10

SPECIFICATION AND ESTIMATION OF HETEROGENEOUS DIFFUSION MODELS

*Henrich R. Greve**
David Strang†
Nancy Brandon Tuma‡

Heterogeneous diffusion models let one combine the analysis of intrinsic propensities with that of intrapopulation contagion, and to disaggregate contagion effects into individual susceptibilities, the infectiousness of prior adopters, and the social proximity of prior-potential adopter pairs. This paper reports the results of a series of Monte Carlo simulation studies that investigate estimation issues for this class of models. Graphical analysis of population-level hazard rates is shown to provide little insight into these processes. We focus on the properties of maximum likelihood estimators, considering variation across parameter values and different forms of model misspecification. When models are correctly specified, we find few conditions under which estimation appears problematic. Difficult cases involve binary networks where network linkages have very strong effects or network density is high. Estimation deteriorates in

This paper discusses work presented in part at the 1993 meetings of the American Sociological Association. Research support was provided by the National Science Foundation (SES 8911666, SES 9213152, and SES 9213258). We thank Eric Bloch for assisting with the development of the program used to generate pseudo-random event histories, David J. Pasta for his comments on the paper and programming assistance, and the editor and anonymous reviewers of *Sociological Methodology* for their helpful suggestions.

*University of Tsukuba
†Cornell University
‡Stanford University

some characteristic ways when models are misspecified. For example, propensity and susceptibility effects are readily confused. An effective model specification strategy is to include variables in all theoretically plausible components of the model rather than to test alternative covariate locations sequentially. Processes where a covariate affects the hazard in multiple ways (for example, has both propensity and infectiousness effects) are successfully parsed in correctly specified models. In general, results offer considerable encouragement for analysts who wish to estimate and test heterogeneous diffusion models.

Considerable effort has been made in recent research to specify the lines along which social influence flows. Interest in this problem stems largely from Burt's (1987) discussion of social cohesion and structural equivalence as alternative ways in which populations are meaningfully connected (also see Burt 1983; Friedkin 1984; Marsden and Friedkin 1993). This discussion carries substantial theoretical interest because it addresses the relative importance of direct social relations, structural location, and competition as bases for social action. Recent empirical work in sociology seeking to specify diffusion processes includes Knoke (1982) on municipal reform as a geographic phenomenon, Strang (1990, 1991*a*) on decolonization within empires and regions, and Davis (1991) on corporate strategies mediated by interlocking directorates.

The effort to specify substantive linkages underlying patterns of diffusion has deflected attention away from deterministic, homogeneous diffusion models (for a review, see Mahajan and Peterson 1985) and toward models that are both more micro-analytic and more relational. In particular, Strang and Tuma (1993) suggest modeling diffusion within an expanded event-history framework (see also Marsden and Podolny [1990] and Strang [1991*b*]). A *heterogeneous diffusion model* permits the testing of specific hypotheses about social structure while meshing well with the intrinsically temporal character of an adoption process. This modeling framework has been applied empirically by Strang and Bradburn (1993) and Greve (1994*a*, 1995). A related but different model has been developed and applied by Tuma and Ingram (1993).

The heterogeneous diffusion model offers an alternative to two current approaches to studying diffusion in a social structure.

Static approaches to modeling social structure, such as the spatial effects models discussed in Doreian (1981), seem less satisfactory given the key role of temporal ordering and the presence of time-varying covariates and right censoring in adoption processes. Log-linear models, such as those used to examine how the diffusion of diseases is channeled by social networks (Morris 1993), are similar to the heterogeneous diffusion model because they are also non-equilibrium models of diffusion. The present framework is more general, however, because it can easily be used to model a multivariate causal structure.

Diffusion models posit interdependent events—what one actor does affects what other actors do thereafter. The analytic difficulties produced by interdependent events are attenuated by the organization of the process through time: earlier events can affect later events, but later events cannot affect earlier events. Despite this simplification, it is unclear *a priori* how well one can empirically estimate models of stochastic processes where events are interdependent.

Concerns about estimation are heightened by the complexity of these models. Excepting recent work, most analyses of diffusion parameterize the process in a fairly simple way. Frequently only deterministic models are considered, and structures of communication and influence are often treated as homogeneous. When heterogeneity is incorporated, it is common to assume a particular "mixing" distribution or to include only a single measure of social proximity. In contrast, the heterogeneous diffusion model we consider permits a highly disaggregated and relationally specific analysis of diffusion. Whether quality of estimation can keep pace with computability becomes an important issue, which this paper addresses.

Strang and Tuma (1993) provide a limited demonstration of the estimability of heterogeneous diffusion models with measured covariates. They generated several populations of event histories under a specific heterogeneous diffusion model and then estimated the model by maximum likelihood using the simulated histories as data. All parameters were recovered without bias and with a relatively low estimated variance. In particular, the quality of the estimated parameters characterizing contagion between pairs in the population was actually slightly better than that of parameters characterizing intrinsic propensities to adopt (Strang and Tuma 1993, table 1).

But an empirical demonstration of successful estimation is hardly sufficient. To be generally useful in empirical research, the quality of estimated parameters needs to be good across a wide range of conditions. This paper explores the quality of estimated parameters in heterogeneous diffusion models, focusing especially on problems common in empirical research. Three main lines of inquiry are pursued.

We first plot nonparametric estimates of the hazard rate and the integrated hazard rate in the population to gauge how variation in parameters affects the overall pattern of events in a population driven by different *types* of diffusion processes. Attention to diffusion as a causal process is often motivated by particular empirical patterns in data, such as a sharply rising hazard over time. Given the generality permitted by individual-level effects within a stochastic framework, however, the reliability of this strategy seems uncertain.

Second, we explore how estimator quality varies across substantively different heterogeneous diffusion models. One main limitation of Strang and Tuma's (1993) Monte Carlo study is its focus on a single set of parameter values. We investigate how estimator quality depends on the coefficients of different types of covariates, focusing on the cases where particular quantities numerically dominate the process.

Third, we examine estimator quality under various forms of model misspecification. We consider situations where models are overspecified (i.e., include variables not actually used to generate the data analyzed) and underspecified (i.e., exclude variables used to generate the data). We further examine how estimator quality and inferences based on standard statistical tests fare under more subtle forms of misspecification, where a covariate having one kind of effect on the diffusion process is modeled as having some other kind of effect. (For example, a covariate that affects susceptibility to others' influence may be incorrectly modeled as affecting the actor's propensity to adopt; see below for a definition of these terms.) We evaluate strategies for exploratory data analysis where particular covariates are correctly tied to adoption but the *location* of their impact on the diffusion process is not well understood.

We should note that the interdependence of events presumed in diffusion analyses also raises issues about estimation from incomplete data on populations. The diffusion models we study (and all those with which we are familiar) presume population rather than sample data. Sampling from a population when events are interdependent leads to incomplete information about variables on the right-hand side, since adoption events enter the estimation equation both as explanatory covariates and as outcomes. This makes it unclear whether, and under what conditions, one can correctly estimate a diffusion model without having data on the entire population. The analysis of sample data is deferred to a companion paper (for preliminary results, see Greve, Strang, and Tuma 1993).

1. HETEROGENEOUS DIFFUSION MODELS

The *heterogeneous diffusion model* proposed by Strang and Tuma (1993) has several distinctive features and advantages relative to the classical homogeneous diffusion model and previous extensions thereof. First, "heterogeneity" here refers to measured covariates. Some previous authors have proposed diffusion models with "mixing distributions" in which social distance is assumed to be a random variable with some postulated distribution (e.g., gamma, logistic); for a review, see Mahajan and Peterson (1985). Such models may fit empirical data better than the classical homogeneous model, but they do not give insight into the social mechanisms affecting adoption.

Second, the model incorporates both "intrinsic" propensities to adopt and contagion resulting from the prior adoptions of others in the population. Intrinsic propensities include internal predispositions of the focal case, and the impact of external sources of diffusion (such as communications from outside the adopting population that directly reach potential adopters). By contrast, contagion operates via social linkages between pairs of population members, one still at risk of adoption and another that has already adopted. Its impact is decomposed into factors describing a focal case's susceptibility to contagion, the infectiousness of the prior adopter, and the pair's social proximity.

Third, like the diffusion models considered by Bartholomew (1982), the heterogeneous diffusion model proposed by Strang and Tuma (1993) is a *stochastic* model. In this respect it contrasts with the many deterministic diffusion models that have been applied empirically in the past. Hence, even for a fixed set of parameter values and for given values of covariates, realizations of the heterogeneous diffu-

sion model exhibit random variation in the timing of events. Moreover, if the covariates have a random distribution (which we assume below), realizations of the covariates introduce further variation in the adoption process. Stochastic variability in the process complicates understanding the implications of the model.

In this paper, we examine the *additive model* defined in Strang and Tuma (1993). Consider a partition of the members of a population into two sets: The set $\mathcal{N}(t)$ consists of all nonadopters at time t, and the set $\mathcal{S}(t)$ consists of prior adopters (termed "spreaders" by Strang and Tuma). A representative member of $\mathcal{N}(t)$ is denoted by n, and one in $\mathcal{S}(t)$ by s. The sizes of the two sets are $N(t)$ and $S(t)$, respectively. The hazard rate for the members of $\mathcal{N}(t)$ (the risk set) is modeled as

$$r_n(t) = \exp(\boldsymbol{\alpha}'\mathbf{x}_n) + \sum_{s\in\mathcal{S}(t)} \exp(\boldsymbol{\beta}'\mathbf{v}_n + \boldsymbol{\gamma}'\mathbf{w}_s + \boldsymbol{\delta}'\mathbf{z}_{ns})$$

$$= \exp(\boldsymbol{\alpha}'\mathbf{x}_n) + \exp(\boldsymbol{\beta}'\mathbf{v}_n) \sum_{s\in\mathcal{S}(t)} \exp(\boldsymbol{\gamma}'\mathbf{w}_s + \boldsymbol{\delta}'\mathbf{z}_{ns}). \quad (1)$$

Here,

- $\mathbf{x}_n$ is a vector of variables describing n's *propensity* to adopt (i.e., net of any contagion via intrapopulation linkages).
- $\mathbf{v}_n$ is a vector of variables describing n's *susceptibility* to contagion.
- $\mathbf{w}_s$ is a vector of variables describing the *infectiousness* of s (for all n).
- $\mathbf{z}_{ns}$ is a vector of variables describing the *proximity of* n and s (the infectiousness of s for a specific n or equivalently, the susceptibility of n for a specific s).

We take the first elements of $\mathbf{x}_n$ and $\mathbf{v}_n$ to be unity and refer to the associated parameters as the propensity and contagion intercepts, respectively. The former indexes the baseline propensity to adopt and the latter the baseline impact of adoptions by others. Note that by convention we locate the contagion intercept in the susceptibility term (i.e., we associate it with $\mathbf{v}_n$). We do not include a leading element of unity in the vectors associated with infectiousness and social proximity because only one intercept in the contagion component is identifiable.

A natural alternative formulation of diffusion treats prior events as multiplying the hazard:

$$r_n(t) = \exp(\boldsymbol{\alpha}'\mathbf{x}_n) \prod_{s\in\mathcal{S}(t)} \exp(\boldsymbol{\beta}'\mathbf{v}_n + \boldsymbol{\gamma}'\mathbf{w}_s + \boldsymbol{\delta}'\mathbf{z}_{ns}). \quad (2)$$

For applications, see Levin, Levin, and Meisel, 1987; Marsden and Podolny 1990; Strang 1991*a;* and Burns and Wholey 1993. More complex models might add temporal heterogeneity to either the additive or multiplicative model so that influence might vary with time since adoption; see Strang and Tuma 1993 for some discussion.[1]

Because of space and resource limitations, we restrict our attention to additive models of diffusion. To our knowledge, this paper provides the first broad analysis of the estimation quality of a microlevel diffusion model via simulation methods, and the range of issues considered is quite large. We anticipate that further statistical and empirical work on this and other diffusion formulations (such as multiplicative diffusion models and models that incorporate explicit time dependencies) will lead toward cross-model comparisons. As a first step, this paper thus seeks a close assessment of estimation issues for one relevant formulation of a diffusion process.

Given this strategy, some review of the motivation for an additive structure for diffusion influences is in order. Like most other hazard models, perhaps the first important property of the functional form in (1) is nonnegativity; the form of the model precludes the estimation of theoretically meaningless negative hazards. Beyond this, however, we would point to two important properties of an additive formulation.

First, an additive formulation does not impose structure on how the influence of prior adoptions varies with their historical ordering. The sth event simply increments the hazard for the previous ($s - 1$)th event by

$$\Delta r \equiv r_s(t) - r_{s-1}(t) = \exp(\boldsymbol{\beta}'\mathbf{v}_n + \boldsymbol{\gamma}'\mathbf{w}_s + \boldsymbol{\delta}'\mathbf{z}_{ns}). \quad (3)$$

In contrast, many formulations imply a systematic relationship between historical order and contagious influence. For example, the multiplicative framework in (2) implies

$$\Delta r = [\exp(\boldsymbol{\beta}'\mathbf{v}_n + \boldsymbol{\gamma}'\mathbf{w}_s + \boldsymbol{\delta}'\mathbf{w}_{ns}) - 1]\, r_{s-1}(t). \quad (4)$$

[1]Either the additive or multiplicative model may include time-varying covariates whether or not there is temporal heterogeneity in influence.

Here the impact of each additional event on the hazard rate is proportional to the accumulated influence of all prior events as summarized in $r_{s-1}(t)$, the rate for the previous, $(s-1)$th event. When contagious influence is positive (probably the most common case), (4) implies that later adopters have larger effects than earlier ones. This is often an undesirable implication; in fact, it is common to seek models that specify decreasing incremental influence (Mahajan, Muller, and Bass 1990).

A second motivation for an additive formulation is that it treats intrinsic propensities and contagious influence as more separable than do alternative formulations. Since modeling contagion in terms of measured covariates is in its infancy relative to research detailing variation in intrinsic propensities, and since we regard analyses of contagion as speaking to central sociological concerns, this separability appears an attractive feature. Models that reduce the likelihood of confusing propensities with contagion seem useful tools for beginning to map out channels of influence within populations.[2] One aim of this paper is to ascertain how well this logic translates into practice, by asking whether in fact estimation can effectively distinguish variations in propensities from variations in contagious influence (see especially Sections 4.3–4.5).

2. SIMULATION METHODOLOGY

We perform a series of Monte Carlo studies to evaluate characteristics of estimation of the heterogeneous diffusion model. In each *condition,* we examine a specific instance of the general model defined in (1). We perform 50 trials for each condition—increasing the number of trials to 100 had little impact on conclusions but doubled computation time. In each trial, event histories for the population are generated using pseudo-random methods.[3] These histories are

[2]We should also note the most obvious limitation of the additive formulation in (1): It allows only positive or reinforcing effects of contagion. Multiplicative formulations, by contrast, permit prior adoptions to decrease as well as increase the hazard of the focal case, capturing empirical contexts where actors seek to avoid the actions of (certain) others in the population. Where the assumption of positive contagion is not warranted, some alternative or amendment to the additive formulation portrayed here must be employed.

[3]We used a program called EHG (Event History Generator), which was begun by James C. Crutchfield and extensively enhanced by Eric Bloch under Tuma's guidance. Later it was modified for the purpose of this study by Bloch and Greve.

then analyzed under some model of interest. Analyses utilize a version of RATE (Tuma 1980) modified to estimate the heterogeneous diffusion model in (1) by the method of maximum likelihood.

These Monte Carlo studies are designed to investigate diffusion under conditions similar to those usually encountered in empirical research. In particular, many empirical analyses of diffusion examine populations with 50 to a few hundred members—such as closed communities of people, organizations within an industry or geographical area, the American states, or countries. It is unclear whether the asymptotic properties of maximum likelihood estimation translate into coefficient estimates close to true parameter values for such population sizes, especially given the information demands made by complex diffusion models. In all trials reported below, we study a population of 100 potential adopters.[4] At the beginning time, $t_0 = 0$, all 100 are at risk of adoption, and no events have occurred. Over time, all population members experience a single, nonrepeatable adoption event.

Covariates measuring propensity, susceptibility, and infectiousness are drawn from statistically independent, standard Gaussian distributions using a pseudo-random number generator. Since Gaussian distributions are symmetric, the impact of a covariate on the diffusion path depends only on the magnitude of its coefficient and not its sign. We therefore examine only positively signed parameters for all covariates, without loss of generality.[5] Conclusions might differ if the covariates had some other type of distribution (e.g., Bernoulli, log-Gaussian) or were correlated (but see Section 4.5 for some evidence on estimation under conditions of perfect collinearity across model components).

Covariates for social proximity are constructed to parallel the ways social researchers typically think about and measure linkages in a social network. The most common version has the form of a *binary incidence matrix,* where a link from one member of the population to another is indicated by a value of one and the absence of a link

[4]Exploratory study of populations with up to 500 members revealed a moderate degree of improvement in estimation as the size grows, so the results below should be relevant to such population sizes. Simulating populations over 500 is extremely time-consuming because the time to create and analyze datasets is quadratic in population size.

[5]This simplification would limit generality if the covariates had an asymmetric distribution—for example, a log-Gaussian distribution.

equals zero. In most analyses reported below, each potential adopter is linked to k randomly chosen alters. The resulting incidence matrix is thus asymmetric with zeros on the diagonal. In most analyses, k is set to three; however, we also report some results for social networks that are denser and more symmetric (i.e., more socially segregated). A link from alter to ego is assumed to be constant over time and to have the same effect for every pair.

We also examine a few conditions where the measure of social proximity is continuous. Like previous researchers,[6] we find it useful to treat social proximity as a nonnegative and bounded quantity. We construct a measure of proximity with these characteristics in a two-step procedure. First, we draw a value u_i from a (0,1) uniform distribution for each population member. Then the social proximity z_{ij} of a representative pair of individuals i and j is defined as $z_{ij} = |u_i - u_j|$. This implies that the probability density of z is $f(z) = 2(1 - z)$.

Parameters in the heterogeneous diffusion model are estimated by the method of maximum likelihood. We maximize the logarithm of the likelihood, rather than the likelihood itself. Under right censoring, this is (Tuma and Hannan 1984, p. 126)

$$\log \mathcal{L} = \sum_{n \in \mathcal{N}(t)} d_n \log r_n(t) + \log G_n(t|t_0), \qquad (5)$$

where d_n is an indicator variable equalling one if the case adopts and $G_n(t|t_0)$ gives the probability that n has not adopted by time t for a process starting at time t_0, and $r_n(t)$ is given by equation (1).

To summarize the distribution of estimates across the 50 trials, we report the mean and standard deviation of the estimated parameters, as well as the average estimated standard error of each parameter. The last comes from the variance-covariance matrix of the maximum likelihood estimates of the parameters. A comparison of the reported standard deviation across trials and the average estimated standard error indicates whether the variability of the estimated coefficient is correctly estimated (if the ratio is approximately unity). We also report the percentage of trials in which one would reject the null hypothesis that the parameter is zero (at the .10 significance level for a two-sided test) using the estimated parameter and estimated standard error in the usual hypothesis-testing procedure. Additional characteristics of the event histories and the estimation process are given when of interest.

[6]For example, Burt (1987, pp. 1295–96, 1330–31) calculates structural equivalence weights as a normalized power transformation of Euclidean distances obtained from an incidence matrix.

As is well known, maximum likelihood estimators have good properties in large samples of independent and identically distributed cases. Under quite general regularity conditions, ML estimators are asymptotically normal, unbiased, and consistent. Tuma and Hannan (1984, ch. 5) demonstrate that these large sample properties translate well in exponentially distributed event histories of moderate size (i.e., a few hundred cases). However, when analyzing a diffusion process, event times are (hopefully) identically distributed but are not statistically independent—indeed, the dependence is often the primary object of study. The quality of maximum likelihood estimation in this situation is unknown, even asymptotically.

We caution that simulation studies of estimator quality (not only ours but those by other investigators) are only suggestive. It is always the case that features of the process not explicitly studied might produce qualitatively different patterns of results. For this reason, analytical results are preferable. However, analytical results for ML estimation generally hold asymptotically, providing unclear guidance for diffusion research on modest-sized populations. And analytical results for heterogeneous diffusion models, even in the asymptotic case, are not readily available. Bartholomew (1982) notes the difficulty of obtaining closed form expressions for simple homogeneous diffusion models. The properties of heterogeneous diffusion processes surely involve a considerable increase in complexity and mathematical difficulty. Results of simulation studies in such situations can be highly informative, even if not definitive.

3. GRAPHICAL EXAMINATION OF DIFFUSION PROCESSES

We begin by examining graphs of the hazard rate and the integrated (or cumulative) hazard rate versus time for various diffusion processes. We use the estimator of the integrated hazard rate proposed by Nelson (1972), which Aalen (1978) proved is unbiased and has minimum variance. The slope of the integrated hazard rate gives a nonparametric estimate of the population-level hazard rate, which

makes visual inspection of the overall time path of the population-level diffusion process relatively easy. We also plot the corresponding estimates of the hazard rate, smoothed to make the overall pattern clearer.[7]

We begin with an arbitrarily chosen homogeneous diffusion process—the simplest interesting version of (1). This model includes only intercept terms for the propensity to adopt and for contagion. The hazard rate $r_n(t)$ is set to equal $\exp(-6) + \Sigma_{s \in \mathcal{S}}(t) \exp(-8) = \exp(-6) + \exp(-8)S(t) = .0025 + .0003S(t)$, where in this study the maximum value of $S(t)$ is 99 (i.e., all members of the population but one have adopted). Figure 1 gives the integrated hazard rate of one realization of this process and the corresponding estimates of the hazard rate.

In Figure 1 the integrated hazard rate tends to curve upward; correspondingly, the hazard rate tends to increase with time. This pattern occurs because each prior event adds $\exp(-8) = .0003$ to the hazard rate according to the model. As events occur, the hazard rate for cases that have not yet adopted rises, stimulating even faster adoption by the cases still at risk. A graph of the cumulative number of events versus t (not shown) yields the familiar S-shaped curve usually associated with diffusion processes.

It is worth noting, however, that *empirical* estimates of the hazard rate do not increase *monotonically,* even though the hazard rate assumed by the model does increase monotonically. Different trials yield plots with different appearances. The occasional dips in the hazard rate plot (corresponding to flat segments of the integrated hazard plot) result from *stochastic* variation.

The second condition we consider (see Figure 2) is governed by a diffusion process that incorporates several kinds of individual heterogeneity. The model is $r_n(t) = \exp(-6+5x_n) + \Sigma_{s \in \mathcal{S}}(t) \exp(-8 + 2v_n + 2w_s + 4z_{ns})$. In contrast to the process displayed in Figure 1, this process does not yield a population-level hazard rate that rises over time. In fact, it falls sharply early in the process, declining to a very low level.

The declining hazard rate in Figure 2 arises from population heterogeneity. The population-level hazard rate declines because

[7]We use the variable-span smoothing algorithm developed by Friedman (1984).

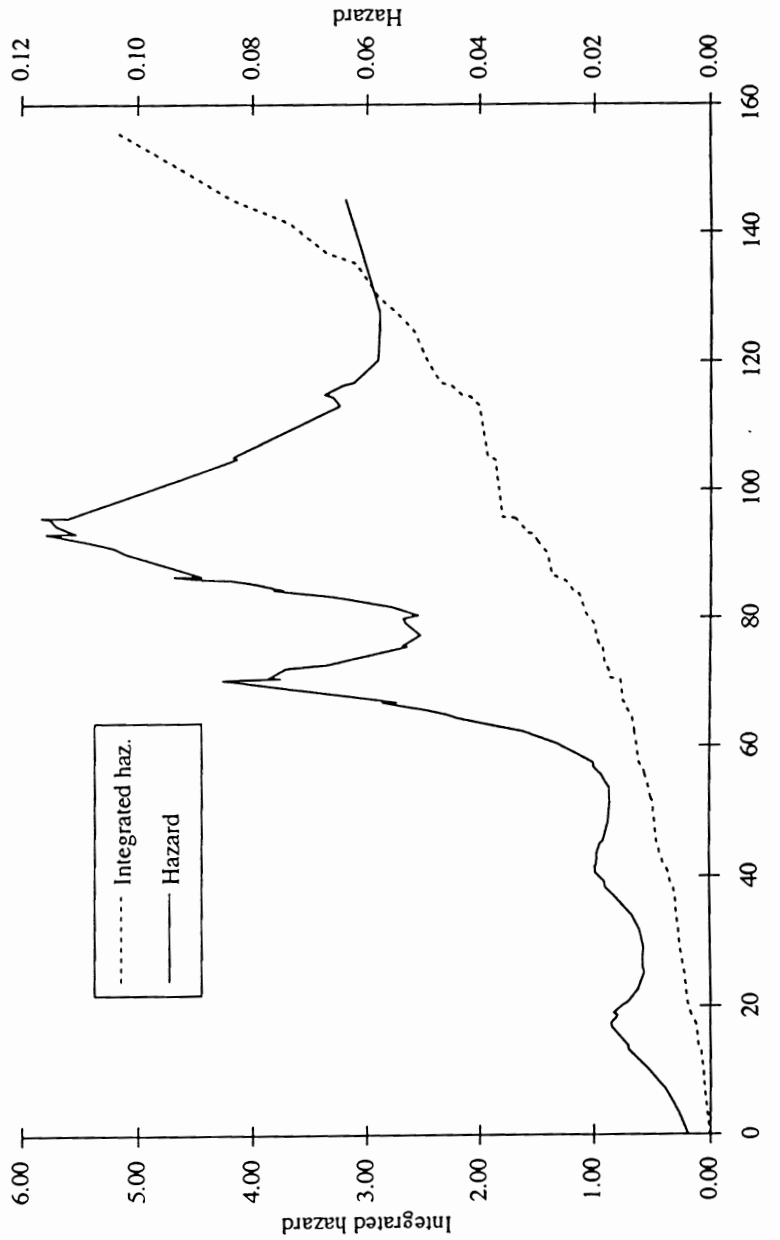

FIGURE 1. Homogeneous diffusion, plots of integrated hazard rate and hazard rate.

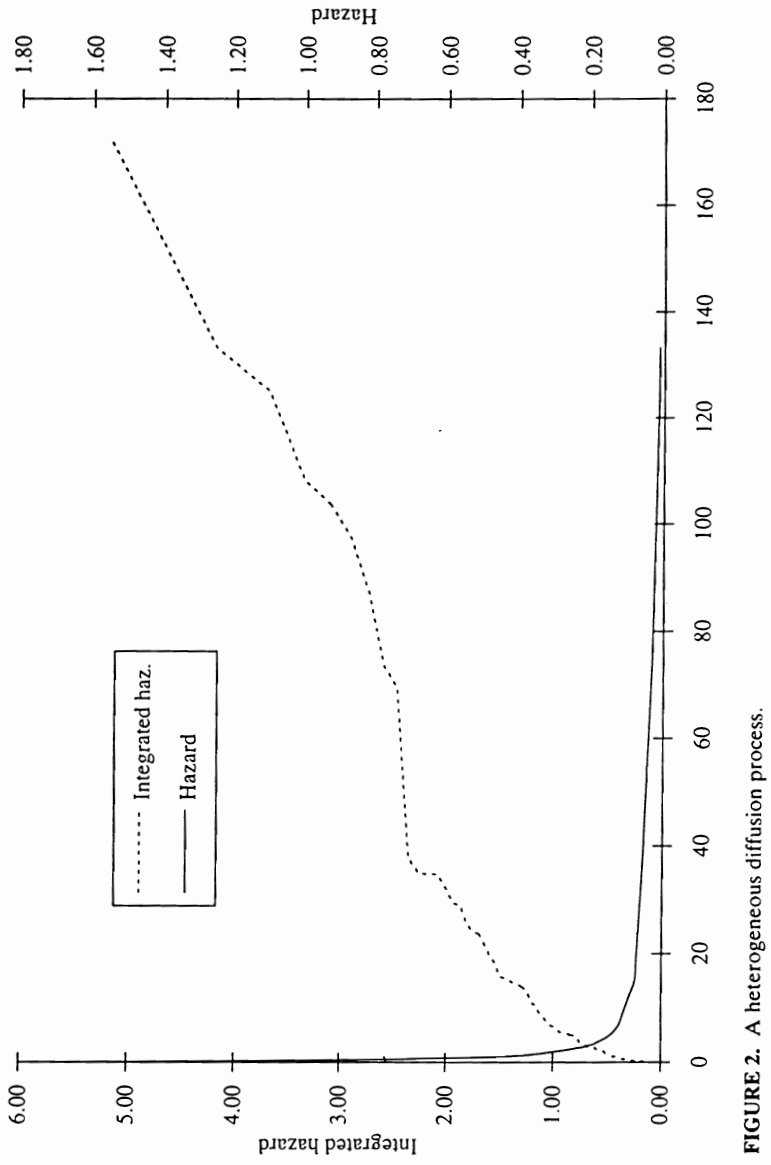

FIGURE 2. A heterogeneous diffusion process.

cases with high propensities adopt relatively early. Since the estimates graphed in Figure 2 do not control for population heterogeneity, the changing composition of the risk set produces a misleading picture of the true temporal structure of the hazard rate for an individual case in the population. Although the hazard rate of each case still at risk is monotonically increasing with time (as intrapopulation influences grow), change in the composition of the population at risk leads the population-level hazard rate to decline over time.

Of the four classes of covariate effects considered in these models, two tend to generate negative time dependence in the population-level hazard rate. The two are heterogeneity in the propensity to adopt (x_n above), and heterogeneity in susceptibility to influence from others (ν_n above). When propensities to adopt are large, cases with high rates experience the event quickly before contagion effects have much overall impact. Similarly, as effects of social contagion accumulate, highly susceptible cases adopt quickly, leaving an increasingly immune population at risk.[8]

Heterogeneity in infectiousness does not produce negative time dependence because w_s refers to the prior adopter rather than the potential adopter. It produces variability in the expansion of contagious influence over time since adopters are differentially infectious and infectiousness is unrelated to the time of adoption. But this form of heterogeneity does not affect the composition of cases at risk. Rather, the overall impact of heterogeneity in infectiousness is to accelerate the diffusion process, *ceteris paribus*.

For example, almost the same model was used to generate Figures 2 and 3; the difference is that in Figure 3 the infectiousness parameter γ equals 6 rather than 2. Consequently, this process is numerically dominated by variation in infectiousness, producing very rapid occurrence of events.

It may seem surprising that variation in infectiousness should have a net effect on the overall speed with which events occur, especially when (as is true here) the covariate is symmetrically distributed around zero. This sort of acceleration occurs because the functional form of the hazard rate is a sum of exponential terms. Since the

[8]These are both cases of the general result that heterogeneity in hazard rates produces observed negative time dependence at the population level, discussed under the rubric of unobserved heterogeneity in the event-history literature (e.g., see Tuma and Hannan 1984, pp. 174–79).

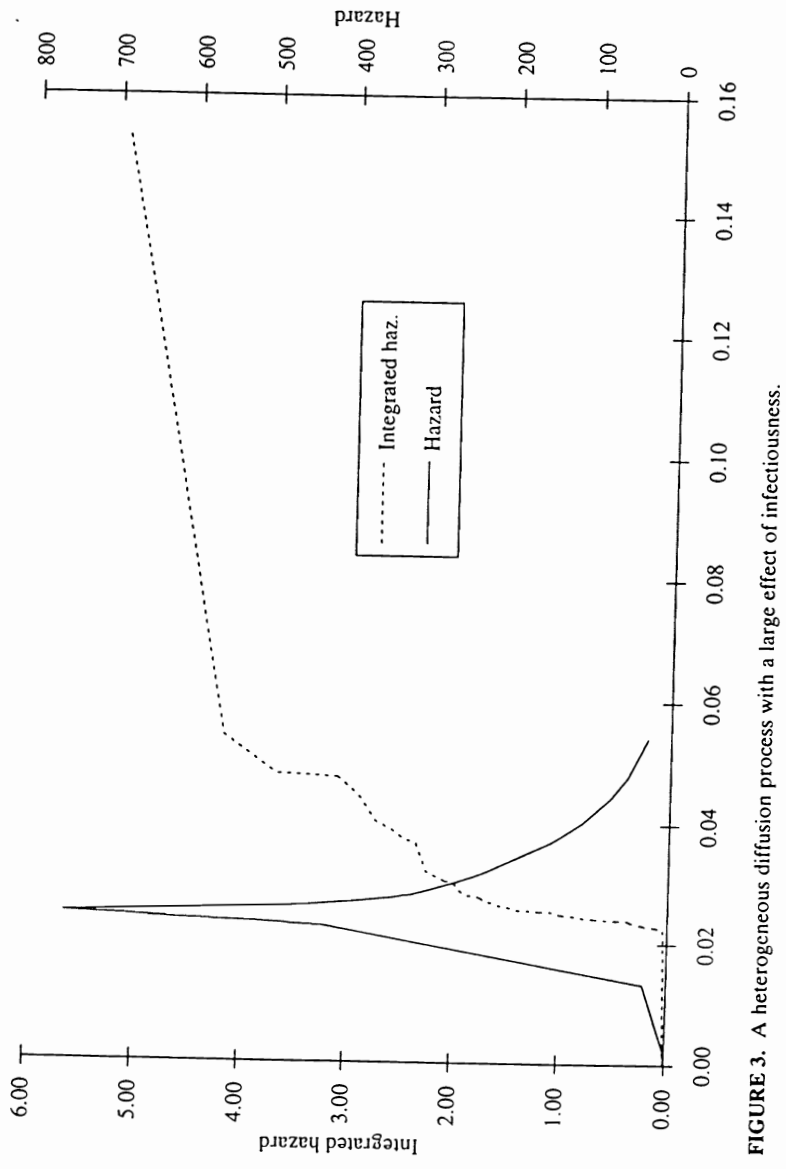

FIGURE 3. A heterogeneous diffusion process with a large effect of infectiousness.

exponential function is everywhere convex, the hazard rate increases as the parameter increases in value. This is easily seen when one considers that $\exp(\alpha + \beta) + \exp(\alpha - \beta) = \exp(\alpha)[\exp(\beta) + \exp(-\beta)] > 2\exp(\alpha)$ for any choice of α and nonzero β. Positive w_s's increase the impact of contagion more than negative w_s's decrease it.

Whether population-level time paths for the hazard rate of events appear to accelerate or decline with time is thus a function of the relative magnitude of different effects. When the contagion intercept or infectiousness dominates the process, the adoption rate of every case rises so steeply over time that the population-level hazard rate rises with time. Where heterogeneity in propensity and susceptibility dominate, cases with exceptionally high hazard rates experience events so quickly that the population-level hazard rate declines with time.

Effects of social proximity on the time path in the population-level hazard rate are more complex and less easily summarized. Since they are a function of both the prior and potential adopter, they do not unambiguously give rise to either positive or negative time dependence. Of broader interest, perhaps, is the pattern produced by proximity in a network where members of cliques are close (connected) to one another and far (disconnected) from members of other cliques. This sort of structure tends to produce a time path in the population-level hazard rate marked by waves, where each wave consists of relatively rapid diffusion within one clique.

This pattern is illustrated in Figure 4, which displays the integrated hazard rate and hazard rate of a population where z_{ns} takes the form of a partition of the population into two cliques of size 50. The hazard is $r_n(t) = \exp(-6) + \Sigma_{s\epsilon\mathcal{S}^t}(t)\exp(-8 + 10z_{ns})$. The population-level time path is made up of two waves, each corresponding to the acceleration of contagion within one of the two cliques. All members of the first clique adopt between roughly $t = 3$ and $t = 9$, while those in the second begin to adopt only after $t = 17$. (The hazard rate for the first wave appears lower than the second because members of the second clique are at risk, but are not adopting, during the period of the first wave.) Although all cases are governed by the same model with the same parameters, within-group contagion reinforces stochastic variations in outcomes.

Figures 1–4 suggest that heterogeneous diffusion processes can generate *any* sort of time path in the population-level hazard

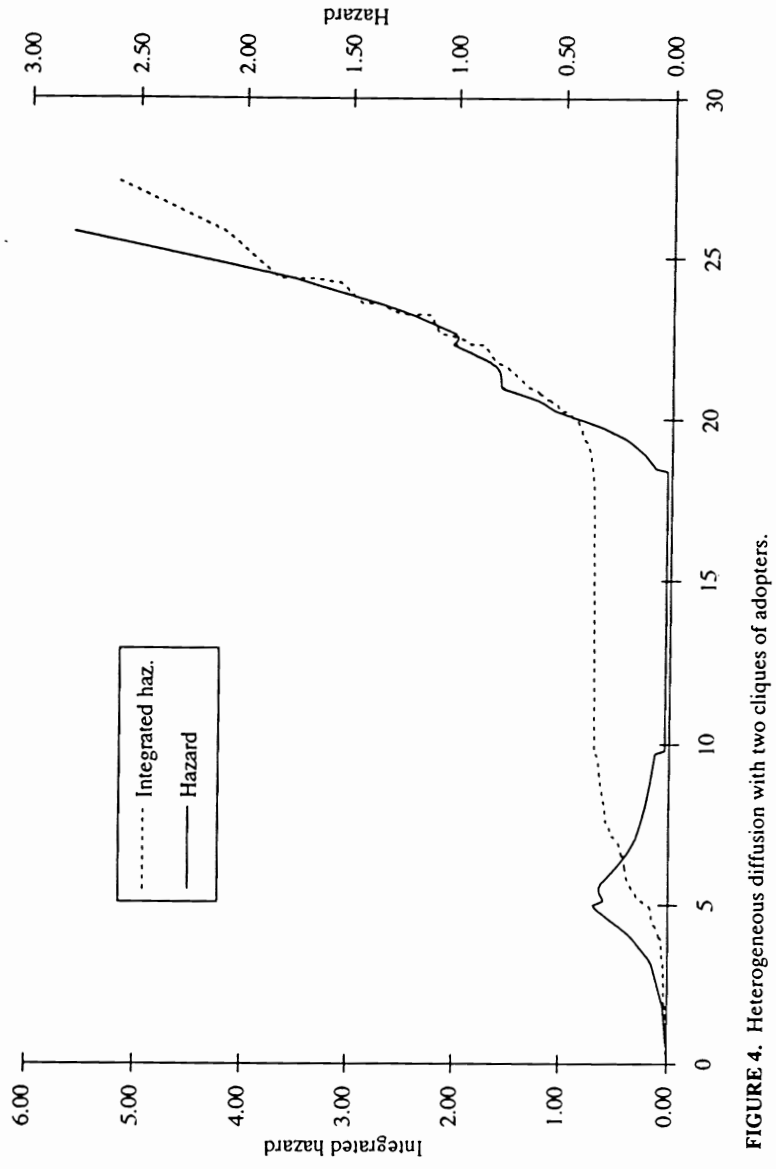

FIGURE 4. Heterogeneous diffusion with two cliques of adopters.

rate. Fully homogeneous diffusion processes do exhibit a rising population-level hazard rate (albeit one with stochastic variation), but the introduction of heterogeneity complicates matters considerably. Heterogeneity in the propensity to adopt and in susceptibility to contagion produce apparent negative time dependence in the population-level hazard rate. Heterogeneity in infectiousness leads to acceleration in the population-level hazard rate. Most interesting of all, social segregation within the population (i.e., cliques) produces population-level hazard rates that vary nonmonotonically over time.

Thus examination of the population-level time path of the integrated hazard rate or the hazard rate is of modest utility in analyzing heterogeneous diffusion processes. Simple descriptive tools such as plotting the hazard rate versus time may be useful in suggesting major features of the process. For example, a wave-like pattern in the hazard rate suggests that, *if contagion is present,* it is probably dominated by a segregated social network structure. But only when the process is fully homogeneous does the time path of the hazard rate at the population level directly reflect the generating model. In general, graphical analysis cannot be used to guide intuition about whether a social process involves intrapopulation influence. Only substantive argument or theory can do this, and only explicit models incorporating individual sources of heterogeneity are of real use in evaluating these conjectures.

4. ESTIMATION OF HETEROGENEOUS DIFFUSION MODELS

We now move from graphical analysis to model estimation. We begin with the parameter set examined in Strang and Tuma (1993), which serves as a convenient starting point for exploring a larger parameter space later. Panel 1 of Table 1 summarizes results from correctly specified analyses of diffusion in a population of size 100.

As in Strang and Tuma (1993), point estimates are unbiased and exhibit rather low variance. For example, the standard deviation of estimates of the contagion intercept across the 50 trials is 0.68, while the average estimated standard error across these trials is 0.70. The similar magnitudes of these two measures of variability in parameter estimates indicates that the standard error computed within trials agrees well with the empirical fluctuation of estimated parame-

TABLE 1
Estimation of Diffusion and Nondiffusion Models

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 1: Heterogeneous Diffusion Model					
Propensity intercept	−6.0	−6.3	0.83	0.75	100
Propensity covariate	5.0	5.2	0.61	0.52	100
Contagion intercept	−8.0	−8.3	0.68	0.70	100
Susceptibility	2.0	2.0	0.15	0.13	100
Infectiousness	2.0	2.1	0.50	0.44	100
Social proximity	4.0	4.1	0.39	0.41	100
Panel 2: Exponential Model					
Intercept	−3.0	−3.0	0.10	0.10	100
Covariate	2.0	2.0	0.09	0.10	100
Panel 3: Gompertz Model					
Intercept	−2.0	−2.0	0.12	0.12	100
Covariate	2.0	2.1	0.13	0.13	100
Time trend intercept	−0.01	−0.010	0.003	0.003	100
Time trend covariate	0.005	0.0057	0.004	0.003	62

Panel 4: Average Correlation of Diffusion Estimates in Panel 1

Parameter	(1)	(2)	(3)	(4)	(5)	(6)
(1) Propensity intercept	1.00	−.93	−.09	.13	.02	.06
(2) Propensity covariate		1.00	.07	−.13	−.02	−.06
(3) Contagion intercept			1.00	−.10	−.89	−.29
(4) Susceptibility				1.00	.04	.16
(5) Infectiousness					1.00	−.02
(6) Social proximity						1.00

ters across trials. (When estimator quality is poor, we usually see high standard deviations across estimates and even higher average standard errors.) The null hypothesis can be rejected for all covariates in all trials.

Panel 4 of Table 1 reports the average correlation of estimated parameters for the heterogeneous diffusion model. Most correlations are close to zero on average, indicating that change in one parameter has little impact on the estimates of the other parameters. In particular, correlations between parameters in the propensity component and those in the contagion component are very low, indicating the separability of these two classes of effects when a heterogeneous diffusion process generates the data. (The magnitude of these correlations would presumably rise if the independent variables were constructed to covary in some systematic way; see Section 4.5 for some discussion.) This separability provides the main estimation advantage of the additive approach to diffusion modeling pursued in this paper.

Several pairs of estimated parameters are strongly related. Parameters for the propensity intercept and covariate, as well as for the contagion intercept and the infectiousness covariate, have a large negative correlation on average ($\approx -.9$). Estimated parameters for the contagion intercept and social proximity are also negatively correlated, but less strongly. All of these correlations follow from the functional form of the model as a sum of exponentials, as discussed for infectiousness above. As we see below, estimation problems can arise from large correlations among estimated parameters.

It is helpful to put the estimation of a heterogeneous diffusion model within the context of more familiar models. Here we briefly compare the above simulation to corresponding analyses of exponential and Gompertz models. When there is no source of time dependence other than that produced by measured covariates, the exponential model is a common baseline model in event-history analysis. In the present context, it is also the special case of equation (1) in which there is only a propensity to adopt and no contagion: $r_n(t) = \exp(\boldsymbol{\alpha}'\mathbf{x}_n)$. The Gompertz model is $r_n(t) = \exp(\boldsymbol{\alpha}'\mathbf{x}_n + \boldsymbol{\beta}'\mathbf{u}_n t)$, where $\mathbf{x}_n$ and $\mathbf{u}_n$ are in general different covariates. It provides the analyst with a ready way of capturing monotonic shifts in the hazard over time and might be used as an alternative way of modeling simple diffusion processes, such as that displayed in Figure 1.

Panels 2 and 3 give results for correctly specified exponential and Gompertz models. As with the heterogeneous diffusion model, all parameters appear estimated without bias. Results for exponentially distributed event histories show very little variance in parameter estimates. The quality of estimates for Gompertz models is high overall, but with some imprecision for the covariate that interacts with time. The mean estimate for this parameter is off by more than 10 percent, the average standard error is large, and the null hypothesis is correctly rejected in only two-thirds of the trials.

Overall, Table 1 suggests that estimator quality for different event-history models is roughly similar. Efficiency declines somewhat as models and processes become more complex. Exponential models of exponentially distributed event times are estimated well because both models and data are simple. Gompertz models can be estimated well overall but have some difficulty estimating interactions with time. Results for the heterogeneous diffusion models show larger variance in estimated effects than exponential models but seem able to parse highly complex interdependencies within populations.

4.1. *Parameter Space Sensitivity*

To explore sensitivity of estimator quality to parameter values, we first consider whether strong effects of one type impede the estimation of other types. For example, can contagion effects be detected well in diffusion processes dominated by heterogeneity in propensities to adopt? Is infectiousness estimated accurately when individuals vary substantially in their susceptibility to contagion? To examine issues like these, we work from the set of parameters estimated in Table 1 and sequentially increase each parameter by 4, which multiplies the effect by about 55 (given the exponential functional form). Table 2 summarizes the results.

Panels 1 and 2 indicate that contagion effects can be estimated well when propensities to adopt (i.e., effects existing independent of intrapopulation linkages) dominate. Panels 4 and 6 reveal little sensitivity to strong variation in susceptibility to contagion and patterns of social proximity. In these conditions, parameter estimates appear unbiased and standard errors are rather small.

A large contagion intercept and substantial heterogeneity in infectiousness (Panels 3 and 5) make it difficult to estimate propensities to adopt. In both conditions, the effects of contagion accelerate so rapidly that estimates of the propensity to adopt are based largely on information about the timing of the first few events, after which contagion effects dominate. What is perhaps most remarkable here, however, is that the parameters in the contagion term continue to be accurately estimated even when the process occurs "in the wink of an eye." (Compare estimator quality with quartile adoption times in Panel 8.)[9]

[9]It is important to note that there is no time aggregation in temporal measurements here, which contrasts with the situation in much empirical research.

TABLE 2
Dominating Effect in Full Model

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 1: Large Propensity Intercept					
Propensity intercept	−2.0	−2.0	0.30	0.30	100
Propensity covariate	5.0	5.1	0.32	0.28	100
Contagion intercept	−8.0	−8.4	1.25	0.96	100
Susceptibility	2.0	2.0	0.18	0.17	100
Infectiousness	2.0	2.2	0.74	0.58	96
Social proximity	4.0	4.1	0.50	0.61	100
Panel 2: Large Effect of Propensity					
Propensity intercept	−6.0	−6.0	0.53	0.47	100
Propensity covariate	9.0	9.1	0.43	0.37	100
Contagion intercept	−8.0	−8.3	1.69	0.87	100
Susceptibility	2.0	2.0	0.14	0.14	100
Infectiousness	2.0	2.2	1.07	0.58	98
Social proximity	4.0	4.1	0.49	0.47	100
Panel 3: Large Contagion Intercept					
Propensity intercept	−6.0	−7.9	6.81	9.81	92
Propensity covariate	5.0	6.0	3.22	4.31	98
Contagion intercept	−4.0	−4.2	0.56	0.51	100
Susceptibility	2.0	2.0	0.14	0.12	100
Infectiousness	2.0	2.1	0.36	0.37	100
Social Proximity	4.0	4.1	0.32	0.35	100
Panel 4: Large Effect of Susceptibility					
Propensity intercept	−6.0	−6.0	0.31	0.32	100
Propensity covariate	5.0	5.0	0.24	0.27	100
Contagion intercept	−8.0	−8.2	0.62	0.58	100
Susceptibility	6.0	6.1	0.19	0.19	100
Infectiousness	2.0	2.1	0.42	0.36	100
Social proximity	4.0	4.1	0.50	0.42	100
Panel 5: Large Effect of Infectiousness					
Propensity intercept	−6.0	−7.9	6.50	5.91	72
Propensity covariate	5.0	6.0	3.26	2.95	94
Contagion intercept	−8.0	−7.9	1.51	1.03	100
Susceptibility	2.0	2.0	0.17	0.13	100
Infectiousness	6.0	6.0	0.65	0.51	100
Social proximity	4.0	3.9	0.62	0.43	100

TABLE 2 (contd.)

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 6(a): Large Effect of Proximity					
Propensity intercept	−6.0	−6.4	2.46	1.40	100
Propensity covariate	5.0	5.3	1.38	0.83	100
Contagion intercept	−8.0	−8.0	0.43	1.14	100
Susceptibility	2.0	2.0	0.11	0.12	100
Infectiousness	2.0	2.0	0.17	0.17	100
Social proximity	8.0	7.9	0.44	0.55	100
Panel 6(b): Very Large Effect of Proximity					
Propensity intercept	−6.0	−6.8	4.67	2.15	100
Propensity covariate	5.0	5.6	2.53	1.14	100
Contagion intercept	−8.0	−10.5	3.74	55.40	62
Susceptibility	2.0	2.0	0.09	0.11	100
Infectiousness	2.0	2.0	0.15	0.14	100
Social proximity	10.0	12.5	3.79	55.41	62
Panel 6(c): Very Large Effect of Proximity, Summed over Proximate Cases Only					
Propensity intercept	−6.0	−3.6	1.96	1.21	92
Propensity covariate	5.0	3.6	1.27	0.75	98
Contagion intercept	−8.0	—	—	—	—
Susceptibility	2.0	2.0	0.11	0.12	100
Infectiousness	2.0	2.0	0.15	0.14	100
(Net)Social proximity	2.0	2.0	0.16	0.13	100
Panel 6(d): Very Large Effect of Proximity (Continuous Measure)					
Propensity intercept	−6.0	−6.2	2.70	2.46	90
Propensity covariate	5.0	5.2	1.45	1.25	98
Contagion intercept	−8.0	−8.4	1.27	0.83	100
Susceptibility	2.0	2.0	0.13	0.12	100
Infectiousness	2.0	2.1	0.67	0.39	100
Social proximity	10.0	10.2	0.56	0.70	100
Panel 7(a): Baseline Model (20% Network Density)					
Propensity intercept	−6.0	−5.9	1.49	1.02	100
Propensity covariate	5.0	5.0	1.45	0.64	100
Contagion intercept	−8.0	−9.7	3.18	20.26	74
Susceptibility	2.0	2.0	0.12	0.13	100
Infectiousness	2.0	2.1	0.42	0.35	100
Social proximity	4.0	5.6	3.55	20.40	68

TABLE 2 (contd.)

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 7(b): Baseline Model (20% Network Density) Summed over Proximate Cases Only					
Propensity intercept	−6.0	−5.9	1.30	1.03	100
Propensity covariate	5.0	5.0	0.83	0.66	100
Contagion intercept	−8.0	—	—	—	—
Susceptibility	2.0	2.0	0.13	0.12	100
Infectiousness	2.0	1.7	0.24	0.24	100
(Net) Social proximity	−4.0	−3.6	0.24	0.28	100

Panel 8: Quartile Adoption Times

	Condition								
Percentage	1	2	3	4	5	6*a*	6*bc*	6*d*	7*ab*
25	0.08	0.14	0.06	0.11	0.08	0.21	0.19	0.03	1.02
50	0.70	1.44	0.12	1.40	0.08	0.46	0.48	0.06	3.68
75	3.31	6.62	0.29	28.36	0.10	1.14	1.16	0.18	12.03
100	418.7	451.9	30.47	10^7	0.93	54.22	119.9	16.11	686.8

These results help index how sensitive estimation is to big sources of variation in specific components of the model. Naturally, they are not an exhaustive exploration of such effects. While we see rather small shifts in estimator quality corresponding to some large shifts in parameter values, we suspect that the effects of any variable can become so strong (or so weak) that estimator quality is degraded. Further simulations suggested one such case of interest—estimation difficulty when social proximity effects become really large. As shown in Panel 6(b), estimation deteriorates markedly when the social proximity parameter is raised to 10. Estimated coefficients for both the contagion and the proximity effect are off target: the former is too small, the latter too large. In addition, neither parameter is reliably estimated, with (correct) rejection of the null hypothesis in only two-thirds of the trials.

These problems arise because the very large effects of contagion among socially proximate actors drown out all other sources of intrapopulation influence. Recall that in these simulations social prox-

imity takes the form of a linkage to three alters (i.e., three contacts) in a social network. Hence, in Panels 6(b) and 6(c), the strength of contagion within the network is 22026 [= exp(−8 + 10)/exp(−8) = exp(10)] times larger than contagion outside the network. Contagion effects reduce in practice to contagion among those directly linked in the network. Estimation suffers because any model in which $\exp(\hat{\beta}_0)$ is small and $\hat{\beta}_0 + \hat{\delta} \approx 2(= -8 + 10)$ fits the data about as well.

When a single binary social proximity measure dominates the process in this way, we find that it is best to allow contagion only from directly connected members of the population (here, for those for whom $z_{ns} = 1$). The new, reduced model is

$$r_n(t) = \exp(\boldsymbol{\alpha}'\mathbf{x}_n) + \sum_{s \in \mathcal{G}(t) \cap (z_{ns}=1)} \exp(\boldsymbol{\beta}'\mathbf{v}_n + \boldsymbol{\gamma}'\mathbf{w}_s + \delta \cdot 1)$$

$$= \exp(\boldsymbol{\alpha}'\mathbf{x}_n) + \exp(\boldsymbol{\beta}'\mathbf{v}_n + \delta) \sum_{s \in \mathcal{G}(t) \cap (z_{ns}=1)} \exp(\boldsymbol{\gamma}'\mathbf{w}_s), \qquad (6)$$

where the contagion term is summed only over the prior adopters linked to n. Then the contagion intercept β_0 and the coefficient of z_{ns}, δ, cannot be distinguished; hence the quantity estimated in Panel 6(c) is $\beta_0 + \delta$, which has a "true" value of $2(= -8 + 10)$. As Panel 6(c) indicates, this composite network effect is estimated without bias and with high efficiency. Further, the effects of susceptibility and infectiousness covariates are also estimated accurately. This is a surprising and useful result since estimators of these effects are based on at most three influential events per potential adopter.

The above treatment of social proximity focuses solely on how coefficient magnitudes affect estimation. In fact, it is reasonable to assume that the results are also contingent on how fine-grained information is, and on characteristics of network structure as well. To investigate the first issue, we employ the continuous proximity measure described above, where each pair of cases draws a value from the distribution $f(z) = 2(1 - z)$. Panel 6(d) indicates that estimator quality is high even when we reexamine the parameter combination that produced problems in Panel 6(b). Both Panels 6(b) and 6(d) have correctly specified models, but the more fine-grained process in Panel 6(d) is estimated better.

Difficulty in estimating social proximity effects also arises in dense networks. We found it difficult to capture effects of social proximity accurately when network density (the ratio of actual to possible ties) rises to 20 percent or more. Panel 7(a) demonstrates the problem: Estimates of the contagion intercept and the proximity effect again show high variance. When network density is high, weak effects of proximity are not easily distinguished from the contagion intercept, while strong effects of proximity hide effects of contagion in the absence of direct ties. It again becomes useful to restrict contagion to socially proximate alters, as Panel 7(b) shows.

Estimating a model in which only proximate cases are allowed to be contagious is strictly speaking a misspecification of the model because nonproximate cases also have tiny contagious influences. For the conditions used to generate the data analyzed in Panels 6(b)–(c) and 7(a)–(b), this model misspecification has few adverse impacts. Indeed, simplifying the model in this way substantially improves the quality of estimates, most importantly by reliably detecting the presence of the social proximity effect. This is an encouraging result, which may also prove useful in research settings where all influences are not reliably observed or where network contacts are sampled rather than exhaustively enumerated. Further work on diffusion processes may suggest other contexts in which particular strategies for simplifying models are effective.

4.2. *Specification Error: Models with Extraneous Variables*

Our study of the impact of model misspecification begins with overspecification—models that include *extraneous* variables, ones not actually used to generate the simulated data. Standard statistical theory implies that the estimated coefficients of extraneous variables in linear models should have an expected value of zero: Including them in the estimated model tends to increase the standard errors of the other parameters (mainly due to correlations between regressors) but not bias parameter estimates.

We use the parameter set in Table 1 as a baseline. In each condition, we set a particular combination of the parameters to zero in the model used to generate the data. Nonzero intercepts (for both propensity and contagion) are retained throughout. To simplify presentation, we report only the percentage of trials where the null hypothesis of no effect is rejected at the .10 significance level. The null hypothesis should be (incorrectly) rejected for extraneous covariates in 10 percent of the trials; it should be (correctly) rejected for

generating covariates in a large but unknown percentage of the trials. (The exact percentage depends on the power of the test, which is difficult to calculate *a priori.*)

Panel 1 of Table 3 indicates that including extraneous variables has little impact on the estimation of the effects of the covariates actually used to generate the event histories. The null hypothesis is rejected in over 90 percent of the trials for 25 of 27 effects. Moreover, coefficients of generating covariates are estimated without any evidence of bias or inefficiency relative to those shown in Table 1 (results not reported for brevity). However, infectiousness effects are underestimated in two conditions, yielding false negatives (Type II errors) in about a fourth of the trials.

Extraneous variables with putative propensity and susceptibility effects are accurately estimated as having no influence on the process. The percentage of false positives hovers around 10 percent and never exceeds 16 percent. But detecting zero effects is less successful in the case of infectiousness and social proximity. In many conditions, false positives arise two or three times as often as the significance level of the hypothesis test suggests. Examination of the actual parameter estimates reveals that estimates of effects of extraneous social proximity measures are particularly poor, with standard deviations usually twice as large as mean estimates. Average standard errors are an order of magnitude larger than standard deviations.

Once again, the problem involves the difficulty of separating the effect of infectiousness from the contagion intercept. Since infectiousness accelerates the population-level hazard rate of adoption rather than the hazard rate at which specific cases adopt, it is sometimes confused with the contagion intercept. The same problem can arise with social proximity effects, as the previous section highlights.

One approach to dealing with this problem is to employ conservative significance levels when evaluating effects of infectiousness and social proximity. In addition to this statistical approach to the problem, our results suggest two modeling considerations. First, Table 3 indicates that estimation difficulties are largest when the generating process is actually homogeneous. The percentages of false positives for infectiousness and social proximity decline when true heterogeneity in the diffusion process exists and is incorporated in the model. In particular, it appears easier to rule out an extraneous social proximity effect when variation in infectiousness is present, and vice versa. In

TABLE 3
Model with Extraneous Variables

	Estimated Variable			
	Propensity (%)	Susceptibility (%)	Infectiousness (%)	Proximity (%)
Panel 1: Binary Ties				
Intercepts only	10	8	24	22
Propensity	**100**	4	24	30
Susceptibility	2	**98**	24	12
Infectiousness	10	4	**90**	8
Social proximity	16	16	12	**100**
Propensity, susceptibility	**100**	**100**	30	24
Propensity, infectiousness	**100**	16	**78**	22
Propensity, proximity	**94**	8	8	**92**
Susceptibility, infectiousness	12	**100**	**98**	16
Susceptibility, proximity	6	**100**	14	**98**
Infectiousness, proximity	8	6	**100**	**100**
No propensity	4	**100**	**100**	**100**
No susceptibility	**100**	16	**100**	**100**
No infectiousness	**100**	**100**	14	**94**
No proximity	**98**	**98**	**72**	20
Panel 2: Continuous Proximity				
Intercepts only	8	12	20	8
Propensity	**100**	16	26	18
Susceptibility	12	**100**	30	12
Contagiousness	10	4	**90**	8
Social proximity	12	10	14	**100**
Propensity, susceptibility	**100**	**100**	34	22
Propensity, infectiousness	**100**	4	**74**	18
Propensity, proximity	**100**	20	14	**100**
Susceptibility, infectiousness	8	**100**	**92**	18
Susceptibility, proximity	6	**100**	8	**100**
Infectiousness, proximity	4	20	**98**	**100**
No propensity	4	**100**	**98**	**100**
No susceptibility	**100**	14	**92**	**100**
No infectiousness	**100**	**100**	22	**98**
No proximity	**100**	**100**	**74**	8

Note: Entries are the percentage of rejections of the null hypothesis that a parameter is zero under a two-tailed 0.10 test. Incorrect rejection of the null hypothesis should occur in 10 percent of the trials. Results for effects are in boldface.

conditions with heterogeneity in social proximity, percentages of false positives for an extraneous infectiousness effect are 12, 8, 14, and 14 percent; where social proximity is homogeneous, the corresponding quantities are 24, 24, 24, and 30 percent.

Second, extraneous social proximity effects can be ruled out better when measures are more fine grained. Panel 2 repeats the simulation studies in Panel 1 except that binary proximity measures are replaced by continuously-distributed proximity measures. Here false positives occur at more appropriate levels. For example, the null hypothesis for the effect of a continuously-distributed proximity measure is falsely rejected in 8 percent of trials when no other covariates are included in the model and an average of 13 percent of the time across all conditions.

4.3. *Specification Error: Models with Omitted Variables*

We now turn to the estimation of models that omit covariates actually used in generating the event-history data. Omitting covariates with nonzero effects from linear models leads to biased parameter estimates and diminishes standard errors if the omitted covariates are correlated with included covariates. Our simulation procedure neither builds in correlations between covariates nor ensures that they are zero in a given trial. The correlation between omitted and included covariates approaches zero on average as the number of trials increases, but a nonzero correlation exists in each trial due to sampling variability. Reasoning by analogy with linear models, the correlation between omitted and included variables may translate across many trials into an increased standard deviation of estimated parameters, a lower average standard error of estimate, and (hopefully) no bias in the mean estimated parameter.

To test how the estimation performs when covariates are omitted, we studied four conditions. In the first three, we generated event histories as in our baseline model, but included one additional covariate with a standard Gaussian distribution. This variable affected the propensity, susceptibility, or infectiousness terms, respectively, with a coefficient of 2.0. In the fourth condition, each case was made socially proximate to five members of the population (rather than three). Estimated models omitted these additional covariates (in the fourth condition, the two additional ties were ignored). Table 4 summarizes results.

TABLE 4
Model with Omitted Variables

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 1: Omitted Propensity					
Propensity intercept	−6.0	−6.5	5.25	0.87	100
Propensity covariate 1	5.0	4.6	2.58	0.55	100
Propensity covariate 2	2.0	—	—	—	—
Contagion intercept	−8.0	−8.0	0.70	0.58	100
Susceptibility	2.0	2.0	0.17	0.13	100
Infectiousness	2.0	1.9	0.45	0.39	100
Social proximity	4.0	4.0	0.43	0.43	100
Panel 2: Omitted Susceptibility					
Propensity intercept	−6.0	−4.5	0.82	0.36	100
Propensity covariate	5.0	4.0	0.71	0.34	100
Contagion intercept	−8.0	−10.9	5.34	1.19	100
Susceptibility 1	2.0	2.4	0.60	0.15	100
Susceptibility 2	2.0	—	—	—	—
Infectiousness	2.0	2.0	2.85	0.70	80
Social proximity	4.0	3.8	1.78	1.20	90
Panel 3: Omitted Infectiousness					
Propensity intercept	−6.0	−7.7	6.90	1.83	98
Propensity covariate	5.0	6.0	4.14	0.99	100
Contagion intercept	−8.0	−7.0	4.45	1.44	98
Susceptibility	2.0	2.0	0.20	0.12	100
Infectiousness 1	2.0	2.3	2.20	0.77	88
Infectiousness 2	2.0	—	—	—	—
Social proximity	4.0	3.1	0.95	0.53	96
Panel 4: Omitted Social Proximity					
Propensity intercept	−6.0	−6.3	1.09	0.87	100
Propensity covariate	5.0	5.2	0.85	0.61	100
Contagion intercept	−8.0	−7.7	1.18	0.72	100
Susceptibility	2.0	1.9	0.12	0.13	100
Infectiousness	2.0	2.0	0.62	0.46	100
Social proximity 1	4.0	3.7	0.40	0.43	100
Social proximity 2	4.0	—	—	—	—

Omitting a measure of the propensity to adopt produces some bias and considerable imprecision in the estimated propensity intercept and in the estimated effect of the propensity measure, both of which are underestimated. But estimated effects in the contagion component of the model are unbiased and have low standard errors. This is an important result because researchers can generally identify many more plausible sources of a propensity to adopt than they can measure. It seems that a heterogeneous diffusion model does not require comprehensive specification of propensities to yield good estimates of the effects of intrapopulation linkages and of contagion (at least when the omitted and included effects are not highly correlated). And even this conclusion probably does not hold when propensity effects are extremely large.

Omitting measures of susceptibility and infectiousness can produce imprecision in almost all components of the hazard rate, especially the contagion intercept, which exhibits a large standard deviation across trials. As expected, the contagion intercept is overestimated when infectiousness measures are omitted and underestimated when susceptibility measures are omitted. Estimates of propensity effects are also considerably affected. By contrast, misspecification of social proximity effects (where two of five alters are ignored) seems to do little to disrupt estimator quality: The contagion intercept shifts upward somewhat, but the estimated effect of social proximity seems to be unbiased. This finding offers considerable encouragement to researchers who are unsure if they can measure all linkages within a social network.[10]

4.4. *Assigning Covariates to Types of Effects*

In addition to the standard forms of misspecification discussed in the previous section (inclusion of extraneous variables and omission of generating variables), heterogeneous diffusion models are subject to a more subtle form of misspecification. Analysts may correctly identify a variable as affecting the diffusion process but not know where to locate the variable within the model. For example, a covariate that

[10]It is important to note that the sensitivity to unobserved linkages probably varies by network structure. For example, it is straightforward to show that the unmodeled influence will be greater when the same proportion of influential alters are unobserved in a denser network.

actually affects the propensity to adopt may be incorrectly postulated to affect susceptibility to contagion or the infectiousness of prior adopters.

Often prior knowledge or theory usefully guides the researcher. Covariates characterizing social relations within a population naturally appear in the contagion component. Greve (1994) suggests that variables characterizing organizational inertia should affect susceptibility because inert organizations resist pressures to innovate, despite the examples provided by others. However, plausible rationales can often be developed for several different types of effects. For example, Strang and Tuma (1993) note that network centrality may increase the propensity to adopt (central actors are often leading innovators), susceptibility to contagion (central actors receive more information than do marginal actors), and infectiousness (central actors are more influential than marginal actors). In such situations, a specification search is required to assign effects to appropriate locations within the heterogeneous diffusion model.

4.4.1. *A Serial Location Search*

We first examine estimator quality when a generating covariate is incorrectly located within the heterogeneous diffusion model. We consider the common modeling strategy where an analyst ventures a plausible specification and retains variables estimated to have significant effects. We separately examine six possible conditions: where propensity is incorrectly specified as susceptibility or infectiousness, where susceptibility is incorrectly specified as propensity or infectiousness, and where infectiousness is incorrectly specified as propensity or susceptibility. In each condition, we also include and correctly model a measure of social proximity to determine the sensitivity of other effects to misspecification.

Table 5 shows that propensity and susceptibility are easily confused. A true propensity effect is usually estimated as a somewhat larger susceptibility effect, and a true susceptibility effect is always estimated as a somewhat smaller propensity effect. Susceptibility is also mistaken for infectiousness with some regularity (in about 40 percent of the trials), but propensity is not. In all four forms of misspecification, the ability to capture the correctly specified effect of social proximity falls dramatically. Estimated effects

TABLE 5
Misspecified Model

Parameter	True Value	ML Estimate Mean	ML Estimate SD	ML Estimate Average SE	ML Estimate Reject H_0 (%)
Panel 1: Propensity Effect Estimated as Susceptibility					
Propensity intercept	−6.0	−3.5	0.48	0.20	100
Propensity covariate x_n	3.0	—	—	—	—
Contagion intercept	−8.0	−11.1	2.67	5.69	96
Susceptibility x_n	0	3.9	2.32	0.83	92
Social proximity	4.0	0.9	3.90	16.1	36
Panel 2: Propensity Effect Estimated as Infectiousness					
Propensity intercept	−6.0	−3.8	0.35	0.25	100
Propensity covariate x_n	3.0	—	—	—	—
Contagion intercept	−8.0	−13.8	6.31	53.4	34
Infectiousness x_s	0	0.1	1.71	5.24	16
Social proximity	4.0	8.6	4.00	38.1	28
Panel 3: Susceptibility Effect Estimated as Propensity					
Propensity intercept	−6.0	−3.7	0.26	0.13	100
Propensity covariate x_n	0	1.3	0.20	0.11	100
Contagion intercept	−8.0	−15.0	1.96	15.8	28
Susceptibility x_n	3.0	—	—	—	—
Social proximity	4.0	5.4	3.41	21.7	0
Panel 4: Susceptibility Effect Estimated as Infectiousness					
Propensity intercept	−6.0	−4.9	0.72	0.36	100
Contagion intercept	−8.0	−30.0	11.6	44.1	22
Susceptibility x_n	3.0	—	—	—	—
Infectiousness x_s	0	8.9	5.73	15.4	42
Social proximity	4.0	3.9	5.03	12.6	12
Panel 5: Infectiousness Effect Estimated as Propensity					
Propensity intercept	−6.0	−7.5	2.05	1.95	94
Propensity covariate x_n	0	−0.0	1.55	1.32	2
Contagion intercept	−8.0	−4.9	0.78	0.20	100
Infectiousness x_s	3.0	—	—	—	—
Social proximity	4.0	2.7	1.46	1.69	94

TABLE 5 (contd.)

Parameter	True Value	ML Estimate Mean	ML Estimate SD	ML Estimate Average SE	ML Estimate Reject H_0 (%)
Panel 6: Infectiousness Effect Estimated as Susceptibility					
Propensity intercept	−6.0	−6.4	0.87	0.96	100
Contagion intercept	−8.0	−5.0	0.90	0.21	100
Susceptibility x_n	0	0.0	0.09	0.10	6
Infectiousness x_s	3.0	—	—	—	—
Social proximity	4.0	3.0	0.64	0.52	98

Note: Both the omitted and incorrectly specified effects are denoted by x to stress that they refer to the same covariate. When the covariate characterizes variation in cases at risk, we subscript it by n; when it characterizes variation in prior adopters (spreaders), we subscript it by s.

are usually biased, and average standard errors are inflated. A similar problem occurs in the estimation of the propensity and contagion intercepts.

By contrast, a true infectiousness effect is hardly ever confused with either propensity or susceptibility. Misspecification here yields estimated parameters centered around zero and low rates of incorrect rejection of the null hypothesis. And incorrect modeling of infectiousness does not strongly disturb estimated effects of social proximity. These remain correctly estimated as positive and statistically significant, though estimated parameters show some downward bias.

These results make sense in view of the functional form of the model. The difference between propensity and susceptibility is that the latter interacts with the cumulative number of prior events, which is some monotonic function of time. The estimation procedure detects the existence of an effect of the covariate on adoption timing; however, misspecifying where its effect is located leads to poor estimates of the effect of both the misspecified variable and other variables in the model. In addition, the relative importance of the contagion and propensity components of the diffusion process is misjudged, with propensity estimated as larger than it really is (see intercept estimates in Panels 1 and 3). By contrast, infectiousness refers to characteristics of the *prior* rather than the *potential* adopter. It thus produces little

overlap with the variation produced by true propensity or susceptibility effects.

Overall, Table 5 makes it clear that an exploratory strategy of serially assigning a covariate to one location within the model and inspecting the results is not viable. False positives are as easily generated as true positives. Where effect location is not well understood on *a priori* grounds, some other strategy is needed.

4.4.2. *A Parallel Location Search*

Table 6 examines the alternative strategy of simultaneously estimating effects in several locations in the model. This approach was employed by Strang and Tuma (1993) to locate the effects of network centrality, but its effectiveness has not previously been evaluated. To assess this approach, we again generate event histories where some covariate has a propensity, susceptibility, or infectiousness effect. But we now estimate models where that covariate is assigned both its true effect and also some other effect that it does not really have. For example, Panel 1 shows event histories generated with a covariate x_n affecting the propensity to adopt and analyzed with that covariate treated as having both a propensity and a susceptibility effect.

This search strategy correctly assigns covariates to the appropriate location in the model. In particular, propensity and susceptibility effects are no longer confused. In conditions 1 and 3, false positives occur in 6 and 14 percent of the trials, respectively, a dramatic decline from the levels of the previous search strategy, where they occur in 92 and 100 percent of the trials, respectively. There continues to be little difficulty in distinguishing infectiousness

TABLE 6
Model Using Parallel Location Search

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 1: Propensity Effect Estimated as Propensity and Susceptibility					
Propensity intercept	−6.0	−6.3	1.08	0.91	100
Propensity covariate x_n	3.0	3.2	0.66	0.59	100
Contagion intercept	−8.0	−8.4	1.77	5.49	96
Susceptibility x_n	0	0.0	0.18	0.18	6
Social proximity	4.0	4.3	1.88	5.72	94

TABLE 6 (contd.)

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 2: Propensity Effect Estimated as Propensity and Infectiousness					
Propensity intercept	−6.0	−6.2	1.00	0.78	100
Propensity covariate x_n	3.0	3.1	0.72	0.53	100
Contagion intercept	−8.0	−8.1	0.48	0.51	100
Infectiousness x_s	0	0.0	0.22	0.24	8
Social proximity	4.0	4.1	0.73	0.78	100
Panel 3: Susceptibility Effect Estimated as Propensity and Susceptibility					
Propensity intercept	−6.0	−6.6	2.26	1.46	96
Propensity covariate x_n	0	0.2	1.43	0.75	14
Contagion intercept	−8.0	−8.0	0.30	0.32	100
Susceptibility x_n	3.0	3.0	0.24	0.19	100
Social proximity	4.0	3.9	0.47	0.46	100
Panel 4: Susceptibility Effect as Susceptibility and Infectiousness					
Propensity intercept	−6.0	−6.0	0.35	0.34	100
Contagion intercept	−8.0	−8.1	0.34	0.34	100
Susceptibility x_n	3.0	3.1	0.20	0.21	100
Infectiousness x_s	0	−0.1	0.29	0.26	16
Social proximity	4.0	4.1	0.40	0.45	100
Panel 5: Infectiousness Effect Estimated as Propensity and Infectiousness					
Propensity intercept	−6.0	−7.4	5.31	2.76	98
Propensity covariate x_n	0	0.80	2.36	1.33	16
Contagion intercept	−8.0	−8.2	0.95	0.67	100
Infectiousness x_s	3.0	3.1	0.40	0.35	100
Social proximity	4.0	4.0	0.45	0.34	100
Panel 6: Infectiousness Effect Estimated as Susceptibility and Infectiousness					
Propensity intercept	−6.0	−5.7	0.64	0.73	100
Contagion intercept	−8.0	−8.1	0.76	0.65	100
Susceptibility x_n	0	0.0	0.11	0.11	12
Infectiousness x_s	3.0	3.0	0.43	0.35	100
Social proximity	4.0	4.0	0.27	0.34	100

Note: Both the omitted and incorrectly specified effects are denoted by x to stress that they refer to the same covariate. When the covariate characterizes variation in cases at risk, we subscript it by n; when it characterizes variation in prior adopters (spreaders), we subscript it by s.

from propensity and susceptibility, and the other parameters in the model are estimated without bias or inefficiency.

We stress the importance of this result. Although an analyst may have plausible grounds for suspecting that a covariate affects the diffusion process in a particular way, such intuitions are often not yet supported by any substantial base of empirical knowledge. An effective exploratory strategy is thus critical. The above results strongly recommend parallel rather than serial analysis of multiple modes of influence.

4.5. *Multiple Effects of a Single Variable*

The above findings immediately draw attention to processes where a single covariate really does exert multiple forms of influence. For example, a covariate might simultaneously increase a case's propensity to adopt *and* its susceptibility to the prior adoptions of others; or it might increase susceptibility *and also* diminish infectiousness. It is important to be able to locate such complex patterns of influence, especially if models treating covariates as having several potential forms of influence are employed frequently in exploratory analyses. Further, a study of such processes also provides a worst case of multicollinearity in explanatory factors.

The literature supports the notion that multiple forms of influence are to be anticipated. For example, Strang and Tuma's (1993) reanalysis of Coleman, Katz, and Menzel's (1966) prescription drug study found that centrally located physicians were both more susceptible to influence and less infectious per tie, though more infectious over all ties (Strang and Tuma 1993, table 5). Further, potential multiple effects extend all the way from propensity to social proximity. Strang and Tuma found that physicians with a science orientation were quicker to prescribe "gammanym" than physicians with a patient-centered orientation, and additionally that physicians are especially influenced by physicians who share their orientation.

Table 7 reports estimates of conditions where a single covariate has multiple effects on the diffusion process. For example, Panel 1 shows results for trials where a covariate x_i multiplies the propensity to adopt by 5 while also increasing susceptibility to diffusion by

TABLE 7
Double Effect Model

Parameter	True Value	ML Estimate Mean	ML Estimate SD	ML Estimate Average SE	Reject H_0 (%)
Panel 1: Propensity and Susceptibility					
Propensity intercept	−6.0	−6.4	1.18	1.31	100
Propensity covariate x_n	5.0	5.2	0.69	0.78	100
Contagion intercept	−8.0	−8.1	0.64	0.67	100
Susceptibility x_n	2.0	2.0	0.18	0.16	100
Infectiousness covariate	2.0	2.0	0.42	0.45	98
Social proximity	4.0	4.0	0.32	0.43	100
Panel 2: Propensity and Infectiousness					
Propensity intercept	−6.0	−5.9	0.69	0.79	100
Propensity covariate x_n	5.0	5.0	0.49	0.52	100
Contagion intercept	−8.0	−8.8	2.26	1.51	96
Susceptibility	2.0	2.0	0.16	0.14	100
Infectiousness x_s	2.0	2.4	1.23	0.79	96
Social proximity	4.0	4.0	0.48	0.47	100
Panel 3: Susceptibility and Infectiousness					
Propensity intercept	−6.0	−6.0	0.82	0.78	100
Propensity covariate	5.0	5.1	0.59	0.52	100
Contagion intercept	−8.0	−8.0	0.90	0.78	100
Susceptibility x_n	2.0	2.0	0.11	0.13	100
Infectiousness x_s	2.0	2.0	0.50	0.46	98
Social proximity	4.0	3.9	0.44	0.44	100
Panel 4: Propensity and Susceptibility					
Propensity intercept	−6.0	−6.3	0.82	0.69	100
Propensity covariate x_n	5.0	5.3	0.63	0.49	100
Contagion intercept	−8.0	−8.1	0.80	0.73	100
Susceptibility x_n	−2.0	2.0	0.21	0.19	100
Infectiousness	2.0	2.1	0.38	0.43	100
Social proximity	4.0	4.0	0.43	0.45	100
Panel 5: Propensity and Infectiousness					
Propensity intercept	−6.0	−5.7	0.59	0.51	100
Propensity covariate x_n	5.0	4.9	0.46	0.40	100
Contagion intercept	−8.0	−8.1	0.43	0.45	100
Susceptibility	2.0	2.0	0.14	0.14	100
Infectiousness x_s	−2.0	−2.0	0.31	0.31	100
Social proximity	4.0	4.0	0.43	0.41	100

TABLE 7 (contd.)

Parameter	True Value	ML Estimate Mean	SD	Average SE	Reject H_0 (%)
Panel 6: Susceptibility and Infectiousness					
Propensity intercept	−6.0	−6.0	0.59	0.59	100
Propensity covariate	5.0	5.0	0.44	0.43	100
Contagion intercept	−8.0	−8.1	0.60	0.59	100
Susceptibility x_n	2.0	2.0	0.15	0.14	100
Infectiousness x_s	−2.0	−2.1	0.45	0.44	100
Social proximity	4.0	4.0	0.52	0.46	100

Note: Both the omitted and incorrectly specified effects are denoted by x to stress that they refer to the same covariate. When the covariate characterizes variation in cases at risk, we subscript it by n; when it characterizes variation in prior adopters (spreaders), we subscript it by s.

2. We examine six sets of "just-specified" models, testing both same-sign and opposite-sign pairs of effects.[11]

Multiple influences are clearly distinguished and accurately estimated in these correctly specified models. Throughout, estimated parameters are close to their true values and indices of variability in the estimates are at normal levels. The matrix of correlations between parameters (not reported here) does not show any unusual patterns, and the null hypothesis of zero effect is correctly rejected in 100 percent of the trials. None of the parameter combinations produces problems. Multiple effects of a single variable (and by implication, high correlations among conceptually distinct measures) do not impede estimation when the effects are located in different parts of the model.[12]

5. DISCUSSION

This paper has explored the estimation of a heterogeneous diffusion process from the perspective of practical research issues. We believe

[11]Covariates that increase the propensity to adopt may diminish susceptibility to contagion. For example, Strang and Bradburn (1993) find that states with high health-care costs have a high propensity to pass health-care legislation, but are less susceptible to the influence of laws passed by other states.

[12]Further work might pursue this problem, giving special attention to correlations between covariates and social proximity effects since much research shows that network ties are likely to be established by similar actors (Marsden 1988).

that substantial theoretical benefits result from developing sociological models that explicitly treat social networks as channeling behavior. While relatively few empirical analyses focus single-mindedly on the structure of intrapopulation influences, a variety of studies employ social contagion as one component of a sociological explanation (for example, see Fligstein 1987; Zhou 1993). But relatively little is known about the conditions under which heterogeneous diffusion models can be effectively estimated.

Overall, our Monte Carlo studies suggest that complex heterogeneous diffusion models can be reliably estimated within an event-history framework. But a number of potential analytic problems arise that empirical research would do well to take into account. In most instances, the studies reported above suggest measurement and modeling strategies that reduce the risk of inappropriate inference to acceptable levels.

Graphical examination of the hazard rate or integrated hazard rate estimated from event histories generated by a heterogeneous diffusion model suggests the flexibility of this model. Apparent time dependence in the population-level hazard rate may be positive, negative, or nonmonotonic, depending on whether the diffusion process is dominated by heterogeneity in propensity, susceptibility, infectiousness, or social proximity. Although graphical analysis can be used to suggest what sorts of effects may be dominant, it should not be employed to decide whether a process involves social contagion. This decision must rely on substantive theory and statistical inference from appropriately specified models.

Maximum likelihood estimation of correctly specified models appears able to decipher the structure of the kinds of heterogeneous diffusion processes analyzed here, with little variation in estimator quality over parameter values. Our simulation studies indicate that continuous measures of social proximity do better than dichotomous indicators, both by avoiding false positives in homogeneous diffusion processes and by correctly inferring strong effects of social proximity. Sparse social networks with powerful contagion effects proved difficult to disentangle from homogeneous diffusion processes. And dense social networks provide too many possible lines of influence to be estimated well. An effective strategy under these conditions is to acknowledge the strong impact of network linkages explicitly and treat all social contagion as channeled through these relations.

To explore the consequences of model misspecification, we examined the effects of including extraneous variables, omitting generating variables, omitting some linkages from a focal case's social network, and misidentifying the location of a covariate's effect. All of these are relatively standard problems for empirical research, where analysts (hopefully) do not write Monte Carlo programs to generate their data.

The inclusion of extraneous covariates does not impede the estimation of covariates with nonzero effects. But false positives for putative infectiousness and effects of binary measures of social proximity arise almost twice as often as standard theory for linear models would suggest. This result warrants closer theoretical inspection—we are unaware of analytic attention to the properties of covariates that characterize cases other than those at risk. A simple strategy when assessing possible effects of infectiousness and social proximity is to employ conservative significance levels.

The exclusion of generating covariates (in the event-history literature, the classic issue of unobserved heterogeneity) mainly produces localized reductions in estimator quality. Omitting a measure of the propensity to adopt degrades the quality of estimates of propensity component but not those in the contagion component. Omitting variables from the contagion component degrades the quality of estimated parameters in the contagion component and has smaller but noticeable effects on the quality of parameter estimates in the propensity component.

Finally, and perhaps most importantly, we considered whether estimation could discover what sort of effect a covariate has within a heterogeneous diffusion model. If locations are examined serially, the answer is no. Most prominently, propensity and susceptibility to contagion cannot be easily separated. False positives result from modeling propensity as susceptibility and also from modeling susceptibility as propensity.

However, models that treat covariates as simultaneously affecting the diffusion process in multiple ways prove a powerful instrument for exposing true relationships. For example, when a covariate that actually affects the propensity to adopt is modeled as affecting both the propensity to adopt and susceptibility to prior adoptions of others, the former effect is accurately captured and the latter is correctly identified as zero. Further, event histories where variables

actually impact the process in multiple ways can be accurately estimated when models are correctly specified. These results strongly recommend an exploratory analysis of multiple effects on the diffusion process where prior theory does not insist on a more parsimonious approach.

REFERENCES

Aalen, Odd. 1978. "Nonparametric Inferences for a Family of Counting Process." *Annals of Statistics* 6:701–26.

Bartholomew, David J. 1982. *Stochastic Models for Social Processes,* 3rd ed. New York: Wiley.

Burns, Lawton R., and Douglas R. Wholey. 1993. "Adoption and Abandonment of Matrix Management Programs: Effects of Organizational Characteristics and Interorganizational Networks." *Academy of Management Journal* 36:106–38.

Burt, Ronald S. 1983. "Cohesion Versus Structural Equivalence as a Basis for Network Subgroups." In *Applied Network Analysis,* edited by R. S. Burt and M. J. Minor, 262–82. Beverly Hills: Sage.

———. 1987. "Social Contagion and Innovation: Cohesion Versus Structural Equivalence." *American Journal of Sociology* 92:1287–335.

Coleman, James S., Elihu Katz, and Herbert Menzel. 1966. *Medical Innovation.* New York: Bobbs-Merrill.

Davis, Gerald F. 1991. "Agents Without Principles? The Spread of the Poison Pill Through the Intercorporate Network." *Administrative Science Quarterly* 36:583–613.

Doreian, Patrick. 1981. "Estimating Linear Models with Spatially Distributed Data." In *Sociological Methodology 1981,* edited by Samuel Leinhardt, 359–88. San Francisco: Jossey-Bass.

Fligstein, Neil. 1987. "The Intraorganizational Power Struggle: The Rise of Finance Personnel to Top Leadership in Large Corporations, 1919–1979." *American Sociological Review* 52:44–58.

Friedkin, Noah E. 1984. "Structural Cohesion and Equivalence Explanations of Social Homogeneity." *Sociological Methods and Research* 12:235–61.

Friedman, Jerome H. 1984. "A Variable Span Smoother." Technical Report No. 5. Stanford: Stanford University, Department of Statistics.

Greve, Henrich R. 1994. "Patterns of Competition: The Diffusion of a Radio Broadcasting Strategy." Academy of Management Best Paper Proceedings 1994.

Greve, Henrich R. 1995. "Jumping Ship: The Diffusion of Strategy Abandonment." *Administrative Science Quarterly*, forthcoming.

Greve, Henrich R., David Strang, and Nancy Brandon Tuma. 1993. "Estimation of Diffusion Processes from Incomplete Populations." Paper presented at the annual meetings of the American Sociological Association, Miami, Florida.

Knoke, David. 1982. "The Spread of Municipal Reform: Temporal, Spatial, and Social Dynamics." *American Journal of Sociology* 87:1314–39.
Levin, Sharon G., Stanford L. Levin, and John B. Meisel. 1987. "A Dynamic Analysis of the Adoption of a New Technology: The Case of Optical Scanners." *The Review of Economics and Statistics* 69:12–17.
Mahajan, Vijay, Eitan Muller, and Frank M. Bass. 1990. "New Product Diffusion Models in Marketing: A Review and Directions for Research." *Journal of Marketing* 54:1–28.
Mahajan, Vijay, and Robert A. Peterson. 1985. *Models for Innovation Diffusion.* Beverly Hills: Sage.
Marsden, Peter V. 1988. "Homogeneity in Confiding Relations." *Social Networks* 10:57–76.
Marsden, Peter V., and Noah E. Friedkin. 1993. "Network Studies of Social Influence." *Sociological Methods and Research* 22:127–51.
Marsden, Peter V., and Joel Podolny. 1990. "Dynamic Analysis of Network Diffusion Processes." In *Social Networks through Time,* edited by H. Flap and J. Weesie, 197–214. Utrecht: ISOR.
Morris, Martina. 1993. "Epidemiology and Social Networks: Modeling Structured Diffusion." *Sociological Methods and Research* 22:99–126.
Nelson, Wayne. 1972. "Theory and Application of Hazard Plotting for Censored Failure Data." *Technometrics* 14:945–65.
Strang, David. 1990. "From Dependency to Sovereignty: An Event History Analysis of Decolonization, 1870–1987." *American Sociological Review* 55:846–60.
Strang, David. 1991*a*. "Global Patterns of Decolonization, 1500–1987." *International Studies Quarterly* 35:429–54.
Strang, David. 1991*b*. "Adding Social Structure to Diffusion Models: An Event History Framework." *Sociological Methods and Research* 19:324–53.
Strang, David, and Ellen M. Bradburn. 1993. "Theorizing Legitimacy or Legitimating Theory? Competing Institutional Accounts of HMO Policy, 1970–1989." Paper presented at the annual meetings of the American Sociological Association, Miami, Florida.
Strang, David, and Nancy Brandon Tuma. 1993. "Spatial and Temporal Heterogeneity in Diffusion." *American Journal of Sociology* 99:614–39.
Tuma, Nancy Brandon. 1980. *Invoking RATE.* Menlo Park, Calif.: SRI International.
Tuma, Nancy Brandon, and Michael T. Hannan. 1984. *Social Dynamics: Models and Methods.* Orlando, Fla.: Academic Press.
Tuma, Nancy Brandon, and Paul Ingram. 1993. "Competition and Heterogeneity in Organizational Populations." Paper presented at the annual meetings of the American Sociological Association, Miami, Florida.
Zhou, Xueguang. 1993. "The Diffusion of Licensing in the American States, 1890 to 1950." *American Sociological Review* 58:536–52.

11

STATISTICAL INFERENCE FOR APPARENT POPULATIONS

*Richard A. Berk**
Bruce Western†
*Robert E. Weiss**

In this paper we consider statistical inference for datasets that are not replicable. We call these datasets, which are common in sociology, apparent populations. We review how such data are usually analyzed by sociologists and then suggest that perhaps a Bayesian approach has merit as an alternative. We illustrate our views with an empirical example.

1. INTRODUCTION

It is common in sociological publications to find statistical inference applied to datasets that are not samples in the usual sense. For the substantive issues being addressed, the data on hand are all the data there are. No additional data could be collected, even in principle. In this paper, we call the complete set of all units comprising such datasets an "apparent population." Consider the following examples.

Thanks go to Jan de Leeuw, David Freedman, Edward Leamer, Roderick Little, William Mason, and Donald Rubin for the many discussions that helped us think about the issues raised in this paper. Additional and special thanks go to David Freedman for his detailed written comments. Finally, we are indebted to several reviewers who held our feet to the fire. If we still do not have it right, it is not for the lack of smart, knowledgeable, and helpful colleagues.

*University of California, Los Angeles
†Princeton University