28	45.35	29.37
	C	8.75	47.47	43.77
African American: P	P	93.85	3.08	3.08
	A	24.66	53.42	21.92
	C	14.89	56.03	29.08
Hispanic: P	P	93.88	0	6.12
	A	35.71	16.67	47.62
	C	0	34.78	65.22

Third, we integrated Matrix A and Matrix C into one matrix. We assumed the row percentages in C applicable to each cell in A, regardless of the partners' race/ethnicity. Therefore, among 659 white female partners of white peripheral male respondents, 95% are peripheral. Also, among the 9 Hispanic female partners of the white peripheral male respondents, 95% are peripheral. We then constructed a 9-by-9 matrix (Matrix M) representing contacts from men to women. In the same way, we computed another 9-by-9 matrix from B and D representing contacts from women to men (Matrix F). Fourth, the two matrices (M and F) must, in principle, give us the same information. For example, if there are 100 partnerships between the white male core and the white female periphery, then there must also be 100 partnerships between the white female periphery and the white male core. What discrepancies we observe arise from sampling variability. We propose to resolve these discrepancies by averaging the male-to-female and female-to-male estimates to obtain a single contact matrix. We then computed the final 9-by-9 symmetric contact matrix (Matrix C) by using the equation  $C = 1/2 [(M + F) + (M + F)']$ .

**APPENDIX B: A Simulation of Different Matching Patterns Between Race/ethnic Groups**

Our goal in this simulation is to show the effects of different matching patterns between racial/ethnic groups, with the two effects (the differential initial infection rates and the different numbers of partners for each subgroup) removed from the pure interracial network effect. First, we transform Matrix C (Table 2) into the transformation Matrix T by the following equation.

$$T = [C_{ij} / \sum_{j=1}^9 C_{ij}] \quad (1)$$

Therefore, equation (1) constructs T as a transposed matrix of the row stochastic matrix of C. Then, every one of the nine groups has the same number of partners because of the

"row stochastic" procedure, and the transformation matrix  $T$  becomes the mixing matrix from column to row.

Using  $T$ , we can calculate each group's proportion of infected people at time 2 from each group's proportion of infected people at time 1 (the sum of the each group's proportions must be 1). Equation (2) shows this relation, where  $r_1$  and  $r_2$  are column vectors representing each group's proportion of infected people at time 1 and time 2, respectively.

$$Tr_1 = r_2 \quad (2)$$

Also, because  $T$  is a "regular" matrix, it must have a "stable" proportion for each group that does not change over time. The following equation shows the relationship between  $T$  and the "stable" proportion for each group,  $s$  (where  $s$  is a column vector representing each group's stable proportion of infected people).

$$Ts = s \quad (3)$$

One of the useful properties of  $s$  is that it does not depend on the initial  $r_1$ . So we also eliminate the differential initial infection rates for each group. We can solve  $s$  from equation (3) and the following equation (4).

$$\sum_{i=1}^9 s_i = 1 \quad (4)$$

After getting  $s$ , we divide  $s$  into 3 parts according to the following equation (5).

$$\sum_{i=1}^3 s_i = y_1, \quad \sum_{i=4}^6 s_i = y_2, \quad \sum_{i=7}^9 s_i = y_3 \quad (5)$$

Using this equation, we get  $y_1$ ,  $y_2$ , and  $y_3$  for each racial/ethnic group, representing each racial/ethnic group's proportion of infected people. So if  $y_1 = 0.6$ , then 60% of the infected people are whites. If we standardize these with regard to each racial/ethnic group's population size, we finally get  $z_1$ ,  $z_2$ , and  $z_3$  for each racial/ethnic group.

This is done by the following equation (6) (each denominator is each group's proportion of the whole population represented in the contact matrix  $C$ ).

$$z_1 = y_2/0.7966, \quad z_2 = y_2/0.1277, \quad z_3 = y_3/0.0756 \quad (6)$$

If  $z_1$  is less than 1, it means that the white community has an under-represented number of infected people; and if it is greater than one, it means there is an over-representation of infected people for whites. We call this the "relative infection prevalence."

We calculated a series of  $z_1$ ,  $z_2$ , and  $z_3$  from a simulation that changes the two matching patterns simultaneously to see what the effects of these changing matching patterns on the infection rates are. First, there is a change in the amount of contacts between the African American core and the Hispanic core. This means there is a change in  $T_{6,9}$  and  $T_{9,6}$  (We assume the same amount of change in  $T_{6,6}$  and  $T_{9,9}$  so that the sum of each column will be one.) Second, the amount of contacts between the white core and the Hispanic core ( $T_{3,9}$  and  $T_{9,3}$ ) are also changed at the same time. (For the same reason, this entails the same amount of change in  $T_{3,3}$  and  $T_{9,9}$ .)

Now we can estimate the differential "relative infection prevalence" for each group based on a set of different matching patterns. These results do not depend on the initial infection rates or the numbers of partners for each subgroup.